PACLIC 34 (2020)

Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation

24–26 October, 2020
University of Science, Vietnam National University
Hanoi, Vietnam

ISSN 2619-7782

Acknowledgments

PACLIC 34 is hosted by University of Science, Vietnam National University, Hanoi in conjunction with The Association for Vietnamese Language and Speech Processing.

Foreword

The 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 34) is organized by the VNU University of Science, October 24–26, 2020. This edition of the PACLIC series of conferences, as its long tradition, also emphasizes the synergy of theoretical analysis and processing of natural language, aiming to enhance the interaction between researchers working in different fields of language study in the Asia-Pacific region as well as around the world.

For the first time in the history of PACLIC series organization, the conference is organized totally online due to the COVID-19 pandemic. We received 112 submissions, out of which 40 were accepted for oral presentations and 22 for poster presentations. The acceptance rate for oral presentations and poster presentations are 36% and 19% respectively. In addition to oral and poster presentations, the conference highlights four keynote talks and one satellite workshop. We are grateful to Alexander Waibel, Harald Baayen, Yunyao Li, Valia Kordoni for accepting to give a keynote talk. We also thank Jong-Bok Kim, Valia Kordoni and Thi Minh Huyen Nguyen for organizing the Workshop on Multi-word Expression in Asian Languages during the conference. Six papers have been accepted to present at this workshop.

PACLIC 34 would not be made possible without the support from many people, especially in the worldwide pandemic situation. We would like to express our sincere gratitude toward program committee members and sub-reviewers whose professional reviews allowed us to maintain the high quality standard of PACLIC. A special thank goes to Giang Son Tran from University of Science and Technology of Hanoi for maintaining the conference website. We are deeply indebted to the local organizing committee Phuong Le-Hong, The Quyen Ngo and My Linh Ha, as well as student staff members from VNU University of Science. We would also like to thank The Association for Vietnamese Language and Speech Processing (VLSP) for their financial and scientific support for the conference.

Le Minh Nguyen
Chi Mai Luong
Sanghoun Song

PACLIC 34 Program Committee Chairs

Organizers

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

Organizing Committee

NGUYEN Thi Minh Huyen, VNU University of Science (Chair)

Ryo Otoguro, Waseda University (Co-chair)

PHAN Xuan Hieu, VNU University of Engineering and Technology (Co-chair)

Yasuhiro Katagiri, Future University Hakodate (Honorary chair)

PHAM Bao Son, Vietnam National University, Hanoi (Honorary chair)

Local Organizing Committee

LE Hong Phuong, VNU University of Science (Chair)

TRAN Giang Son, University of Science and Technology of Hanoi (Co-chair)

HA My Linh, VNU University of Science

NGO The Quyen, VNU University of Science

TRAN Mai Vu, VNU University of Engineering and Technology

TRAN Thi Oanh, VNU International School

Program Committee Chairs

LUONG Chi Mai, IOIT, Vietnam Academy of Science and Technology

NGUYEN Le Minh, Japan Advanced Institute of Science and Technology

Sanghoun Song, Korea University

Reviewers

Wirote Aroonmanakun
Nguyen Bach
Philippe Blache
Thanh Hung Bui
Jasper Kyle Catapang
Alvin Cheng-Hsien Chen
Emmanuele Chersoni
Sung-Kwon Choi
Anh-Hien Dao
Dien Dinh
Alex Chengyu Fang
Helena Gao
Thanh-Le Ha
Yasunari Harada
Jeffrey J. Holliday
Munpyo Hong
Miao-Ling Hsieh
Shu-Kai Hsieh
Chu-Ren Huang
Shin'Ichiro Ishikawa
Jong-Bok Kim
Valia Kordoni
Pei-Jung Kuo
Oi Yee Kwong
Huei-Ling Lai
Huong Thanh Le
Yong-Hun Lee
Phuong Le-Hong
Dongsik Lim
Te-Hsin Liu
Wei Lu
Chi Mai Luong
Erlyn Manguilimotan
Yuji Matsumoto
James Myers
Ponrudee Netisopakul
Xuan Bach Ngo
Le Minh Nguyen
Minh Thuan Nguyen
Minh-Tien Nguyen
Thi Minh Huyen Nguyen
Thi Thu Trang Nguyen
Vinh Van Nguyen
Hoang Ky Nguyen
Tien Ha Nguyen
Tien Huy Nguyen
Jian-Yun Nie
Nathaniel Oco
Ethel Ong
Chutamanee Onsuwan
Jong C. Park
Hien Pham
Luan Pham
Quang Nhat Minh Pham
Anh Phan
Nattama Pongpairoj
Rodolfo Jr Raga
Lavinia Salicchi
Masashi Saraki
Shu-Ing Shyu
Melanie Siegel
Pornsiri Singhapreecha
Sanghoun Song
Zhiyang Teng
Oanh Tran
Vu Tran
Hong Viet Tran
Yuen-Hsien Tseng
Benjamin Tsou
Yasushi Tsubota
Sinh Vu
Hiroko Wakamatsu
Xinyu Wang
Tak-Sum Wong
Jiun-Shiung Wu
Rong Xiang
Cheng-Zen Yang
Daisuke Yokomori
Satoru Yokoyama
Liang-Chih Yu
Zhang Yu
Niina Ning Zhang
Yuxiang Zhou

Invited Speakers:

Alexander Waibel, Carnegie Mellon University, Karlsruhe Institute of Technology
Harald Baayen, University of Tübingen
Yunyao Li, IBM Almaden Research Center
Valia Kordoni, Humboldt-Universität zu Berlin

Invited Talks

Alexander Waibel: Organic Machine Learning for "Intelligent" Language Interfaces

There is good news and bad news in Speech and Language Processing: The good news is: Performance rates have dramatically improved and reach human parity (at least on matched test conditions), and Speech, Dialog, and Translation systems have gone mainstream and have become features of modern Tech Interfaces. The bad news, however: they are still barely usable and certainly not "intelligent". What explains this discrepancy? Intelligence is the ability to respond to change and new situations. Rather than batch learning in static conditions on aggregated data and testing in matched conditions, human intelligence excels by learning and adapting continuously, incrementally and interactively, from mismatched new testing data. They must exploit multimodal information and advance with very little or no data. Learning must be a life-long process with local, personal data. We call this "Organic Machine Learning".

In this talk, I share observations on where the technology is and where it isn't and discuss some early research results with OML. We develop architectures for OML learning and apply them to AI language tasks such as Speech Translation and Speech Dialogs with Humanoid Robots.

Harald Baayen: How long you make your words crucially depends on their meanings

Traditional approaches to human lexical processing assume that words have static form and meaning representations in the lexicon. Measures such as word frequency, number of neighbors, and word length are typically used to probe how word forms are processed. Measures such as number of synonyms or number of synonym sets in WordNet have been found to be useful for gauging semantic effects on lexical processing. Effectively, in research on the mental lexicon, measures of word form play a dominant role. For instance, the Chinese Lexical Database (Sun et al., 2018) makes available more than 200 measures of word form, but no measures of words' meanings. Thus, the role of meaning in lexical processing is still not well understood.

A radically different approach to the mental lexicon is developed within the framework of the "Discriminative Lexicon" (Baayen et al., 2019). Central to this framework are simple fully connected two-layer networks (without hidden layers) that define mappings between high-dimensional numeric representations of word forms and high-dimensional numeric representations of word meanings (using semantic vectors aka word embeddings). These simple networks, formally equivalent to the mathematics underlying multivariate multiple regression, turn out to be surprisingly effective for predicting a wide range of lexical phenomena. In this presentation, the focus will be on predicting the acoustic durations with which words' are realized in speech production. Evidence from English, Vietnamese, and Mandarin Chinese will be presented clarifying that how well a word's form can be learned and predicted from its meaning is the crucial factor shaping its acoustic duration. Since learnability measures substantially out-perform measures such as word frequency as predictors of acoustic duration, the theory of the Discriminative Lexicon appears to provide a useful and productive new framework for understanding human lexical processing.

Yunyao Li: Towards Universal Natural Language Understanding

Understanding the semantics of the natural language is a fundamental task in artificial intelligence. English semantic understanding has reached a mature state and successfully deployed in multiple IBM AI products and services, such as Watson Natural Language Understanding and Watson Compare and Comply. However, scaling existing products/services to support additional languages remain an open challenge. In this talk, we will discuss the open challenges in supporting universal natural language under-

standing. We will share our work in addressing these challenges in the past few years to provide the same unified semantic representation across languages. We will also showcase how such universal semantic understanding of natural languages can enable cross-lingual information extraction in concrete domains (e.g. insurance and compliance) and show promise towards seamless scaling existing NLP capabilities across languages with minimal efforts.

Valia Kordoni: Figurative Language in Big Data

This talk focuses on metaphor analysis in big data, mainly in the area of education, that is, in multi-genre and heterogeneous course material, varying from video lectures, assignments, tutorial text to social web text posted on MOOC blogs and fora. While metaphor has been tackled in Natural Language Processing before, the focus of that research has never simultaneously been on the analysis of multilingual, multi-genre and heterogeneous texts for applications like Machine Translation. The work we will be presenting in this talk has been mainly carried out in TraMOOC (Translation for Massive Open Online Courses), an EU-funded Horizon 2020 collaborative project which has developed reliable Neural Machine Translation for Massive Open Online Courses (MOOCs).

Table of Contents

Regular Papers

Poster Papers

Workshop on Multiword Expressions in Asian languages

Regular Papers

Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese

BLOUIN Baptiste
Aix-Marseille University, IrAsia
Aix-Marseille University, LIS
ENP-China
baptiste.blouin@lis-lab.fr

MAGISTRY Pierre
Aix-Marseille University, IrAsia
ENP-China
pierre@magistry.fr

Abstract

Named Entity Recognition (NER) is a typical sequence labeling task. It remains challenging for Chinese, partly because of the lack of clear typographic word boundaries. Decisions have to be made regarding the choice of basic units which constitute the sequence to be labeled, and their vectorized representation. Recent approaches have shown that character-based models lack the information about larger units (words) which is useful for NER, while word-based models may suffer from the propagation of word segmentation errors and a higher rate of Out-of-Vocabulary (OOV) tokens. In this paper, we propose a new representation of sinograms (Chinese characters) enriched with word boundary information, for which different types of embeddings can be built. Experiments show that our solution outperforms other state-of-the-art models. We also took great care to propose a fully retrainable pipeline, which is available at https://github.com/enp-china/CCSR-NER . It does not rely on pretrained models and can be trained in few days on common hardware.

1 Introduction

The present work explores the task of Named Entity Recognition (NER) in Mandarin Chinese, specifically for cases when relying on large pre-trained models is not an option. This can occur when one has to process domain specific data, or in our case[1], historical texts where language is quite different from the language of the corpora used to pretrain publicly available models, especially words and characters embeddings. The models we propose can be trained in a reasonable time (days) from a relatively small amount of raw data (few hundred millions of characters) on affordable hardware (such as a single GTX 1080 ti).

[1]ENP China, https://www.enpchina.eu/ (ERC No 788476)

Chinese script does not provide a clear and frequent typographic marker for word boundaries. As a result, when addressing the case of Chinese(s) language(s) in NER, we have to face the issue of word segmentation. Recent models proposed in the literature can be divided into character-based, word-based or hybrid models, but every work had to take a stance regarding Chinese Word Segmentation (CWS). The importance and methods for CWS have a long history in Chinese NLP, a recent work Li et al. (2019) makes the strong claim that the neural era of NLP is turning CWS into an irrelevant or even harmful step in a pipeline. However Li et al. (2019) did not provide experimental results on the NER task and our own experiments presented in this paper tend to show that CWS can be either harmful or beneficial, depending on how much care is given to consistency in segmentation and to the way word embeddings are built and used. Our main findings are that off-the-shelf embeddings for Mandarin Chinese must be used carefully, but it is possible to improve on the state-of-the-art by retraining everything from raw and labeled corpora, as we achieve 77.27 (+2.84) of f-score on OntoNotes 4 (Hovy et al., 2006) and 80.64 (+1.04) on OntoNotes 5 with a model simpler than previous state-of-the-art which requires dependency parsing.

The second focus of our study is a comparison between supervised and unsupervised CWS. When targeting a specific downstream NLP task, we ran experiments to decide whether we should follow a specific segmentation guideline by the mean of supervised machine learning, or if consistency brought by an unsupervised system is enough to improve on the downstream (here NER) task. This question is crucial for us to face more ancient texts, for which training data for CWS may not be available. We show that using CWS for the task of named entity recognition allows to provide useful information compared

to using only characters.

In summary, the contributions of this paper are as follows:

- We propose a novel method to combine CWS information and a character-level representation which can be used by a BiLSTM-CRF (Lample et al., 2016) model to improve on Chinese NER task.
- In an attempt to explain this improvement, we study the impact of our new representation on the OOV issue compared to other possible representations.
- We investigate two different strategies of supervised and unsupervised CWS, to assess for the need of manually segmented training corpus.
- The experimental results demonstrate that our proposed method significantly outperforms the current state-of-the-art performance on five different Chinese NER datasets. Our proposed solution does not rely on any pre-trained models, and can be fully trained from corpora of relatively small size on affordable hardware.

2 Related works

Our work relates to existing methods on multiple tasks, including NER, segmentation and embeddings.

2.1 Named Entity Recognition

Our model architecture is similar to that proposed by Huang et al. (2015), which is a bidirectional recurrent neural network (BiLSTMs) with a subsequent conditional random field (CRF) decoding layer. For this kind of architecture we have to choose a level of tokenization for the input. It can result in word-based models, character-based models and hybrid models. A word-based BiLSTM-CRF model applied to Chinese NER will suffer from segmentation errors. Zhang and Yang (2018) and Liu et al. (2019) showed that using a hybrid model to integrate words in character sequence leads to better results for character-based Chinese NER. The main difference between those models is that Zhang and Yang (2018) uses a DAG-structured LSTM to put every potential words that match a lexicon into their model, this requires them to process sentences one by one, whereas Liu et al. (2019) add word information into the input vector. This second approach selects a single segmentation and choose one word for each character without ambiguity.

Another approach to integrate the word segmentation information to the model was proposed by Cao et al. (2018) which involves using multitask on Chinese segmentation to transfer this information to the NER task.

Jie and Lu (2019) propose a more complex approach which integrates dependency parses to the LSTM and relies on pre-trained ELMo contextual embeddings. They obtain promising results on the OntoNotes 5 corpus, but they do not discuss the issue of word segmentation (for which they use the gold segmentation).

2.2 Word Segmentation

Word-level information can be introduced into a NER system in various ways, as a first step of processing or to build an external resource such as a word embeddings lexicon. In any case, it relies on a Chinese Word Segmentation (CWS) system and training corpus in the supervised case. When using pre-trained word embeddings, one implicitly relies on the CWS system which has been used to prepare the embeddings. In our case we conduct two kinds of experiments, the first one is based on supervised CWS for which we use *zpar* (Zhang and Clark, 2007) trained on the Chinese Treebank[1]. Since training data for word segmentation is not available for all domains, languages (to adapt to other sinitic languages, such as Cantonese) or more ancient documents, and can be time consuming or costly to obtain, we also run experiments based on an unsupervised CWS system using *eleve* (Magistry and Sagot, 2012) which requires only an unannotated corpus. We use texts from the Chinese Wikipedia to train the segmenter, which we sampled from the corpus prepared by Majliš and Žabokrtský (2012) down to a size we think consistent to what will be available for future adaptations of our system.

2.3 Embeddings

Vectorized word representations (Turian et al., 2010; Mikolov et al., 2013), especially known as word embeddings, are a key element for multiple NLP

[1] https://catalog.ldc.upenn.edu/LDC2013T21

tasks including NER (Collobert et al., 2011). Today there are three distinct embedding types. Classical word embedding (Pennington et al., 2014; Mikolov et al., 2013) , character-level features (Ma and Hovy, 2016; Zhang and Yang, 2018) and contextualized word embeddings (Peters et al., 2017; Zhang and Yang, 2018). Contextualized word embeddings as been shown to be effective for improving many natural language processing tasks including NER. In our work we use FastText (Bojanowski et al., 2016a) to generate our non-contextual embeddings and Flair Akbik et al. (2018) for the contextual ones. We decided not to use BERT (Devlin et al., 2018) because in our situation we will have to train new embeddings on multiple historical subcorpora of a limited size, which makes BERT either unusable or not affordable. It remains worth noting that we outperform the systems tested in (Jie and Lu, 2019) which rely on ELMo (Peters et al., 2018) and for which the authors report it obtained performances similar to BERT in preliminary experiments.

3 Datasets

The larger project for which we design our models introduces constraints in terms of corpus size and retrainability. We limit ourselves to a reasonable amount of data. Nevertheless, for the experiments presented in this paper, we rely on standard datasets of Modern Chinese, widely used in the literature to be able to provide a comprehensive evaluation.

We limit our raw data to a random sample of 324 millions tokens (243 millions sinograms) taken from the Wikipedia in Mandarin Chinese. We make this sample available for the sake of reproducibility.

For word segmentation, we finally used the Chinese Treebank (CTB) [1] and compare it to an unsupervised word segmentation.[2]

For Named Entities, we use the OntoNotes4 Corpus (Hovy et al., 2006) and follow the de facto standard split and entity types selection from Che et al. (2013). We also evaluate our system against the popular MSRA (Levow, 2006) Weibo NER (Peng and Dredze, 2015) and corpus of resume in Chinese (Zhang and Yang, 2018). Those four datasets represent three different domains, OntoNotes and MSRA datasets are in the news domain, the Chinese resume dataset contains resumes of senior executives from listed companies in the Chinese stock market and the Weibo NER dataset is drawn from the social media website Sina Weibo. Another difference between those datasets is that MSRA , Weibo and Chinese resume did not provide word segmentation for all the sections, unlike OntoNotes4 which has a gold-standard segmentation for the training, development and test sections. We also provide results on OntoNotes5 (Weischedel et al., 2013) to compare our system with Jie and Lu (2019). We summarize the datasets in Table 1.

Dataset	Type	Train	Test	Dev
OntoNotes4 18 classes	Sent	15.7k	4.3k	4.3k
	Char	491.9k	208.1k	200.5k
	Entities	13.4k	7.7k	6.95k
OntoNotes5 18 classes	Sent	38.3k	4.3k	6.3k
	Char	1212k	145k	175k
	Entities	64.1k	7.6k	9.2k
Weibo 4 classes	Sent	1.4k	0.27k	0.27k
	Char	73.8k	14.8k	14.5k
	Entities	1.89k	0.42k	0.39k
Resume 8 classes	Sent	3.8k	0.48k	0.46k
	Char	124.1k	15.1k	13.9k
	Entities	1.34k	0.15k	0.16k
MSRA 3 classes	Sent	46.4k	4.4k	-
	Char	2169.9k	172.6k	-
	Entities	74.8k	6.2k	-

Table 1: Statistics of the datasets

4 Methods

4.1 Contextual Character Embeddings

Contextual word embeddings have shown to improve state-of-the-art on several NLP tasks. One of our contribution is to propose two new kinds of contextual embeddings at the character level which can take into account word boundary information.

Referring to Akbik et al. (2018) paper which introduces a word-level embeddings based on a character-level language model, we introduce a sinogram embedding using their character language model (LM). Where the LM allows the text to be treated as a sequence of characters passed to an LSTM which at each point in the sequence is trained

[2] we also tried to use the dataset from Peking University (PKU) and Microsoft Research (MSR) provided for the CWS Bakeoff 2 (http://sighan.cs.uchicago.edu/bakeoff2005/) but it did not make any noticeable differences.

to predict the next character. In our system, we train the LM to produce characters with segmentation information. Given a sequence of characters ($C_0, C_1, ..., C_N$) we learn $P(C_i|C_0, ..., C_{i-1})$, an estimate of the predictive distribution over the next character given past characters. We utilize the hidden states of a forward-backward recurrent neural network to create contextualized character embeddings. The final contextual character representation is given by :

$$C_i^{LM} = \begin{bmatrix} C_i^f \\ C_{T-i}^b \end{bmatrix}$$

Where C_i^f denote the hidden state at position i of the forward LM and C_{T-i}^b denote the hidden state at position $T - i$ of the backward LM.

4.2 Contextual Character with segmentation information Embeddings

In this work, we investigate the different ways to inject the CWS information into a NER pipeline. Several approaches propose to directly use the word-tokens as segmented by a CWS system, they showed that discrepancies between the output of the CWS and the NE annotation can be harmful for NER. Out-of-Vocabulary (OOV) tokens is another common issue for NER. In order to tackle those issues, we designed a new kind of sinogram representation which contains the information of the chosen word segmentation at the character level. We decide to use the BIES format to represent the CWS (as introduced in Xue and Shen (2003), originally as an intermediary step for CWS) and we train a language model to produce embeddings of those character with BIES tag. As we use a BI-LSTM to process the NER task and as we stay at a character level, our new representation allows us to reconstruct the entire word according to the BIES tag. But in the case of a mismatching segmentation between NE and word, the model can still learn to use this wrong segmentation as the right delimiter of an entity.

4.3 Model Description

We use the Flair framework (Akbik et al., 2019) to create our model (Figure 2.1). The main difference with other existing NER models is that we use stacked embeddings to represent our input. With this kind of architecture we can combine our different kinds of embeddings. Character, word information and bichar embeddings are concatenated to represent each character. The final character representation is given by

$$c_i = \begin{bmatrix} r_i^{char} \\ r_i^{bichar} \\ r_i^{word} \end{bmatrix}$$

The fact that we use character as neural units allows us to give word information associated to a character. In our case, the word information is given by the contextual character with segmentation embeddings. We denote a Chinese sentence as s = $\{c_1, c_2, ..., c_n\}$.

We use an extra linear layer between the input layer and the LSTM's to make the stacked representation trainable. Figure 1 shows the structure of our model. The blue part of the model shows how we use the embeddings. The symbol $\bigoplus$ indicates the possibility to concatenate different kinds of embeddings. Using this approach, we can then add other types of embeddings related to characters. The red part is a BiLSTM-CRF.

5 Experiments

We conducted several experiments to evaluate the effectiveness of our approach across different domains. In addition, we evaluate the importance of the segmentation for our representations by using supervised and non-supervised segmentation approaches. We also investigate on the usefulness of the bichar representation for Chinese Natural Language Processing. Evaluations are reported using standard metrics of precision (P), recall (R) and F1-score (F).

5.1 Experimental Settings

We used the datasets presented in the section 3, including the OntoNotes gold segmentation to evaluate the distance between our supervised/unsupervised segmentations and whether this distance makes a difference to our overall process.

Embeddings. We used FastText (Bojanowski et al., 2016b) to pretrain characters and bi-characters embeddings on a subset of 7 millions sentences from Chinese Wikipedia dump. for both of these representations we used a context of bi-character.

Hyper-parameter. Table 2 shows the values of hyper-parameters for our models, which were fixed

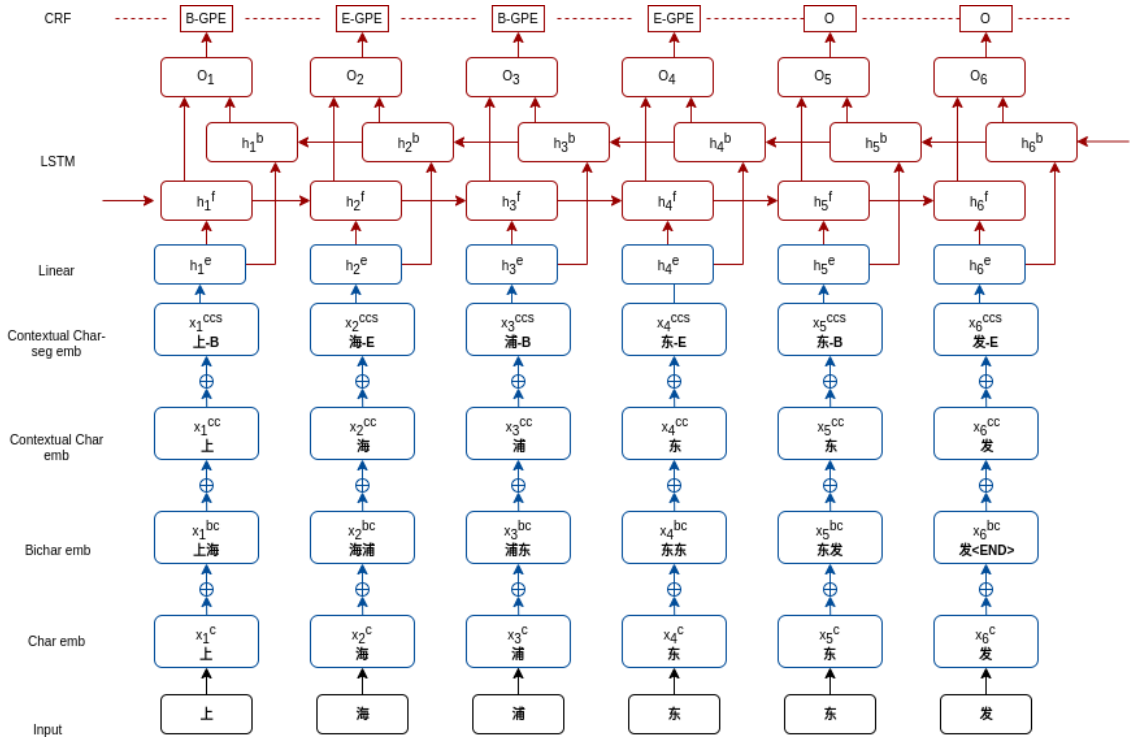


Figure 1: Architecture of the model and representation of our embeddings

Parameter	value	Parameter	value
Char emb size	50	Bichar emb size	50
LSTM hidden	256	LSTM layer	1
Learning rate	0.1	Anneal factor	0.5
Emb dropout	0.05	batch size	16

Table 2: Hyper-parameter values.

without specific grid search adjustments for each individual dataset. Stochastic gradient descent (SGD) is used for optimization, with an initial learning rate of 0.1 and we divide its value by two if the f-score does not increase on the development corpus during 5 epochs. In that case, we reload the previous best model before dividing the learning rate.

Configurations. In order to evaluate the importance of the different representations, we have set up 8 configurations of embeddings.

- **Char** For this configuration we only use character embeddings. (size : 50)
- **Bichar** For this configuration we only use bi-character embeddings. (size : 50)
- **Bichar + Char** For this configuration we concatenate bi-character and character embeddings. (size : 50 + 50)
- **Char ctx** For this configuration we only use contextual character embeddings. (size : 1024)
- **Char-seg unsup** For this configuration we only use contextual character with segmentation information embeddings where the segmentation comes from the unsupervised segmenter. (size : 1024)
- **Bichar + Char-seg unsup** For this configuration we only concatenate bi-character embeddings to the previous configuration. (size : 50 + 1024)
- **Char-seg ctb** for this configuration we only use contextual character with segmentation information embeddings where the segmentation comes from the supervised segmenter trained on the Chinese Treebank. (size : 1024)
- **Bichar + Char-seg ctb** or this configuration we only concatenate bi-character embeddings to the previous configuration. (size : 50 + 1024)

Input	Models	P	R	F
Gold seg	Wang et al. (2013)	76.43	72.32	74.32
	Che et al. (2013)	77.71	72.51	**75.02**
	Yang et al. (2017)	65.59	71.84	68.57
No seg	Zhang and Yang (2018)	76.35	71.56	73.88
	Char baseline	70.08	60.53	64.95
	Liu et al. (2019)	76.09	72.85	74.43
	Char	67.31	64.33	65.79
	Bichar	72.25	72.18	72.21
	Bichar + Char	74.11	72.75	73.42
	Char ctx	76.79	75.66	76.22
	Char-seg unsup *CSU*	77.54	75.91	76.72
	Bichar + *CSU*	76.3	76.77	76.53
	Char-seg ctb *CSC*	77.81	76.21	77
	Bichar + *CSC*	77.67	76.89	**77.27**[3]

Table 3: NER results for named entities on the OntoNotes 4 dataset. There are three blocks. The first two blocks contain the previous state-of-the-art models where "Gold seg" means that they used the reference segmentation proposed by the dataset and "No seg" means that they used other approaches that do not rely on reference segmentation. The last block lists the performance of our proposed model.

5.2 Experimental results

OntoNotes. Table 3 shows the experimental results on OntoNotes 4 dataset. The first column (Input) shows the representations of input sentence that was used. "Gold seg" means that they used the segmentation provided by the corpus to represent the word in the sentence, "No seg" means that we used only the character as input and other approaches that do not benefit from the reference segmentation to provide information about the word level.

The first part of table 3 are the results of Wang et al. (2013); Che et al. (2013); Yang et al. (2017) . These three approaches rely on gold segmentation at the word level, with character embeddings. Che et al. (2013) achieve good performance with 75.02 F-score. Here we exceed this score without using the gold segmentation.

The second part shows the performances of more recent approaches (Zhang and Yang, 2018; Liu et al., 2019) and a character baseline which is the original character-based BILSTM-CRF model. Zhang and Yang (2018) proposes a lattice LSTM to exploit word information in character sequence and Liu et al. (2019) use a new word-character LSTM model to add word information on the first or on the last character of each word. These two approaches show a significant improvement compared to the character baseline, which illustrates the importance of the word information in character sequence.

The last part of the table 3 shows the results of our configurations. The first three rows show results where we only used the character information. Through these results we show that bichar representations are very efficient for Chinese. This may be explained by the fact that bichars have a length closer to the average word length and provide more contextual information than single characters. The last four rows show the results of using our contextual char-seg representations. Those configurations achieve very good results, improving the state of the art, beating both models that do not use gold segmentation and even those that do. Firstly, these results show that the information about the boundaries of a word is useful. Secondly, on this corpus, we can see that there is only a slight difference between using supervised and unsupervised segmentation. Which is very encouraging to address situations where we do not have adequate CWS training data.

Weibo NER. Table 4 shows the experimental results on Weibo NER dataset. This dataset proposes two kinds of annotations, named entities and nominal entities. For our experiments we only evaluated the combination of these two annotations. Compared to the other corpus, this one offers few annotated data, that is why different approaches have been proposed. Peng and Dredze (2015, 2016); Cao et al. (2018) use multitask learning and He and Sun (2017) use semi-supervised learning. As a result of these approaches, they use cross-domain or semi-supervised additional data. In contrast, Zhang and Yang (2018); Liu et al. (2019) and our model do not need any additional data.

These results exhibit similar patterns as those on OntoNotes. However in this case the unsupervised CWS can even lead to higher scores. This may be the result of Weibo Corpus being drawn from social media. A CWS system trained on the CTB is better suited for the news domain and less reliable in the Weibo case.

[3]This result is the average of 20 runs. The results of these runs have a variance of 4.10^{-2}

Models	P	R	F
Peng and Dredze (2015)	-	-	56.05
Peng and Dredze (2016)	-	-	58.99
He and Sun (2017)	-	-	58.23
He and Sun (2017)	-	-	54.82
Cao et al. (2018)	-	-	58.70
Zhang and Yang (2018)	-	-	58.79
Liu et al. (2019)	-	-	59.84
char baseline	-	-	52.88
Char	72.14	34.69	46.85
Bichar	72.63	33.01	45.39
Bichar Char	69.73	49.04	57.58
Char ctx	66.67	52.63	58.82
Char-seg unsup	66.48	55.98	60.78
Bichar + Char-seg unsup	70.37	59.09	**64.24**
Char-seg ctb	71.25	55.74	62.55
Bichar + Char-seg ctb	67.24	56.46	61.38

Table 4: Weibo NER results

Models	P	R	F
Zhang and Yang (2018)	94.81	94.11	94.46
Liu et al. (2019)	95.27	95.15	95.21
char baseline	93.26	93.44	93.35
Char	92.76	94.36	93.55
Bichar	93.64	94.79	94.21
Bichar Char	93.93	94.97	94.45
Char ctx	94.39	95.03	94.71
Char-seg unsup	94.77	95.58	95.17
Bichar + Char-seg unsup	94.56	94.91	94.73
Char-seg ctb	94.84	94.66	94.75
Bichar + Char-seg ctb	95.07	95.83	**95.45**

Table 5: Chinese resume results

Models	P	R	F
Zhang et al. (2006)	92.20	90.18	91.18
Zhou et al. (2013)	91.86	88.75	90.28
Dong et al. (2016)	91.28	90.62	90.95
Cao et al. (2018)	91.73	89.58	90.64
Zhang and Yang (2018)	93.57	92.79	93.18
Liu et al. (2019)	94.33	93.11	93.71
char baseline	89.61	86.98	88.37
Char	84.95	84.37	84.66
Bichar	87.3	83.74	85.48
Bichar Char	90.13	89.74	89.93
Char ctx	90.6	88.58	89.58
Char-seg unsup	94.77	93.43	94.1
Bichar + Char-seg unsup	94.93	93.38	**94.15**
Char-seg ctb	93.63	91.42	92.51
Bichar + Char-seg ctb	93.73	91.78	92.74

Table 6: MSRA results

Models	P	R	F
Zhang and Yang (2018)	76.34	77.01	76.67
Jie and Lu (2019)			
BiLSTM-CRF	77.94	75.33	76.61
BiLSTM-CRF + ELMo	79.20	79.21	79.20
DGLSTM-CRF + ELMo	78.86	81.00	79.92
without Gold dep.			79.59
Bichar + Char-seg ctb	80.70	80.60	**80.65**

Table 7: Ontonotes 5 results. Jie and Lu (2019) provide detailed results on gold segmentation and parsing only. An F-measure of 79.59 is obtained with non-gold dependencies, but the authors did not report experiments related to the quality of the word segmentation.

Resume Table 5 shows the experimental results on Resume dataset. These are consistent with the observations made on OntoNotes and Weibo NER. Our model achieves good results on this dataset, but unlike the other corpora, very good results were already obtained by other systems. It does not allow us to highlight our approach as much as the other corpora.

MSRA Table 6 shows the experimental results on MSRA dataset. The best results are obtained with the unsupervised segmentation.

Ontonotes 5 To complete our evaluation, we run our best model from the Ontonotes 4 experiment on Ontonotes 5 to provide comparison with Jie and Lu (2019). Results are shown Table 7. Note that the comparison is somewhat unfair as Jie and Lu (2019) rely on gold segmentation. Nevertheless, our system obtains the highest results, without the need for a dependency parser.

The embeddings we propose achieve state-of-the-art results on a diversity domains such as news, social media, and Chinese resume.

5.3 Out Of Vocabulary analysis

When using a model with word-level features, one of the most common problems comes from unknown words. Our approach which injects segmentation information at the characters level allows to rebuild the words from characters and leads to fewer unknowns.

To do so, we used two types of segmentation, word level and char-seg level, in a supervised and unsupervised way to segment our Wikipedia sample. Once our four Wikipedia samples were segmented,

embeddings	OntoNotes seg	OOV
Word ctb	supervised ctb	18.89 %
Word ctb	gold	18.96 %
Word unsup	unsupervised	32.34 %
Word unsup	gold	35.81 %
Char-seg ctb	supervised ctb	0.67 %
Char-seg ctb	gold	0.28 %
Char-seg unsup	unsupervised	0.95 %
Char-seg unsup	gold	1.78 %

Table 8: OOV statistics on OntoNotes 4 with supervised and unsupervised segmentation.

we trained four different FastText to obtain 4 lexicons for each of them. To evaluate the OOV rate on OntoNotes, we segmented it in three different ways in order to compare for each case the presence or not of words in the lexicons generated by our embeddings. We segmented OntoNotes in a supervised and unsupervised way with the same two models we used to segment Wikipedia and in a last step we left the "gold" segmentation in words proposed by OntoNotes. Results of this experiments are shown in table 8.

For the embeddings column, we have two levels of segmentation, in word and char-seg, and two levels of supervision, "ctb" for the supervised part trained on the Chinese TreeBank and "unsup" for the unsupervised part. The OntoNotes seg column represents the three types of segmentation used to segregate OntoNotes into words. Because OntoNotes is segmented into words and because our lexicon for our char-seg embeddings contains only characters with segmentation information, for a given word coming from OntoNotes, we try to reconstruct the char-seg sequence constituting this word from our embedding lexicon. For example, for the word 越南 we are looking for char-seg 越-B and 南-E in our embedding lexicon. If a char-seg is missing, then the whole word is missing too.

The results show our representations greatly decrease the unknown word rates. it allows us to have a representation for most of the words. Moreover, unlike traditional word representations, we do not have fixed representations of our words, which makes it easier to have representations for new words, but which can then call into question the quality of our representations.

6 Discussions

Annotation ambiguity. The named entity recognition task combines a step of segmentation with one of classification. We feel the need to question some cases of ambiguity from the data. By using the guideline from OntoNotes we annotated in-house data and we found it difficult in some cases to choose between Geopolitical Entity (GPE) and Location (LOC). This case of ambiguity has a direct impact on our predictions. we noted that more than $\frac{1}{3}$ of LOC that has been detected is annotated as a GPE, which is consistent with the difficulties encountered in our annotation experiment.

Another issue arises from the conversion of the OntoNotes 4 corpus from 18 classes to 4. Most notably the entity types NORP (Nationality, Other, Religion, Political) and FAC (Facility). These classes are discarded in the 4-classes version, but are typical cases of nested entities containing a GPE, LOC or ORG, which is also discarded in the process, creating erroneous annotation.

Entity segmentation against word segmentation. Our results show that although staying at the character level allows us to tackle the OOV issue, the information brought by CWS is still what enables us to reach the highest scores. In the cases when the CTB segmentation guidelines are consistent with the NER corpus, supervised segmentation performs better. However NER with unsupervised segmentation is close in these cases and can perform better in other cases. So our answer to Li et al. (2019) could be that Word Segmentation is actually necessary, but unsupervised CWS may be enough.

7 Conclusion and future works

In this paper, we propose new sinogram embeddings which includes word information at the character level for Chinese NER. Our proposed approach shows that adding CWS label to a character allows to give word level information while reducing considerably the number of OOV compared to a word sequence. Our experiments on multiple datasets, in different domains, show that our system outperforms previous state-of-the-art approaches. This paves the road to NER in more challenging situations such as historical documents or less-resourced situations.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. https://www.aclweb.org/anthology/N19-4010 FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. https://www.aclweb.org/anthology/C18-1139 Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016a. http://arxiv.org/abs/1607.04606 Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016b. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. https://www.aclweb.org/anthology/D18-1017 Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Wanxiang Che, Mengqiu Wang, Christopher D. Manning, and Ting Liu. 2013. https://www.aclweb.org/anthology/N13-1006 Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 52–62, Atlanta, Georgia. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. http://dl.acm.org/citation.cfm?id=2078183.2078186 Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 999888:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. http://arxiv.org/abs/1810.04805 BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. https://doi.org/10.1007/978-3-319-50496-4_20 Character-based lstm-crf with radical-level features for chinese named entity recognition. volume 10102, pages 239–250.

Hangfeng He and Xu Sun. 2017. https://www.aclweb.org/anthology/E17-2113 F-score driven max margin neural network for named entity recognition in Chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718, Valencia, Spain. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. http://dl.acm.org/citation.cfm?id=1614049.1614064 Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. http://arxiv.org/abs/1508.01991 Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Zhanming Jie and Wei Lu. 2019. https://doi.org/10.18653/v1/D19-1399 Dependency-guided LSTM-CRF for named entity recognition. 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 3862–3872, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. https://doi.org/10.18653/v1/N16-1030 Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Gina-Anne Levow. 2006. https://www.aclweb.org/anthology/W06-0115 The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. https://doi.org/10.18653/v1/P19-1314 Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252, Florence, Italy. Association for Computational Linguistics.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019. https://www.aclweb.org/anthology/N19-1247 An encoding strategy based word-character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2379–2389. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. http://arxiv.org/abs/1603.01354 End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv e-prints*, page arXiv:1603.01354.

Pierre Magistry and Benoît Sagot. 2012. https://www.aclweb.org/anthology/P12-2075 Unsupervized word segmentation: the case for Mandarin Chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–387, Jeju Island, Korea. Association for Computational Linguistics.

Martin Majliš and Zdeněk Žabokrtský. 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/267_Paper.pdf Language richness of the web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2927–2934, Istanbul, Turkey. European Language Resources Association (ELRA).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Nanyun Peng and Mark Dredze. 2015. https://doi.org/10.18653/v1/D15-1064 Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2016. http://arxiv.org/abs/1608.02689 Multi-task multi-domain representation learning for sequence tagging. *CoRR*, abs/1608.02689.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. https://doi.org/10.18653/v1/N18-1202 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. Association for Computational Linguistics.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. http://arxiv.org/abs/1705.00108 Semi-supervised sequence tagging with bidirectional language models. *CoRR*, abs/1705.00108.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. http://dl.acm.org/citation.cfm?id=1858721 Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013. http://dl.acm.org/citation.cfm?id=2891460.2891588 Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 919–925. AAAI Press.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. https://catalog.ldc.upenn.edu/LDC2013T19 Ontonotes release 5.0. *Linguistic Data Consortium*.

Nianwen Xue and Libin Shen. 2003. https://doi.org/10.3115/1119250.1119278 Chinese word segmentation as lmr tagging.

Jie Yang, Zhiyang Teng, Meishan Zhang, and Yue Zhang. 2017. http://arxiv.org/abs/1708.07279 Combining discrete and neural features for sequence labeling. *CoRR*, abs/1708.07279.

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. https://www.aclweb.org/anthology/W06-0126 Word segmentation and named entity recognition for SIGHAN bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161, Sydney, Australia. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2007. https://www.aclweb.org/anthology/P07-1106 Chinese segmentation with a word-based perceptron algorithm. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 840–847, Prague, Czech Republic. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018. https://doi.org/10.18653/v1/P18-1144 Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.

J. Zhou, W. Qu, and F. Zhang. 2013. Chinese named entity recognition via joint identification and categorization. *Chinese Journal of Electronics*, 22:225–230.

Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models

The Viet Bui[1]
vietbt6@fpt.com.vn

Thi Oanh Tran[1,2]
oanhtt@isvnu.vn

Phuong Le-Hong[1,2]
phuonglh@vnu.edu.vn

[1] FPT Technology Research Institute, FPT University, Hanoi, Vietnam
[2] Vietnam National University, Hanoi, Vietnam

Abstract

This paper describes our study on using mutilingual BERT embeddings and some new neural models for improving sequence tagging tasks for the Vietnamese language. We propose new model architectures and evaluate them extensively on two named entity recognition datasets of VLSP 2016 and VLSP 2018, and on two part-of-speech tagging datasets of VLSP 2010 and VLSP 2013. Our proposed models outperform existing methods and achieve new state-of-the-art results. In particular, we have pushed the accuracy of part-of-speech tagging to 95.40% on the VLSP 2010 corpus, to 96.77% on the VLSP 2013 corpus; and the F_1 score of named entity recognition to 94.07% on the VLSP 2016 corpus, to 90.31% on the VLSP 2018 corpus. Our code and pre-trained models viBERT and vELECTRA are released as open source to facilitate adoption and further research.

1 Introduction

Sequence modeling plays a central role in natural language processing. Many fundamental language processing tasks can be treated as sequence tagging problems, including part-of-speech tagging and named-entity recognition. In this paper, we present our study on adapting and developing the multilingual BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) models for improving Vietnamese part-of-speech tagging (PoS) and named entity recognition (NER).

Many natural language processing tasks have been shown to be greatly benefited from large network pre-trained models. In recent years, these pre-trained models has led to a series of breakthroughs in language representation learning (Radford et al., 2018; Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Clark et al., 2020). Current state-of-the-art representation learning methods for language can be divided into two broad approaches, namely *denoising auto-encoders* and *replaced token detection*.

In the denoising auto-encoder approach, a small subset of tokens of the unlabelled input sequence, typically 15%, is selected; these tokens are masked (e.g., BERT (Devlin et al., 2019)), or attended (e.g., XLNet (Yang et al., 2019)); and then train the network to recover the original input. The network is mostly transformer-based models which learn bidirectional representation. The main disadvantage of these models is that they often require a substantial compute cost because only 15% of the tokens per example is learned while a very large corpus is usually required for the pre-trained models to be effective. In the replaced token detection approach, the model learns to distinguish real input tokens from plausible but synthetically generated replacements (e.g., ELECTRA (Clark et al., 2020)) Instead of masking, this method corrupts the input by replacing some tokens with samples from a proposal distribution. The network is pre-trained as a discriminator that predicts for every token whether it is an original or a replacement. The main advantage of this method is that the model can learn from all input tokens instead of just the small masked-out subset. This is therefore much more efficient, requiring less than $1/4$ of compute cost as compared to RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019).

Both of the approaches belong to the fine-tuning method in natural language processing where we first pretrain a model architecture on a language modeling objective before fine-tuning that same model for a supervised downstream task. A major advantage of this method is that few parameters need to be learned from scratch.

In this paper, we propose some improvements over the recent transformer-based models to push the state-of-the-arts of two common sequence labeling tasks for Vietnamese. Our main contributions in this work are:

- We propose pre-trained language models for Vietnamese which are based on BERT and ELECTRA architectures; the models are trained on large corpora of 10GB and 60GB uncompressed Vietnamese text.
- We propose the fine-tuning methods by using attentional recurrent neural networks instead of the original fine-tuning with linear layers. This improvement helps improve the accuracy of sequence tagging.
- Our proposed system achieves new state-of-the-art results on all the four PoS tagging and NER tasks: achieving 95.04% of accuracy on VLSP 2010, 96.77% of accuracy on VLSP 2013, 94.07% of F_1 score on NER 2016, and 90.31% of F_1 score on NER 2018.
- We release code as open source to facilitate adoption and further research, including pre-trained models viBERT and vELECTRA.

The remainder of this paper is structured as follows. Section 2 presents the methods used in the current work. Section 3 describes the experimental results. Finally, Section 4 concludes the papers and outlines some directions for future work.

2 Models

2.1 BERT Embeddings

2.1.1 BERT

The basic structure of BERT (Devlin et al., 2019) (*Bidirectional Encoder Representations from Transformers*) is summarized on Figure 1 where Trm are transformation and E_k are embeddings of the k-th token.

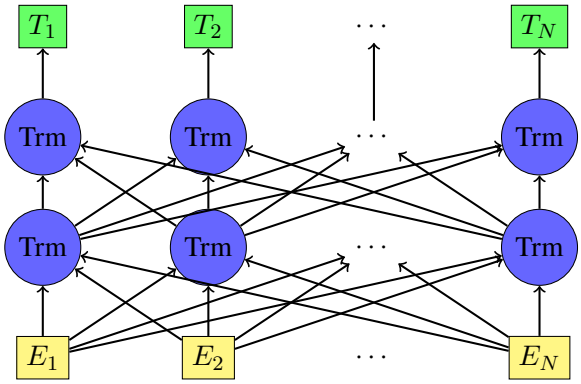


Figure 1: The basic structure of BERT

In essence, BERT's model architecture is a multilayer bidirectional Transformer encoder based on the original implementation described in (Vaswani et al., 2017). In this model, each input token of a sentence is represented by a sum of the corresponding token embedding, its segment embedding and its position embedding. The WordPiece embeddings are used; split word pieces are denoted by ##. In our experiments, we use learned positional embedding with supported sequence lengths up to 256 tokens.

The BERT model trains a deep bidirectional representation by masking some percentage of the input tokens at random and then predicting only those masked tokens. The final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary. We use the whole word masking approach in this work. The masked language model objective is a cross-entropy loss on predicting the masked tokens. BERT uniformly selects 15% of the input tokens for masking. Of the selected tokens, 80% are replaced with [MASK], 10% are left unchanged, and 10% are replaced by a randomly selected vocabulary token.

In our experiment, we start with the open-source mBERT package[1]. We keep the standard hyper-parameters of 12 layers, 768 hidden units, and 12 heads. The model is optimized with Adam (Kingma and Ba, 2015) using the following parameters: $\beta_1 = 0.9, \beta_2 = 0.999$, $\epsilon = 1e-6$ and L_2 weight decay of

[1] https://github.com/google-research/bert/blob/master/multilingual.md

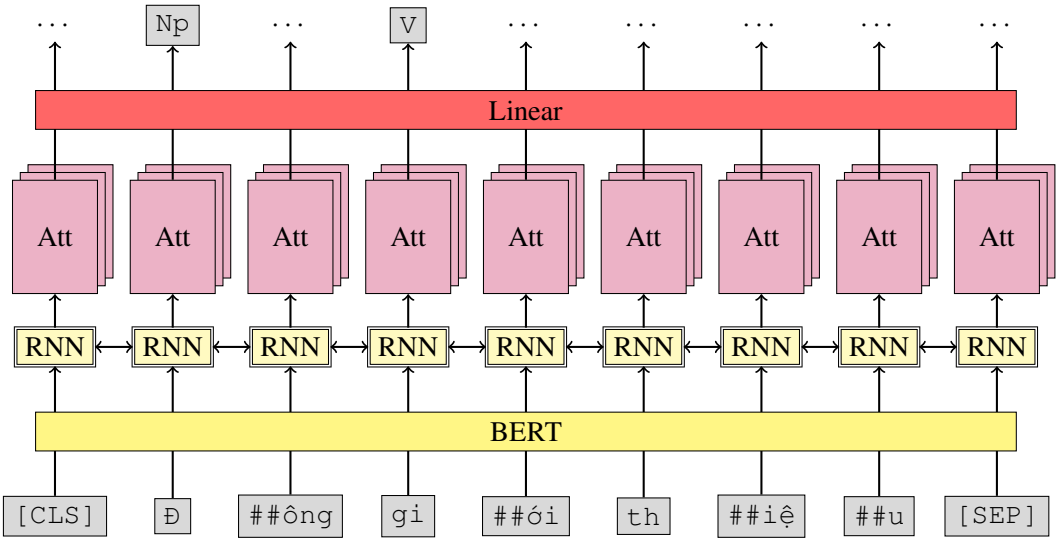


Figure 2: Our proposed end-to-end architecture

0.01.

The output of BERT is computed as follows (Peters et al., 2018):

$$B_k = \gamma \left(w_0 E_k + \sum_{k=1}^{m} w_i h_{ki} \right),$$

where

- B_k is the BERT output of k-th token;
- E_k is the embedding of k-th token;
- m is the number of hidden layers of BERT;
- h_{ki} is the i-th hidden state of of k-th token;
- $\gamma, w_0, w_1, \ldots, w_m$ are trainable parameters.

2.1.2 Proposed Architecture

Our proposed architecture contains five main layers as follows:

1. The input layer encodes a sequence of tokens which are substrings of the input sentence, including ignored indices, padding and separators;
2. A BERT layer;
3. A bidirectional RNN layer with either LSTM or GRU units;
4. An attention layer;
5. A linear layer;

A schematic view of our model architecture is shown in Figure 2.

2.2 ELECTRA

ELECTRA (Clark et al., 2020) is currently the latest development of BERT-based model where a more sample-efficient pre-training method is used. This method is called replaced token detection. In this method, two neural networks, a generator G and a discriminator D, are trained simultaneously. Each one consists of a Transformer network (an encoder) that maps a sequence of input tokens $\vec{x} = [x_1, x_2, \ldots, x_n]$ into a sequence of contextualized vectors $h(\vec{x}) = [h_1, h_2, \ldots, h_n]$. For a given position t where x_t is the masked token, the generator outputs a probability for generating a particular token x_t with a softmax distribution:

$$p_G(x_t|\vec{x}) = \frac{\exp(x_t^\top h_G(\vec{x})_t)}{\sum_u \exp(u_t^\top h_G(\vec{x})_t)}.$$

For a given position t, the discriminator predicts whether the token x_t is "real", i.e., that it comes from the data rather than the generator distribution, with a sigmoid function:

$$D(\vec{x}, t) = \sigma \left(w^\top h_D(\vec{x})_t \right)$$

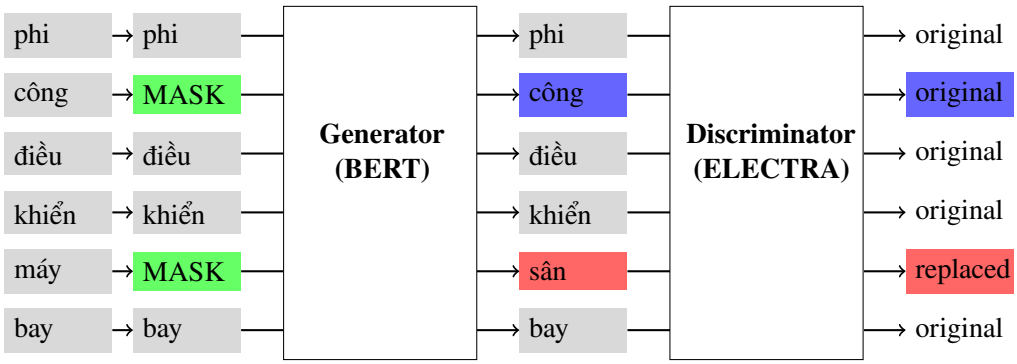


Figure 3: An overview of replaced token detection by the ELECTRA model on a sample drawn from vELECTRA

An overview of the replaced token detection in the ELECTRA model is shown in Figure 3. The generator is a BERT model which is trained jointly with the discriminator. The Vietnamese example is a real one which is sampled from our training corpus.

3 Experiments

3.1 Experimental Settings

3.1.1 Model Training

To train the proposed models, we use a CPU (Intel Xeon E5-2699 v4 @2.20GHz) and a GPU (NVIDIA GeForce GTX 1080 Ti 11G). The hyper-parameters that we chose are as follows: maximum sequence length is 256, BERT learning rate is $2E-05$, learning rate is $1E-3$, number of epochs is 100, batch size is 16, use apex and BERT weight decay is set to 0, the Adam rate is $1E-08$. The configuration of our model is as follows: number of RNN hidden units is 256, one RNN layer, attention hidden dimension is 64, number of attention heads is 3 and a dropout rate of 0.5.

To build the pre-training language model, it is very important to have a good and big dataset. This dataset was collected from online newspapers[2] in Vietnamese. To clean the data, we perform the following pre-processing steps:

- Remove duplicated news
- Only accept valid letters in Vietnamese
- Remove too short sentences (less than 4 words)

We obtained approximately 10GB of texts after collection. This dataset was used to further pre-train the mBERT to build our viBERT which better represents Vietnamese texts. About the vocab, we removed insufficient vocab from mBERT because its vocab contains ones for other languages. This was done by keeping only vocabs existed in the dataset.

In pre-training vELECTRA, we collect more data from two sources:

- NewsCorpus: 27.4 GB[3]
- OscarCorpus: 31.0 GB[4]

Totally, with more than 60GB of texts, we start training different versions of vELECTRA. It is worth noting that pre-training viBERT is much slower than pre-training vELECTRA. For this reason, we pre-trained viBERT on the 10GB corpus rather than on the large 60GB corpus.

3.1.2 Testing and evaluation methods

In performing experiments, for datasets without development sets, we randomly selected 10% for fine-tuning the best parameters.

To evaluate the effectiveness of the models, we use the commonly-used metrics which are proposed by the organizers of VLSP. Specifically, we measure the accuracy score on the POS tagging task which is calculated as follows:

$$Acc = \frac{\#of\_words\_correcly\_tagged}{\#of\_words\_in\_the\_test\_set}$$

[2] vnexpress.net, dantri.com.vn, baomoi.com, zingnews.vn, vitalk.vn, etc.

[3] https://github.com/binhvq/news-corpus

[4] https://traces1.inria.fr/oscar/

No.	VLSP 2010		VLSP 2013			
Existing models						
1.	MEM (Le-Hong et al., 2010)	93.4	RDRPOSTagger (Nguyen et al., 2014)	95.1		
2.			BiLSTM-CNN-CRF (Ma and Hovy, 2016)	95.4		
3.			VnCoreNLP-POS (Nguyen et al., 2017)	95.9		
4.			jointWPD (Nguyen, 2019)	96.0		
5.			PhoBERT_base (Nguyen and Nguyen, 2020)	96.7		
Proposed models						
	Model Name	**mBERT**	**viBERT**	**vELEC**	**mBERT**	**viBERT**	**vELEC**
1.	+Fine-Tune	94.34	95.07	95.35	96.35	96.60	96.62
2.	+BiLSTM	94.34	95.12	95.32	96.38	96.63	**96.77**
3.	+BiGRU	94.37	95.13	95.37	96.45	96.68	96.73
4.	+BiLSTM_Attn	94.37	95.12	**95.40**	96.36	96.61	96.61
5.	+BiGRU_Attn	94.41	95.13	95.35	96.33	96.56	96.55

Table 1: Performance of our proposed models on the POS tagging task

and the F_1 score on the NER task using the following equations:

$$F_1 = 2 * \frac{Pre * Rec}{Pre + Rec}$$

where *Pre* and *Rec* are determined as follows:

$$Pre = \frac{NE\_true}{NE\_sys}$$

$$Rec = \frac{NE\_true}{NE\_ref}$$

where *NE_ref* is the number of NEs in gold data, *NE_sys* is the number of NEs in recognizing system, and *NE_true* is the number of NEs which is correctly recognized by the system.

3.2 Experimental Results

3.2.1 On the PoS Tagging Task

Table 1 shows experimental results using different proposed architectures on the top of mBERT and viBERT and vELECTRA on two benchmark datasets from the campaign VLSP 2010 and VLSP 2013.

As can be seen that, with further pre-training techniques on a Vietnamese dataset, we could significantly improve the performance of the model. On the dataset of VLSP 2010, both viBERT and vELECTRA significantly improved the performance by about 1% in the F_1 scores. On the dataset of VLSP 2013, these two models slightly improved the performance.

From the table, we can also see the performance of different architectures including fine-tuning, BiLSTM, biGRU, and their combination with attention mechanisms. Fine-tuning mBERT with linear functions in several epochs could produce nearly state-of-the-art results. It is also shown that building different architectures on top slightly improve the performance of all mBERT, viBERT and vELECTRA models. On the VLSP 2010, we got the accuracy of 95.40% using biLSTM with attention on top of vELECTRA. On the VLSP 2013 dataset, we got 96.77% in the accuracy scores using only biLSTM on top of vELECTRA.

In comparison to previous work, our proposed model - vELECTRA - outperformed previous ones. It achieved from 1% to 2% higher than existing work using different innovation in deep learning such as CNN, LSTM, and joint learning techniques. Moreover, vELECTRA also gained a slightly better than PhoBERT_base, the same pre-training language model released so far, by nearly 0.1% in the accuracy score.

3.2.2 On the NER Task

Table 2 shows experimental results using different proposed architectures on the top of mBERT, viBERT and vELECTRA on two benchmark datasets from the campaign VLSP 2016 and VLSP 2018.

No.	VLSP 2016				VLSP 2018		
Existing models							
1.	TRE+BI (Le-Hong, 2016)			87.98	VietNER		76.63
2.	BiLSTM_CNN_CRF (Pham and Le-Hong, 2017a)			88.59	ZA-NER		74.70
3.	BiLSTM (Pham and Le-Hong, 2017b)			92.02			
4.	NNVLP (Pham et al., 2017)			92.91			
5.	VnCoreNLP-NER (Vu et al., 2018)			88.6			
6.	VNER (Nguyen, 2019)			89.6			
7.	ETNLP (Vu et al., 2019)			91.1			
8.	PhoBERT_base (Nguyen and Nguyen, 2020)			93.6			
Proposed models							
	Model Name	**mBERT**	**viBERT**	**VELEC**	**mBERT**	**viBERT**	**VELEC**
1.	+Fine-Tune	91.28	92.84	94.00	86.86	88.04	89.79
2.	+BiLSTM	91.03	93.00	93.70	86.62	88.68	89.92
3.	+BiGRU	91.52	93.44	93.93	86.72	88.98	**90.31**
4.	+BiLSTM_Attn	91.23	92.97	**94.07**	87.12	89.12	90.26
5.	+BiGRU_Attn	90.91	93.32	93.27	86.33	88.59	89.94

Table 2: Performance of our proposed models on the NER task. ZA-NER (Luong and Pham, 2018) is the best system of VLSP 2018 (Huyen et al., 2018). VietNER is from (Nguyen et al., 2019)

These results once again gave a strong evidence to the above statement that further training mBERT on a small raw dataset could significantly improve the performance of transformation-based language models on downstream tasks. Training vELECTRA from scratch on a big Vietnamese dataset could further enhance the performance. On two datasets, vELECTRA improve the F_1 score by from 1% to 3% in comparison to viBERT and mBERT.

Looking at the performance of different architectures on top of these pre-trained models, we acknowledged that biLSTM with attention once a gain yielded the SOTA result on VLSP 2016 dataset. On VLSP 2018 dataset, the architecture of biGRU yielded the best performance at 90.31% in the F_1 score.

Comparing to previous work, the best proposed model outperformed all work by a large margin on both datasets.

3.3 Decoding Time

Figure 4 and 5 shows the averaged decoding time measured on one sentence. According to our statistics, the averaged length of one sentence in VLSP 2013 and VLSP 2016 datasets are 22.55 and 21.87 words, respectively.

For the POS tagging task measured on VLSP 2013 dataset, among three models, the fastest decoding time is of vELECTRA model, followed by viBERT model, and finally by mBERT model. This statement holds for four proposed architectures on top of these three models. However, for the fine-tuning technique, the decoding time of mBERT is faster than that of viBERT.

For the NER task measured on the VLSP 2016 dataset, among three models, the slowest time is of viBERT model with more than 2 milliseconds per sentence. The decoding times on mBERT topped with simple fine-tuning techniques, or biGRU, or biLSTM-attention is a little bit faster than on vELECTRA with the same architecture.

This experiment shows that our proposed models are of practical use. In fact, they are currently deployed as a core component of our commercial chatbot engine FPT.AI[5] which is serving effectively many customers. More precisely, the FPT.AI platform has been used by about 70 large enterprises, and of over 27,000 frequent developers, serving more than 30 million end users.[6]

[5] http://fpt.ai/

[6] These numbers are reported as of August, 2020.

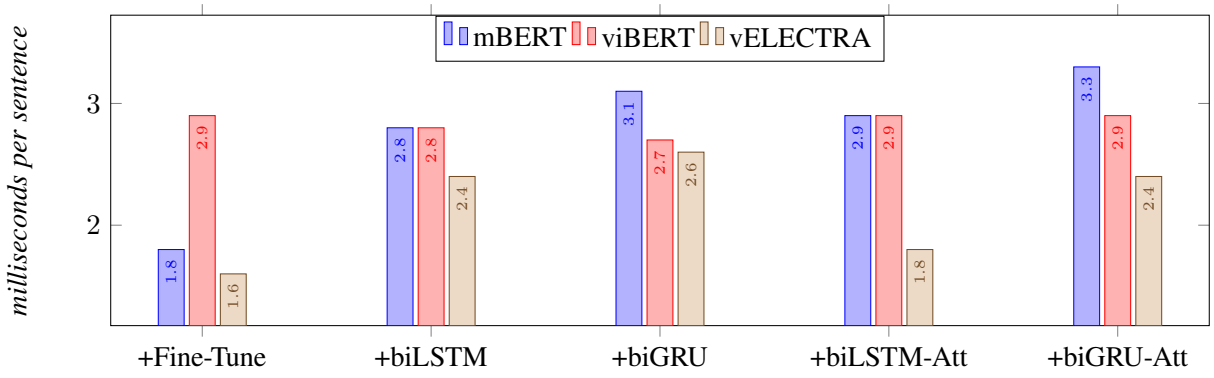


Figure 4: Decoding time on PoS task – VLSP 2013

4 Conclusion

This paper presents some new model architectures for sequence tagging and our experimental results for Vietnamese part-of-speech tagging and named entity recognition. Our proposed model vELECTRA outperforms previous ones. For part-of-speech tagging, it improves about 2% of absolute point in comparison with existing work which use different innovation in deep learning such as CNN, LSTM, or joint learning techniques. For named entity recognition, the vELECTRA outperforms all previous work by a large margin on both VLSP 2016 and VLSP 2018 datasets.

Our code and pre-trained models are published as an open source project for facilitate adoption and further research in the Vietnamese language processing community.[7] An online service of the models for demonstration is also accessible at `https://fpt.ai/nlp/bert/`. A variant and more advanced version of this model is currently deployed as a core component of our commercial chatbot engine FPT.AI which is serving effectively millions of end users. In particular, these models are being fine-tuned to improve task-oriented dialogue in mixed and multiple domains (Luong and Le-Hong, 2019) and dependency parsing (Le-Hong et al., 2015).

Acknowledgement

We thank three anonymous reviewers for their valuable comments for improving our manuscript.

[7] viBERT is available at `https://github.com/fpt-corp/viBERT` and vELECTRA is available at `https://github.com/fpt-corp/vELECTRA`.

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 *Proceedings of ICLR*.

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*, pages 1–16, Minnesota, USA.

Nguyen Thi Minh Huyen, Ngo The Quyen, Vu Xuan Luong, Tran Mai Vu, and Nguyen Thi Thu Hien. 2018. VLSP shared task: Named entity recognition. *Journal of Computer Science and Cybernetics*, 34(4):283–294.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Phuong Le-Hong, Azim Roussanaly, Thi Minh Huyen Nguyen, and Mathias Rossignol. 2010. An empirical studyof maximum entropy approach for part-of-speech tagging of Vietnamese texts. In *Traitement Automatique des Langues Naturelles – TALN, Jul 2010, Montréal, Canada*, pages 1–12.

Phuong Le-Hong, Thi-Minh-Huyen Nguyen, Thi-Luong Nguyen, and My-Linh Ha. 2015. Fast dependency parsing using distributed word representations. In *Trends and Applications in Knowledge Discovery and Data Mining*, volume 9441 of *LNAI*. Springer.

Phuong Le-Hong. 2016. Vietnamese named entity recognition using token regular expressions and bidirectional inference. In *VLSP NER Evaluation Campaign*.

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. In *Preprint*.

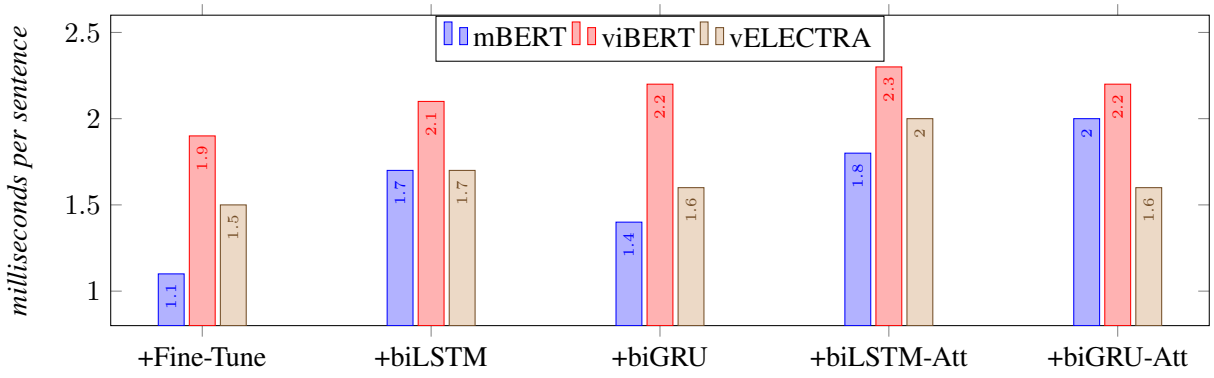


Figure 5: Decoding time on NER task – VLSP 2016

Chi-Tho Luong and Phuong Le-Hong. 2019. Towards task-oriented dialogue in mixed domains. In *Proceedings of the International Conference of the Pacific Association for Computational Linguistics*, pages 267–266. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-15-6168-9_22.

Viet-Thang Luong and Long Kim Pham. 2018. ZA-NER: Vietnamese named entity recognition at VLSP 2018 evaluation campaign. In *In the proceedings of VLSP workshop 2018*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *In Proceedings of ACL*, pages 1064–1074.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *https://arxiv.org/pdf/2003.00744.pdf*.

Dat Quoc Nguyen, Dai Quoc Nguyen, and and Son Bao Pham Dang Duc Pham. 2014. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *In Proceedings of the Demonstrations at EACL*, pages 17–20.

Dat Quoc Nguyen, Thanh Vu, Dai Quoc Nguyen, Mark-Dras, and Mark Johnson. 2017. From word segmentation to POS tagging for Vietnamese. In *In Proceedings of ALTA*, pages 108–113.

Kim Anh Nguyen, Ngan Dong, , and Cam-Tu Nguyen. 2019. Attentive neural network for named entity recognition in Vietnamese. In *In Proceedings of RIVF*.

Dat Quoc Nguyen. 2019. A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing. In *In Proceedings of ALTA*, pages 28–34.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*, pages 1–15, Louisiana, USA.

Thai Hoang Pham and Phuong Le-Hong. 2017a. End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character-level. In *PACLING - Conference of the Pacific Association of Computational Linguistics*, pages 219–232.

Thai Hoang Pham and Phuong Le-Hong. 2017b. The importance of automatic syntactic features in Vietnamese named entity recognition. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31 (2017)*, pages 97–103.

Thai Hoang Pham, Xuan Khoai Pham, Tuan Anh Nguyen, and Phuong Le-Hong. 2017. Nnvlp: A neural network-based Vietnamese language processing toolkit. In *The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017). Demonstration Paper*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

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 *Proceedings of NIPS*.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark-Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *In Proceedings of NAACL: Demonstrations*, pages 56–60.

Xuan-Son Vu, Thanh Vu, Son Tran, and Lili Jiang. 2019. ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *In Proceedings of RANLP*, pages 1285–1294.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS*, pages 5754–5764.

A new look at Pattani Malay Initial Geminates: a statistical and machine learning approach

Francesco Burroni
Department of Linguistics and
Cognitive Science Program
Cornell University
203 Morrill Hall,
Ithaca, NY, USA 14850
fb279@cornell.edu

Sireemas Maspong
Department of Linguistics and
Southeast Asia Program
Cornell University
203 Morrill Hall,
Ithaca, NY, USA 14850
sm2627@cornell.edu

Pittayawat Pittayaporn
Department of Linguistics and
Southeast Asian Linguistics Research Unit
Faculty of Arts, Chulalongkorn University
Phayathai Road, Pathumwan,
Bangkok, Thailand 10330
pittayawat.p@chula.ac.th

Pimthip Kochaiyaphum
Department of Linguistics and
Southeast Asian Linguistics Research Unit
Faculty of Arts, Chulalongkorn University
Phayathai Road, Pathumwan,
Bangkok, Thailand 10330
pimthip.ko@student.ac.th

Abstract

In this paper, we present a statistical and machine learning approach to the acoustic discrimination of a cross-linguistically unusual phonological contrast, initial geminates vs. singletons in Pattani Malay. We show that the only statistically significant difference between geminates and singletons is the duration of the consonant itself. No differences in F0 and intensity were observed on the following vowel, *contra* earlier reports. We further investigated the robustness of this contrast using linear discriminant analysis. Results show that discrimination is above chance, but poor (~62%). The large overlap between the two categories may be partly due to the naturalistic nature of our speech samples. However, we also found that the contrast is neutralized in some minimal pairs. This merger is surprising since initial geminates are often the sole realization of lexical and morphosyntactic contrasts. We suggest that the singleton/initial geminate contrast is now best characterized as a marginal contrast. We hypothesize that this marginally contrastive status may be the result of an on-going sound change, perhaps connected with the more modest role that initial geminates play in Pattani Malay morphophonological alternations.

1 Introduction

Pattani Malay (PM), an Austronesian language spoken in Southern Thailand (Uthai 2011), exhibits a cross-linguistically unusual phonological 'length' contrast for all word-initial consonants, e.g., [matɔ] 'eye' *vs* [mːatɔ] 'jewelry'. The long forms of initial consonants, usually termed initial geminates (IGs), have been reported to differ from singletons along multiple acoustic dimensions. With regards to duration, PM IGs have been reported to be, on average, three times longer than their singleton counterparts (Abramson 1987). Durational differences are hardly a surprising finding since closure duration is usually considered the most reliable acoustic correlate of phonological length cross-linguistically (Ladefoged and Maddieson

1996). If previous work is representative, however, the IG/singleton duration ratio of 3:1 in PM would be on the extreme side of the spectrum (Ladefoged and Maddieson 1996).

Interestingly, duration is not the only cue that distinguishes IGs from singletons in PM. IGs have been reported to produce acoustic effects on the following vowel as well. In particular, previous research has reported that vowels following IGs display longer duration, higher fundamental frequency (F0), and higher intensity (Abramson 1987; Abramson 1998; Phuengnoi 2010). These F0 and intensity cues alone have been shown to be reliable enough for native speakers to correctly identify IGs vs singleton onsets; even in environments where durational cues are ambiguous, such as in absolute utterance-initial position where closure duration cannot be distinguished from preceding silence (Abramson 2003). Similar acoustic features in production and perceptual results have been reported for another closely related variety, Kelantan Malay (Hamzah et al. 2019; Hamzah et al. 2020).

The concomitant manifestation of IGs in the form of local durational differences and of effects on intensity and F0 of the following vowel has led scholars to hypothesize that PM speakers may be in the process of reanalyzing consonantal length as a prosodic contrast based on stress/pitch accent, or that the language may even be on its way to tonogenesis (Abramson 2004).

The possibility that IGs may be the target of ongoing sound change warrants by itself a fresh look at the realization of this unusual phonological contrast. However, we should be cautious in considering previous work the last word on PM IGs. For one thing, previous studies were based on a limited number of speakers (4 for Abramson, 7 for Phuengnoi). Moreover, the difference between IGs and singletons was studied only in words produced in isolation or in words that appeared in a carrier sentence. Finally, in previous studies, speakers were explicitly instructed about the production of the contrast in question. All these factors combined may have led to an exaggeration of the differences between IGs and singletons.

Given such limitations in previous studies, we investigate again the acoustic correlates of IGs in PM by comparing words with and without IGs, but we do so in more ecologically valid speech, which was elicited outside the lab using natural sounding sentences. To characterize the differences between IGs and singletons we make use of both statistical and machine learning techniques.

Statistical analyses showed that IGs are longer than their singleton counterparts, but the difference is much smaller than reported by previous studies. We also found no difference in F0 and intensity on the vowel following IGs vs singletons, *contra* the reports of previous studies.

Additionally, to quantify the robustness of the IG/singleton contrast and to find out which dimensions best discriminate the two categories, we performed classification using linear discriminant analysis (LDA) with a variety of models that employ different combinations of acoustic features. We found that the model performances are above chance, but still poor, peaking at only about 62% accuracy for the best feature combinations.

We speculate that the limited statistical differences and low accuracy of the LDA may be partly due to the naturalistic nature of the speech materials we collected and to ongoing neutralization of the contrast in some minimal pairs. We conclude by discussing several hypotheses concerning the mechanisms that may be at the heart of the observed neutralization.

2 Acoustic Analyses

2.1 Methodology

14 native speakers of PM (6M; 8F) were asked to pronounce 13 disyllabic minimal pairs differing only for their word-initial onsets, which were either geminate or singleton, as shown in Table 1. Stimuli were presented orally with natural-sounding Thai sentences containing the target words. Participants were asked to translate the sentence into PM. Each sentence was repeated six times.

singleton (CVCV)	**gloss**	**geminate (C:VCV)**	**gloss**
pagi	‘morning’	*p:agi*	‘early morning’
paka	‘to use/wear’	*p:aka*	‘usable’
tanɔh	‘land’	*t:anɔh*	‘outside’
dapo	‘kitchen’	*d:apo*	‘at the kitchen’

singleton (CVCV)	gloss	geminate (C:VCV)	gloss
katoʔ	'hammer'	*kːatoʔ*	'frog'
kabo	'Java kapok'	*kːabo*	'beetle'
gaɟɨ	'wage'	*gːaɟɨ*	'saw'
ɟalɛ	'path'	*ɟːalɛ*	'to walk'
ɟuɣi	'to steal'	*ɟːuɣi*	'thief'
misa	'mustache'	*mːisa*	'to grow a moustache'
labɔ	'profit'	*lːabɔ*	'spider'
bulɛ	'moon'	*bːulɛ*	'month'
buŋɔ	'flower'	*bːuŋɔ*	'to bloom'

Table 1. Stimuli

Audio was collected at 44.1 kHz in Praat (Boersma and Weenink 2020). All recordings were made in quiet rooms at the Prince of Songkla University Pattani Campus.

Segmental boundaries were obtained in Praat TextGrids by forced alignment using the Montreal Forced Aligner (McAuliffe et al. 2017). The TextGrids were inspected and manually corrected when necessary. The corrected TextGrids containing segmental boundaries and the audio signals of each word were read back in MATLAB® for analysis.

Eight acoustic measurements were collected:

(1) Duration of initial segments (ms)
(2) Duration of initial syllables (ms)
(3) F0 mean of initial syllables (semitone)
(4) Intensity peak of initial syllables (dB)
(5) F0 mean over initial 10% of vowel following target consonants (semitone)
(6) Intensity mean over initial 10% of vowel following target consonants (dB)
(7) Difference between semitone transformed mean F0 of initial and final syllable
(8) Ratio of mean RMS amplitude of initial to final syllable

F0 was calculated using a MATLAB® implementation of Talkin's robust algorithm for pitch tracking (Talkin 1995) contained in the Voicebox toolbox for MATLAB®. (Brookes 1997). F0 was further processed within all trials and separately by participant by removing all data points with standard deviation scores greater than 2 from the mean; datapoints deviating ±10 Hz from neighboring samples were also excluded. When the F0 vector of a word contained less than 5 datapoints per each syllable, the contour was no longer processed, as interpolation over the entire word would not be reliable. In the other cases, F0 was subsequently interpolated using spline interpolation and smoothed using a median filter. F0 was transformed by converting from Hz to semitones according to the equation $\frac{12}{log_{10}2} \times log_{10}\left(\frac{Hz}{\mu Hz}\right)$ in Zhang (2018).

Sound Pressure Level (SPL) normalized intensity was calculated by transforming the root mean squared intensity of the signal to dB and normalizing to human auditory threshold using the formula $20 \times log_{10}\frac{P}{I0}$. In this formula *P* represents the power of the signal and *P0* represents the normalizing term for the auditory threshold of a 1000 Hz sine wave, equal to 2×10^{-5} (Huang et al. 2001).

Statistical analyses were conducted by fitting linear mixed effect regressions. We compared a model where the fixed effect was the presence/absence of IGs to an intercept-only model. Random effects were subject, word, and position of the word in the phrase (medial or final). Random intercepts were present in the model for each random effect. Random slopes were added when they resulted in a better fit as determined *via* a loglikelihood ratio test. Loglikelihood ratio tests were, thus, used to assess statistical significance and to determine the random effect structure.

2.2 Results

Consonant Duration: Comparing the initial segment in the IG and no IG condition, we found that IGs are significantly longer than singletons ($\chi^2(1) = 4.03$, $p = .04$) with an effect size estimated at 17 ms, as illustrated in Figure 1.

Syllable Duration: The presence of IGs does not significantly affect the duration of the initial syllable ($\chi^2(1) = 1.34$, $p = .24$), as illustrated in Figure 2.

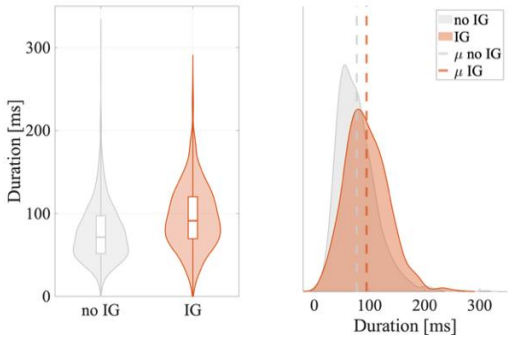


Figure 1. Comparison of initial segment duration (ms)

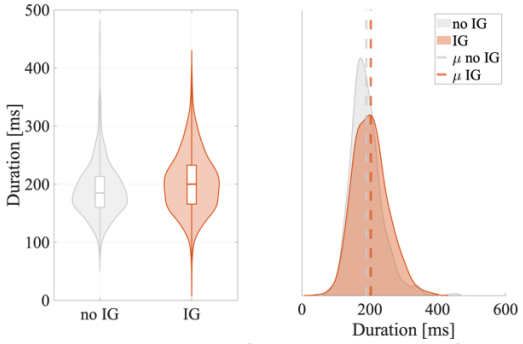


Figure 2. Comparison of syllable duration of initial syllables (ms)

F0: The presence of IGs does not significantly affect the mean F0 of the initial syllable ($\chi^2(1)$ = 0.16, p= .69), as illustrated in Figure 3.

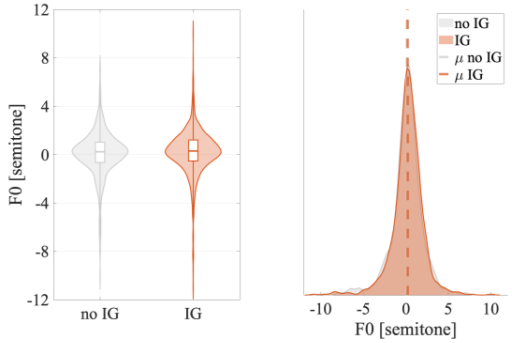


Figure 3. Comparison of mean F0 of initial syllables (semitones)

Intensity: IGs do not significantly affect the maximum SPL normalized intensity of the initial syllable ($\chi^2(1)$ = 0.49, p= .48), as illustrated in Figure 4.

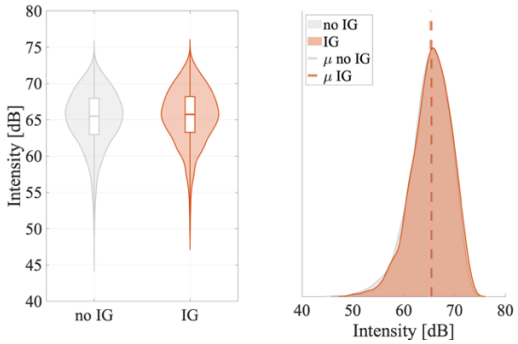


Figure 4. Comparison of maximum SPL normalized intensity of initial syllables (dB)

To further investigate whether the effects of IGs on the following vowel may be limited to the region immediately following the release, we also examined mean F0 and intensity over the first 10% of the vowel, following previous work on Kelantan Malay (Hamzah et al. 2020).

We found no significant differences between mean F0 over the initial 10% of the vowel following IGs vs. singletons ($\chi^2(1)$ = 0.06, p= .79). F0 contours over the vowel are presented in Figure 5.

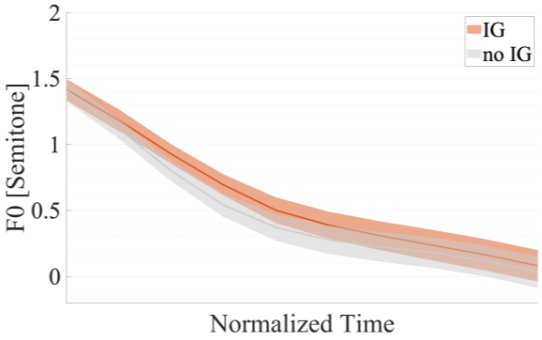


Figure 5. Comparison of time normalized F0 trajectory of initial vowel in semitone. Shaded areas represent ±2 Standard Errors

We also found no significant difference between mean SPL normalized intensity over the initial 10% of a vowel following IGs vs. singletons ($\chi^2(1)$ = 0.95, p= .33). The intensity contours of the following vowel are presented in Figure 6.

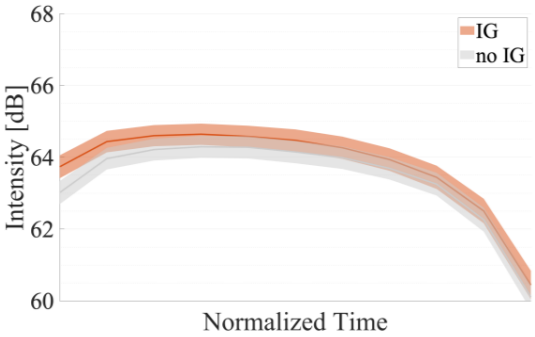


Figure 6. Comparison of time normalized SPL normalized intensity trajectory in dB. Shaded areas represent ±2 Standard Errors

Finally, also following previous work (Abramson 1998; Hamzah et al. 2020), we examined whether differences between IGs and singletons may be manifested more globally in the F0 difference and RMS amplitude ratios of the two syllables. We found no differences for both F0 ($\chi^2(1) = 0.007$, p= .93) and RMS amplitude ($\chi^2(1) = 0.07$, p= .79), as illustrated in Figure 7.

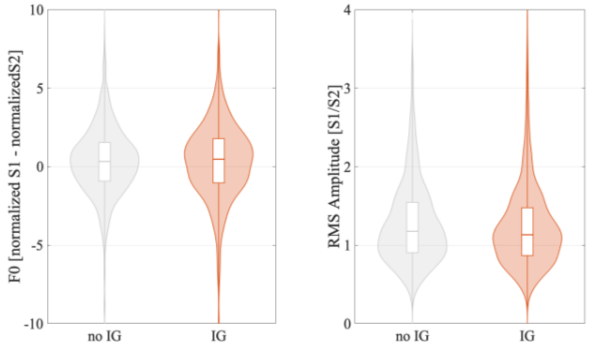


Figure 7. F0 difference and RMS amplitude ratio

2.3 Summary

We found that the durations of IGs and singletons are significantly different, but, unlike in previous studies, the duration of IGs is not three times longer than singletons. The durational differences are estimated at about 17 ms. Furthermore, there is a significant overlap between the two distributions. Contrary to previous descriptions, the presence or absence of IGs does not have a significant effect on syllable duration, mean F0, or peak intensity of the following vowel; no effect is observed even if only 10% of the vowel is examined. We also observed no significant differences in the F0 difference and amplitude ratios of the two syllables.

In sum, we found only very small durational differences between IGs and singletons and the other acoustic measurements do not display significant differences.

3 Linear Discriminant Analysis

To further address the question of whether the singleton/IG contrast in PM is comparable in terms of its magnitude to the singleton/geminate contrast of other languages, we performed classification of IGs vs. singletons using linear discriminant analysis (LDA). In a nutshell, LDA is a classification technique (and also a dimensionality reduction technique) that uses linear combinations of features to maximize the separation between two or more categories. LDA is of interest here because it has been successfully applied to the study of various phonetic contrasts, including geminate vs non-geminate contrasts in both word-medial, in Japanese (Idemaru and Guion-Anderson 2010) and Lebanese Arabic (Khattab and Al-Tamimi 2014), and word-initial position, in Salentino (Burroni and Maspong to appear). We tried to extend this methodology to characterize the word-initial geminate contrast of PM.

3.1 Methodology

We fitted LDA models using cross-validation to evaluate the accuracy of our models. We randomly assigned 80% of the data to a training set and the remaining 20% to a test set. 10,000 such LDA models were fitted for each combination of predictors. The mean accuracy and standard deviations reported here were taken over these 10,000 iterations.

To determine which acoustic dimensions were more apt to discriminate the singleton/IG contrast, we considered that duration of the first segment (CDur) and ratio of the duration of the first segment to the entire word (CDur / WordDur) are the only two statistically significant differences present in our data. We then tested whether adding information concerning the duration (σ_i Dur), mean F0 (σ_i MeanF0), and maximum intensity (σ_i MaxInt) of the target syllable would improve LDA classification. All features were z-scored by participants before performing LDA, as this procedure is known to improve LDA classification.

3.2 Results

We found that the model performance is above chance (that is, above 50%), but still quite poor, as summarized in Table 2, peaking at only about 62% accuracy for the best linear combination of features: the duration of the first segment (CDur) alone or in combination with the duration ratio of the first segment to the entire word (CDur / WordDur).

Model Structure	Mean Accuracy	Standard Deviation
CDur + CDur/WordDur + σ_iDur + σ_iMaxInt + σ_iMeanF0	58.84%	2.18%
CDur + CDur/WordDur + σ_iDur + σ_iMeanF0	58.20%	2.07%
CDur + CDur/WordDur + σ_iDur + σ_iMaxInt	58.88%	2.20%
CDur + CDur/WordDur + σ_iDur	58.19%	2.10%
CDur + CDur/WordDur	**61.40%**	2.06%
CDur/WordDur	59.84%	2.14%
CDur	**62.36%**	2.11%

Table 2. Accuracy of LDA models for different combinations of features

Optimizing the hyperparameters of the model does not greatly improve performance in the identification of IGs as is clear from the confusion matrix of the optimized model presented in Figure 8.

If we inspect the predicted boundary between the two classes, as shown in Figure 9, the reason for the low performance of the model becomes clear: IGs and singletons are not linearly separable in the investigated acoustic dimensions, thus, they cannot be captured by an LDA classifier.

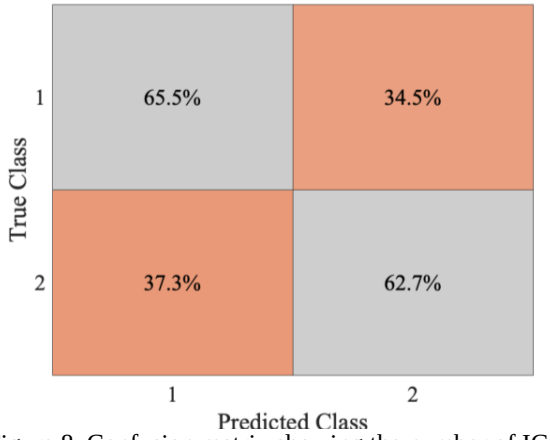


Figure 8. Confusion matrix showing the number of IGs (class 1, top) and singletons (class 2, bottom) classified correctly (gray diagonal) and incorrectly (orange diagonal).

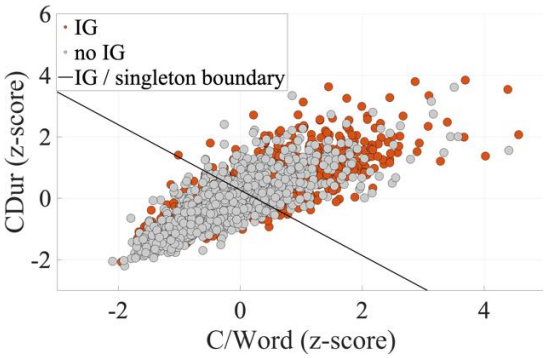


Figure 9. Output of LDA showing large overlap between categories

The low LDA accuracy for geminates contrasts sharply with high accuracy reported for other languages. For instance, for medial geminates in Japanese, accuracy is at ~85-95% (Idemaru and Guion-Anderson 2010) and, for IGs in Salentino, accuracy is at ~80% (Burroni and Maspong to appear).

3.3 Summary

In sum, the discrimination above chance shows that there is indeed a contrast between words with and those without IGs that can be picked up by a simple model, such as an LDA classifier. This is in line with previous phonetic and phonological research on PM and justifies looking for contrasts between words with and without IGs. On the other hand, the low classification accuracy suggests that the contrast is subtle.

We now discuss what factors may be responsible for the observed overlap between IGs and singletons.

4 Discussion

We have three non-mutually exclusive hypotheses to explain why the contrast between IGs and singletons looks much less robust than previously reported.

The first possibility that comes to our mind is that the differences between the result of our study and previous work is due to different methods of data collection. Previous work (Abramson 1987; Abramson 1998) examined IGs only in isolation and in a carrier sentence. Our data, on the other hand, presented IGs and their singleton counterparts in naturalistic sentences. Accordingly, the difference could be due to less carefully articulated speech.

A second possibility is that the contrast may be neutralized for some speakers. The size of our dataset does not allow for a full quantitative assessment of this claim; however, our impression is that almost all speakers produce IGs that are longer than singletons on average, as illustrated in Figure 10.

A third possibility is that the contrast only exists for a subset of minimal pairs. This means that, for many lexical items, the contrast between singletons and IGs is not realized.

Indeed, our data suggests that closure duration of the initial consonants is distinct only for a subset of minimal pairs, as illustrated in Figure 11.

Given this observation, we ask what generalizations may explain the observed neutralizations, as well as the non-neutralizations.

In the framework of Evolutionary Phonology (EP), IGs have been hypothesized to be diachronically unstable (Blevins 2004). Furthermore, EP holds that the stability of phonetic cues to IGs may be related to their wider role in the grammar. IGs survive only in languages where they represent the only cue to lexical contrasts and produce "sentential minimal pairs". In other words, IGs survive only when they compete lexically with singletons and cannot be disambiguated by context (Blevins and Wedel 2009; Burroni and Maspong to appear).

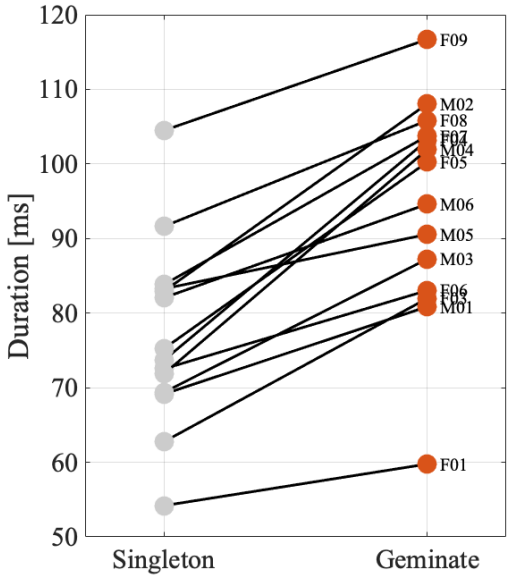


Figure 10. Mean duration of singletons (left) and IGs (right) by speaker (ms)

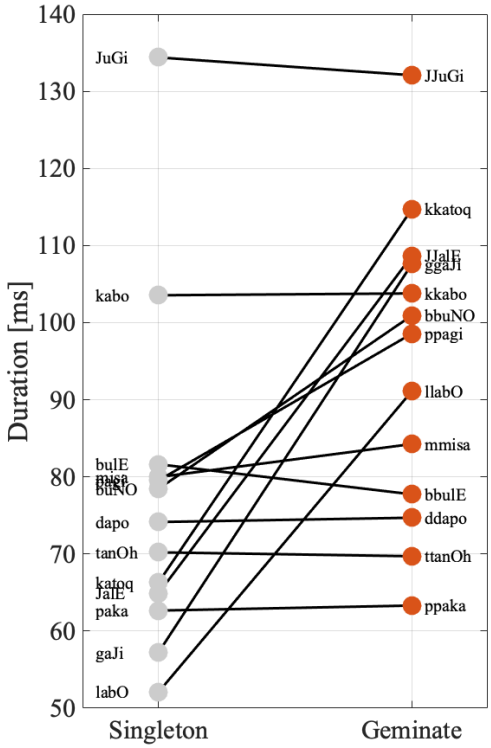


Figure 11. Mean duration of singletons (left) vs IGs (right) by word (ms)

Interestingly, PM seems a counterexample to this generalization, as IGs are being lost in this language, even though they are the unique realization of morphosyntactic contrasts. For instance, under an EP approach, the observed

merger of [dapo] ‘kitchen’ and [dːapo] ‘at the kitchen’ is expected, since these forms appear in different positions and can be disambiguated by context. Similarly, the non-merger of [katoʔ] ‘hammer’ and [kːatoʔ] ‘frog’ is expected since these forms appear in the same context and the IG or lack thereof is the only cue distinguishing them. However, other mergers, such as [kabo] ‘Java kapok (type of plant)’ and [kːabo] ‘beetle (type of bug)’, are not expected, because context does not allow for disambiguation, thus, the neutralizing IG would be one that is a unique cue to the contrast, just like the non-merging one in [katoʔ]/[kːatoʔ]. However, the merger of [kabo]/[kːabo] may suggest some role for word frequency effects. Phillips (2006) explained that retrieving low-frequency word is a challenge for the learner. These difficulties, in turn, may lead to alterations of the phonetic forms of low frequency words on the model of unmarked patterns, that IGs may be altered to singletons. At any rate, for another counterexample to the EP claim that cues to IGs are dependent on lexical competition, we refer the reader to Burroni and Maspong (to appear). Since lexical competition alone does not explain the paradox of IGs merging with singletons in PM, other factors need to be considered.

It has been reported that PM speakers no longer make use of IGs for the purpose of morphological derivation due to contact with Thai (Uthai 1993), accordingly, it is possible that the contrastive phonological status of IGs is being eroded in connection with their reduced ‘functional’ role in the grammar. If IGs and singletons will be merging at evolutionary timescales, the loss of PM IG contrasts would be a striking example of sound change via lexical diffusion connected with a reduced functional load, an information theoretic measurement that has been argued to correlate with geminate to singleton ratio (Tang and Harris 2014) and resistance to merger (Wedel et al. 2013). Further research is necessary to test the merits of these hypotheses on the basis of a larger PM dataset. Corpus frequencies also need to be obtained in order to calculate information theoretic measurements, such as functional load (Surendran and Niyogi 2006).

At any rate, since the contrast between IGs and singletons is only observed for some minimal pairs, it may be best interpreted as a quasi-phonemic or marginal contrast (Hall 2013; Renwick and Ladd 2016). If this interpretation is correct, our acoustic results would align with recent work demonstrating that marginal phonological contrasts may display large overlaps when data is collected outside the lab, in more naturalistic contexts (Cohn and Renwick 2019).

5 Conclusion

In this paper, we have shown that the only significant difference between PM IGs and singletons in naturalistic speech is the duration of the consonants themselves. We have further shown that an LDA model is able to discriminate between syllables with and without IGs slightly above chance level (~62%). This is much below usual LDA performances for geminates in other languages.

The striking difference between our findings and earlier reports regarding the robustness of cues to IGs in PM calls for an explanation. One possibility is that previous experimental work may have exacerbated the difference between IGs and singletons. After all, highly controlled lab speech is very different from less carefully articulated naturalistic speech. IGs in PM could then be an example showing that a more nuanced characterization of phonological contrasts requires an integrated analysis of both laboratory and more naturalistic phonetic data, as advocated by Cohn and Renwick (2019).

However, we have also shown that, although speakers on average produce longer IGs than singletons, they produce the contrast only for a subset of minimal pairs. We have speculated that an appropriate characterization of the subsets that undergo and resist merger will require further collection of information theoretic measurements, such as functional load. One thing is relatively clearer: IGs are moving towards a more marginally contrastive role in the grammar of PM, a fact that may be reflected in their phonetic realization.

Acknowledgements

We would like to thank Santhawat Thanyawong for his help with data collection, Sam Tilsen, and members of the Cornell Phonetics Lab for their feedback. We are grateful to the Southeast Asian Linguistics Research Unit, Faculty of Arts, Chulalongkorn University for financial support.

References

Abramson, Arthur S. 1987. Word-initial consonant length in Pattani Malay. *International Congress of Phonetic Science* (ICPhS) 6, 68–70.

Abramson, Arthur S. 1998. The complex acoustic output of a single articulatory gesture: Pattani Malay word-initial consonant length. *Annual meeting of the Southeast Asian Linguistics Society* (SEALS) 4, 1–20.

Abramson, Arthur S. 2003. Acoustic cues to word-initial stop length in Pattani Malay. *International Congress of Phonetic Science* (ICPhS) 15, 387–390.

Abramson, Arthur S. 2004. Toward prosodic contrast: Suai and Pattani Malay. In *International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages*, 1–4.

Blevins, Juliette. 2004. *Evolutionary Phonology: the emergence of sound patterns*. Cambridge; New York: Cambridge University Press.

Blevins, Juliette & Andrew B. Wedel. 2009. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica* 26(2), 143–183.

Boersma, Paul & David Weenink. 2020. *Praat: doing phonetics by computer*. [Computer software].

Brookes, Mike. 1997. *Voicebox: Speech processing toolbox for* MATLAB. [Computer Software].

Burroni, Francesco & Sireemas Maspong. to appear. Re-examining Initial Geminates: Typology, Evolutionary Phonology, and Phonetics. In *Historical Linguistics 2019*.

Cohn, Abigail C. & Margaret E. L. Renwick. 2019. Doing Phonology in the age of big data. In *Cornell Working Papers in Phonetics and Phonology 2019*, 1–36.

Hall, Kathleen Currie. 2013. A typology of intermediate phonological relationships. *The Linguistic Review* 30(2), 215 – 275.

Hamzah, Mohd Hilmi, J Fletcher & J Hajek. 2019. The secondary roles of amplitude and F0 in the perception of word-initial geminates in Kelantan Malay. *International Congress of Phonetic Science* (ICPhS) 19, 2735-2757.

Hamzah, Mohd Hilmi, John Hajek & Janet Fletcher. 2020. Non-durational acoustic correlates of word-initial consonant gemination in Kelantan Malay: The potential roles of amplitude and f0. *Journal of the International Phonetic Association*. Cambridge University Press 50(1), 23–60.

Huang, Xuedong, Alex Acero, Hsiao-Wuen Hon & Raj Foreword By-Reddy. 2001. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Idemaru, Kaori & Susan Guion-Anderson. 2010. Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica* 67(1–2), 25–46.

Khattab, Ghada & Jalal Al-Tamimi. 2014. Geminate timing in Lebanese Arabic: The relationship between phonetic timing and phonological structure. *Laboratory Phonology* 5(2), 231–269.

Kraehenmann, Astrid. 2011. Initial geminates. *The Blackwell companion to phonology*. Wiley Online Library.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. Oxford; Cambridge, MA: Blackwell Oxford.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. *Montreal Forced Aligner*. [Computer software].

Phillips, Betty S. 2006. *Word frequency and lexical diffusion*. New York: Palgrave Macmillan.

Phuengnoi, Nattaphon. 2010. *An acoustic study of stressed and unstressed syllables in Pattani Malay and Urak Lawoi'*. Bangkok: Chulalongkorn University thesis.

Renwick, Margaret E. L. & D. Robert Ladd. 2016. Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology* 7(1). 1–29.

Surendran, Dinoj & Partha Niyogi. 2006. Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In Ole Nedergaard Thomsen (ed.), *Competing Models of Linguistics Change: Evolution and Beyond. In commemoration of Eugenio Coseriu (1921-2002)*, 43–58. Amsterdam & Philadelphia: Benjamins.

Talkin, David. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*. 495-518. New York, NY: Elsevier Science.

Tang, Kevin & John Harris. 2014. A functional-load account of geminate contrastiveness: a meta-study. In *Linguistics Association of Great Britain 2014*.

Uthai, Ruslan. 1993. *A comparison of word formation in Standard Malay and Pattani Malay*. Bangkok: Chulalongkorn University thesis.

Uthai, Ruslan. 2011. *Keistimewaan dialek Malayu Patani*. Bangi: Penerbit Universiti Kebangsaan Malaysia.

Wedel, Andrew, Abby Kaplan & Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128(2). 179–186.

Zhang, Jingwei. 2018. A Comparison of Tone Normalization Methods for Language Variation Research. *Pacific Asia Conference on Language, Information and Computation* (PACLIC) 32, 823-831.

Sketching the English Translations of Kumārajīva's *The Diamond Sutra*: A Comparison of Individual Translators and Translation Teams

Xi Chen
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Department of English, University of Macau
yb77703@um.edu.mo

Vincent Xian Wang
Department of English
University of Macau
vxwang@um.edu.mo

Chu-Ren Huang
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
The HK PolyU-PKU Research Centre on Chinese Linguistics
churen.huang@polyu.edu.hk

Abstract

This is a corpus-based study of four English translations of Kumarajiva's *The Diamond Sutra* (401/2002). We sketched the four translated English sutras made by both individual translators and translation teams in terms of the profile of their word and sentence use and readability, using a range of corpus tools. Our results reveal that there are major differences between the individual translators and the translation teams in terms of word repertoire, sentence length and readability. The translation teams produced the English Buddhist texts as easy to read and strict with key concept terms to facilitate their missionary work. The individual translators' renditions tend to differ remarkably based on the translators' identities. Our study would shed light on the future research on language studies of English Buddhist texts and the dissemination of Buddhism from East to West through translation.

1 Introduction

Kumārajīva was a monk from Kucha (龜茲 Qiūcí in Chinese), the current Aksu Prefecture in China. He started to translate the Buddhist scriptures into Chinese when he arrived in Chang'an (the present-day Xi'an), China, in 401 CE and the translation activity lasted till his death in 409 CE. With the assistance of his translation team, he translated over 30 sutras containing 313 volumes. Regarding the scope, style, sophistication, popularity and influence, Kumārajīva's translated Buddhist scriptures are often considered best in Chinese history (Cheung, 2014, p. 93; Hung, 2005, p. 80).

The previous studies on Kumārajīva and his team's translation activity are situated in the field of translation history. Ma (1999) and Wang (2006) include their translation activity in the historical research on the translation of Buddhist scriptures. Wang (1984) elaborates Kumārajīva and his team's sophisticated translation process and Siu (2010) depicts their translation institutes in Chang'an. Kumārajīva's translated Buddhist scriptures are regarded as Buddhist classics in China. They were retranslated into English by different translators with the spread of Buddhism from East to West, especially the renowned *The Diamond Sutra*. Being able to access different translation versions presents a rare chance to compare the divergent images of the Buddhist philosophy in the English world. Although they were produced from the same source text, the diversity of these translated

texts would exert uneven influences on the audience varying from the mission of the religion to the study of the philosophy. It is thus of great value to investigate these English translations of *The Diamond Sutra*.

2 Translation Versions

We sorted out the English translations of *The Diamond Sutra* from Chinese (Table 1). The translators can be roughly divided into two groups, namely the individual translators and the translation teams. For the individual translators, they hold different professions like a physician (William Gemmel[1]), professors (Samuel Beal[2] and Daisetz Teitaro Suzuki[3]) and Buddhists (Bhikshu Wai-Tao, Dwight Goddard[4] and Pia Giammasi[5]). The translation teams, on the other hand, were made up of Buddhist monks. It is apparent that they produced the translated English sutras for the international preachment of their temples.

N	First Published Year	Translator	Translation Title	Publisher
1	1864	Samuel Beal	Vajra-chhediká, the "Kin Kong King," or Diamond Sútra	Journal of Royal Asiatic Society
2	1912	William Gemmel	The Diamond Sutra (Chin-kang-ching) or Prajna-paramita	Kegan Paul, Trench, Trübner & Co., Ltd.
3	1935	Bhikshu Wai-Tao and Dwight Goddard	The Diamond Sutra: A Buddhist Scripture	Dwight Goddard
4	1935	Daisetz Teitaro Suzuki	The Kongokyo or Vajracchedika	Eastern Buddhist Society
5	1947	A.F. Price	The Diamond Sutra or The Jewel of Transcendental Wisdom	The Buddhist Society
6	1974	Buddhist Text Translation Society	The Diamond Sutra: A General Explanation of the Vajra prajna Paramita Sutra	Sino-American Buddhist Association
7	2004	Pia Giammasi	Diamond Sutra Explained	Primodia Media
8	2005	Cheng Kuan	The Diamond Prajna-paramita Sutra (The Diamond Sutra)	Vairocana Publishing
9	2009	Chung Tai Translation Committee	The Diamond of Perfect Wisdom Sutra	Chung Tai Chan Monastery
10	2016	Fo Guang Shan International Translation Center	Diamond Prajnaparamita Sutra	Fo Guang Shan International Translation Center

Table 1: English Translations of *The Diamond Sutra* from Chinese

3 Methodology

We focus on four English translations of *The Diamond Sutra* from Chinese for this paper, namely Gemmel (1912), Hsuan (2002)[6], Giammasi (2004) and Chung Tai Translation Committee (2009) (Table 2). These texts were selected for three reasons. First, they explicitly state in the texts that their translations were rendered from the Chinese version of Kumārajīva's *The Diamond Sutra*. Second, they are still in circulation today. The translations made by Gemmel and Giammasi are still reprinted and sold on Amazon. The other two are distributed to the believers and disciples of their temples. Third, TT1 and TT2 were produced by the individual translators; TT3 and TT4 were made by the Buddhist translation teams. These four texts form the comparison groups as the Table 2 shows. We built a corpus of these four translated English sutras after digitalizing them for further analysis.

Source Text	ST	金剛般若波羅蜜經 *Vajracchedikā Prajñāpāramitā Sūtra*
Group 1 (Individual Translators)	TT1	*The Diamond Sutra (Chin-Kang-Ching) or Prajna-Paramita* translated by William Gemmel in 1912
	TT2	*Diamond Sutra Explained* translated by Pia Giammasi in 2004

1 Mattoon (2010)
2 Ockerbloom (n.d.)
3 Abe (1986)
4 Wai-Tao and Goddard (1935)
5 Giammasi (2004)

6 Hsuan (2002) is the second edition of the translation made by Buddhist Text Translation Society in 1974. The Buddhist Master Hsuan Hua was put in the position of author. This kind of arrangement follows the tradition of Buddhist translation activity that the translated Buddhist scripture is authored by the Buddhist Master who chaired the translation activity. That Buddhist Master is called 主譯 zhǔ yì (Master Translator) in Chinese (Wang, 1984).

Group 2 (Translation Teams)	TT3	*The Vajra Prajna Paramita Sutra: A General Explanation* translated by Buddhist Text Translation Society in 2002
	TT4	*The Diamond of Perfect Wisdom Sutra* translated by Chung Tai Translation Committee in 2009

Table 2: The Selected Four English translations of Kumārajīva's *The Diamond Sutra*

In order to sketch the profile of these English translations, we adopted four corpus tools to compare the texts in three dimensions: word, sentence and readability. The tools employed and their corresponding functions are listed in Table 3.

Tools	Functions
WordSmith 8.0	STTR, Mean Word Length, Sentences, Sentence Length
BFSU HugeMind Readability Analyzer 2.0	Readability Tests
NVivo 12 Plus	Word List, Word Cloud
AntConc 3.5.7	Collocation

Table 3: Corpus Tools

By describing the four selected English translations of Kumārajīva's *The Diamond Sutra* with the assistant of corpus tools, this study aims to answer the following two research questions.

In terms of the profile of word, sentence and readability:

1. What differences exhibit between the individual translators and the translation teams, if there are?
2. Are there any differences within each group – i.e. the group of individual translators and the group of translation teams (cf. Table 2)?

4 Results

4.1 STTR, Words and Sentences

We used the WordSmith Tools 8.0 (Scott, 2020) to examine the lexical complexity and sentential patterns of the four translations regarding the STTR, word length, number of sentences and sentence length (Table 4). As the text size of TT1 (7,068 tokens) is much larger than the other three. The standard type-token ratio (STTR), which calculates the type-token ratio (TTR) on every 1,000 words, is adopted here as one indicator to compare the lexical diversity of these four texts. The STTR of Group 1 (TT1: 30.20%; TT2: 28.94%) is notably higher than Group 2 (TT3: 25.36%; TT4: 27.44%), which suggests the individual translators employ a wider range of vocabulary than the translation teams. With respect to the mean word length in words, the varieties in each group do not show the same tendency (TT1: 5.03; TT2: 4.66; TT3: 4.52; TT4: 4.76). At the sentential level, the individual translators (TT1: 333; TT2: 322) used fewer sentences than the translation teams (TT3: 359; TT4: 377). But the average length of the former (TT1: 21.23; TT2: 16.11) is greater than the latter (TT3: 15.17; TT4: 13.90).

Indicators	TT1	TT2	TT3	TT4
Tokens	7,068	5,189	5,447	5,243
Types	1,059	761	619	693
TTR	14.98%	14.67%	11.36%	13.22%
STTR	30.20%	28.94%	25.36%	27.44%
STTR std.dev	58.51	57.80	60.89	59.09
STTR basis	1,000	1,000	1,000	1,000
Mean Word Length (characters)	5.03	4.66	4.52	4.76
Word Length std.dev	2.71	2.53	2.63	2.60
Sentences	333	322	359	377
mean (in words)	21.23	16.11	15.17	13.90
std.dev.	14.97	12.51	12.65	11.52

Table 4: WordSmith Tools 8.0 Statistics List

4.2 Frequent Words and Collocations

The "word frequency query" function of NVivo 12 Plus (QSR International Pty Ltd, 2020) generated the content-word lists with word frequencies for each text and visualized the content words with word clouds. We set the query criteria as "display the 100 most frequent words with minimum length of 3 letters". The full top 100 frequent word lists of each file with their counts and weighted percentages are placed in the Appendices (A-D). To concisely illustrate the main differences of the four translations in terms of the frequently-occurring words, four word clouds of the 30 most frequent words are presented below (Figures 1-4). Visually, a considerable number of the highly frequently-used words in the four texts differ, although the main characters "Subhuti" and "Buddha" are consistently on the top of the lists. These differences largely result from the translators' varied renditions of some repetitive key terms in *The Diamond Sutra*, for example, "如來", "法" and "阿耨多羅三藐三菩提" (Table 5).

First, "如來 rú lái" is the honorific title of Buddha. Its literal translation from Sanskrit is

"Tathāgata" (Ding, 2016). TT1 does not distinguish it from "佛 fó" (Buddha) that Gemmel translated it as "Lord Buddha" too. The other three follow the literal translation "Tathagata", which is different from "Buddha". Second, "法 fǎ" is a key concept in Buddhism and it has multiple meanings. It can refer to the universe's truth or law (Ding, 2016). TT1 substitutes it with "Law", while the other three adopts the literal translation "D/dharma". Third, "阿耨多羅三藐三菩提 ā nòu duō luó sān miǎo sān pú tí" is the transliteration of Sanskrit "anuttara-samyak-sambodhi", which was also translated into Chinese as "無上正等正覺 wú shàng zhèng děng zhèng jué" (supreme perfect enlightenment). It represents the highest wisdom of all truth in Buddhism (Ding, 2016). TT2 and TT3 retain the transliteration. TT1 substitutes it with "supreme spiritual wisdom", and TT4 literally translates it while providing the transliteration at the first time.

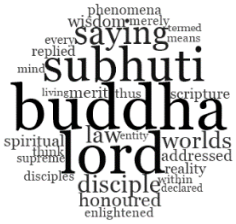


Figure 1: Word Cloud of TT1

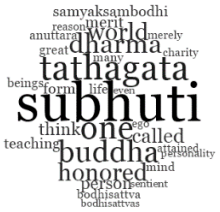


Figure 2: Word Cloud of TT2

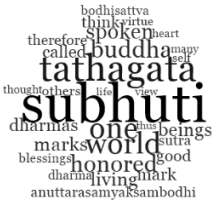


Figure 3: Word Cloud of TT3

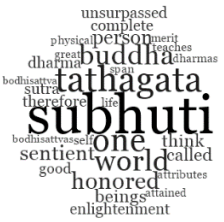


Figure 4: Word Cloud of TT4

Terms	TT1	TT2	TT3	TT4
如來 rú lái	Lord Buddha	Tathagata	Tathagata	Tathagata
法 fǎ	Law	Dharma	dharma	dharma
阿耨多羅三藐三菩提 ā nòu duō luó sān miǎo sān pú tí	supreme spiritual wisdom	anuttara-samyaksambodhi	Anuttarasamyaksambodhi	unsurpassed complete enlightenment (anuttara-samyak-sambodhi)

Table 5: Translations of Key Terms

As *The Diamond Sutra* is the dialogue between Buddha and his disciple Subhuti, we further explore the verbs collocated with these two characters by virtue of AntConc 3.5.7 (Anthony, 2018). We set the span from 3L to 3R and the collocate measure as MI + Log-likelihood ($p > 0.05$). We list the frequently collocated verbs with high statistical scores in Tables 6-7. It can be seen that TT1 has the varied verbs (addressed, declared, enquired, etc.) collocated with Buddha and Subhuti, while TT2, TT3 and TT4 use the simple verbs, such as "said", "told" and "called".

Texts	Collocate	Stat.	Freq	Freq(L)	Freq(R)
TT1	saying	4.10404	46	1	45
	addressed	5.16544	37	3	34
	declared	4.68002	15	0	15
	enquired	4.66294	12	0	12
TT2	said	5.85739	14	4	10
	says	4.35936	6	1	5
TT3	said	5.98352	23	12	11
	told	6.62989	7	0	7
TT4	said	5.62832	22	6	16

Table 6: Buddha's Collocated Verbs

Texts	Collocate	Stat.	Freq	Freq(L)	Freq(R)
TT1	saying	5.33639	79	2	77
	addressed	5.57804	36	33	3
	replied	5.47073	28	4	24
	enquired	5.61757	17	12	5
TT2	said	4.92450	15	10	5
	called	3.85050	15	13	2
	replied	5.14001	11	0	11
TT3	said	4.78715	25	4	21

	called	3.95064	14	14	0
TT4	told	5.31321	7	7	0
	said	4.92734	29	13	16
	called	3.95688	14	12	2

Table 7: Subhuti's Collocated Verbs

4.3 Readability Tests

Finally, we tested the readability of these four translations via the BFSU HugeMind Readability Analyzer 2.0. It is a corpus tool developed by the FLERIC team of Beijing Foreign Studies University (http://corpus.bfsu.edu.cn/TOOLS.htm). It can do six different readability tests for the texts. The calculation formulae are listed in the Appendix E. Apart from the Flesch Reading Ease test, the higher the score is, the less understandable the text is (Coleman & Liau, 1975; Flesch, 1981; Gunning, 1952; Kincaid, Fishburne, Rogers, & Chissom, 1975; Mclaughlin, 1969; Smith & Senter, 1967). The scores of the four texts are listed in Table 8. The results infer that Group 1 (TT1 and TT2) are generally weaker than Group 2 (TT3 and TT4) in readability except that the score of TT2 is a little lower than TT4 in Gunning Fog Index test. Especially, the score of TT1 has marked gap between it and the other three texts.

Tests	**TT1**	**TT2**	**TT3**	**TT4**
Automated Readability Index	13.04	9.87	8.28	9.02
Coleman–Liau Index	13.42	11.36	9.91	11.4
Flesch Reading Ease	39.47	51.31	61.7	52.53
Flesch–Kincaid Readability Test	12.67	10.13	8.42	9.41
Gunning Fog Index	39.71	31.11	25.72	31.51
SMOG (Simple Measure of Gobbledygook)	24.73	20.30	17.95	19.49

Table 8: Readability Tests by BFSU HugeMind Readability Analyzer 2.0

5 Individual Translators versus Translation Teams

We have compared these four English translations of Kumārajīva's *The Diamond Sutra* using a range of corpus tools. In response to our research questions, results show that there are major differences between the individual translators and the translation teams. The individual translators employed a wider range of vocabulary than the translation teams. The former is inclined to use fewer but longer sentences than the latter. In terms of readability, the translated Buddhist texts made by the translation teams are easier to read than the ones rendered by the individual translators. For the inner group comparison, the two translations done by teamwork appear to be consistent with each other in our corpus-based sketching except for the rendition of "阿耨多羅三藐三菩提 ā nòu duō luó sān miǎo sān pú tí". TT4 provides both the literal translation and transliteration, and this seems to be a strategy to stay faithful to the original while facilitate the readers' reading. Although TT1 and TT2 show some similar tendencies as both belong to the group of individual translators, they markedly differ in word diction, i.e. the frequent words and verb collocations. TT1's readability tests scores are much distinct from the other three. This can be explained from the identities of these translators. William Gemmel was a physician, who had a "lifelong interest in history and archeology" (Mattoon, 2010). Although Pia Giammasi is an individual translator, she is the disciple of the Buddhist Master Nan Huai-Chin (Giammasi, 2005). Such a Buddhist background would situate herself in line with the Buddhist groups, which is revealed by her choices of the frequently used words. Therefore, William Gemmel as a translator outside the religious circle of Buddhism tends to enjoy the

highest subjectivity when translating the Buddhist sutra. Our description of the four texts by corpus tools can also offer a glimpse of the mechanism behind each group's translating practice on the Buddhist sutras. The Buddhist translation teams employ sterile and plain words and shorter sentences to reduce the difficulty of the Buddhist texts availing the preachment. They are strict with the key terms and concepts, which maintain a high level of faithfulness to the original text. The individual translators vary owing to their identities.

6 Conclusion

In this paper, we have compared four different English translations of Kumārajīva's *The Diamond Sutra* with corpus tools and demonstrated the differences between the individual translators and translation teams. Our study showed that the two groups clearly differ from each other in terms of the profile of the words and sentences they used and also in readability. The translated sutras rendered by the translation teams tend to be easy for reading while rigorous with the expressions of key concepts. The individual translators performed differently based on their own identities. However, we only compared four texts and did not involve the textual analysis of the Chinese source text as both the classic Chinese and religious language of Buddhist sutras are not supported by the mainstream corpus tools. That is the area in which subsequent studies can work on in the domain of this special textual genre (cf. Lee & Wong, 2016; Wong & Lee, 2018).

Acknowledgements

We are thankful to the four anonymous reviewers of this article for their valuable comments and suggestions, especially one reviewer providing us the resources for future research. The authors would like to acknowledge the support of the research projects "A Comparative Study of Synaesthesia Use in Food Descriptions between Chinese and English" (G-SB1U) of The Hong Kong Polytechnic University and MYRG2018-00174-FAH of the University of Macau.

References

Abe, M. (ed.). (1986). *A Zen life: D.T. Suzuki remembered*. New York and Tokyo: Weatherhill.

Anthony, L. (2018). AntConc (Version 3.5.7) [Software]. Available from https://www.laurenceanthony.net/software

Beal, S. (1864). Art. I.—Vajra-chhediká, the "Kin Kong King," or Diamond Sútra. *Journal of the Royal Asiatic Society*, *1*(1), 1-24. https://doi.org/10.1017/S0035869X00160800

Cheng, K. (trans.). (2005). *The Diamond Prajna-paramita Sutra (The Diamond Sutra).* Taipei: Vairocana Publishing.

Cheung, M. P. Y. (ed.). (2014). *An anthology of Chinese discourse on translation volume 1: From earliest times to the Buddhist project.* Oxon: Routledge.

Chung Tai Translation Committee. (trans.). (2009). *The Diamond of Perfect Wisdom Sutra.* Nantou: Chung Tai Chan Monastery.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283–284. https://doi.org/10.1037/h0076540

Ding, F. (ed.). (2016). *Dictionary of Buddhist studies*. Taipei: Chinese Buddhist Electronic Text Association.

Flesch, R. F. (1981). *How to write plain English.* New York, NY: Barnes & Noble

Fo Guang Shan International Translation Center. (trans.) (2016). *Diamond Prajnaparamita Sutra.* Hacienda Heights, CA: Fo Guang Shan International Translation Center.

Gemmel, W. (trans.). (1912). *The Diamond Sutra (Chin-kang-ching) or Prajna-paramita*. London: Kegan Paul, Trench, Trübner & Co., Ltd..

Giammasi, P. (trans.). (2004). *Diamond Sutra explained.* Florham Park, NJ: Primodia Media.

Gunning, R. (1952). *The technique of clear writing.* New York, NY: McGraw-Hill.

Hsuan, H. (trans.). (1974). *The Diamond Sutra: A general explanation of the Vajra Prajna Paramita Sutra*. San Francisco, CA: Sino-American Buddhist Association.

Hsuan, H. (trans.). (2002). *The Vajra Prajna Paramita Sutra: A general explanation.* Burlingame, CA: Buddhist Text Translation Society.

Hung, E. (2005). *Rewriting Chinese translation history*. Hong Kong: Research Centre for

Translation, The Chinese University of Hong Kong.
Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease formula) for navy enlisted personnel. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.
Kumārajīva. (trans.). (401/2002). *Vajracchedikā Prajñāpāramitā Sūtra*. Taipei: Chinese Buddhist Electronic Text Association. Retrieved from http://buddhism.lib.ntu.edu.tw/BDLM/sutra/chi_pdf/sutra3/T08n0235.pdf
Lee, J., & Wong, T. (2016). Conversational network in the Chinese Buddhist canon. *Open Linguistics, 2*(1), 427–436. doi: 10.1515/opli-2016-0022
Ma, Z. (1999). *A history of translation in China*. Wuhan: Hubei Education Press.
Mattoon, N. (2010). Dancing with death: A Scottish doctor's macabre obsession. Retrieved from http://www.booktryst.com/2010/10/dancing-with-death-scottish-doctors.html
Mclaughlin, G.H. (1969). SMOG Grading - A new readability formula. J*ournal of Reading, 12*(8), 639–646.
Ockerbloom, J. M. (Ed.). (n.d.). Online books by Samuel Beal. Retrieved from http://onlinebooks.library.upenn.edu/webbin/book/lookupname?key=Beal%2C%20Samuel%2C%201825%2D1889
Price, A. F. (trans.). (1947). *The Diamond Sutra or The Jewel of Transcendental Wisdom*. London: The Buddhist Society.
QSR International Pty Ltd. (2020). NVivo (Version 12) [Software]. Available from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home
Scott, M. (2020). WordSmith Tools (Version 8) [Software]. Available from https://lexically.net/wordsmith/downloads/
Siu, S. (2010). *Kumārajīva's translation team in Chang'an*. Kaohsiung: Fo Guang Publications.
Smith, E. A., & Senter, R. J. (1967). Automated Readability Index. *AMRL-TR*, 1–14.
Suzuki, D. T. (1935). *Manual of Zen Buddhism*. Kyoto: Eastern Buddhist Society.
Wai-Tao, B., & Goddard, D. (trans.). (1935). *The Diamond Sutra: A Buddhist scripture.* Santa Barbara, CA: Dwight Goddard.
Wang, T. (2006). *Translation history of Chinese Buddhist scriptures*. Beijing: Central Compilation & Translation Press.
Wang, W. (1984). *Fodian hanyi zhi yanjiu* [A study of Chinese translation of Buddhist scriptures]. Taipei: Heavenly Lotus Publishing.
Wong, T., & Lee, J. (2018). Vernacularization in medieval Chinese: A quantitative study on classifiers, demonstratives, and copulae in the Chinese Buddhist canon. *Digital Scholarship in the Humanities, 34*(1), 64-81. doi: 10.1093/llc/fqy012

Appendices

Appendix A. Top 100 Frequent Word List of TT1

Word	Length	Count	Weighted Percentage (%)
buddha	6	197	5.21
lord	4	179	4.74
subhuti	7	144	3.81
saying	6	96	2.54
disciple	8	66	1.75
law	3	62	1.64
worlds	6	62	1.64
honoured	8	48	1.27
wisdom	6	40	1.06
merit	5	38	1.01
spiritual	9	38	1.01
addressed	9	37	0.98
scripture	9	34	0.90
reality	7	33	0.87
phenomena	9	31	0.82
replied	7	31	0.82
thus	4	30	0.79
think	5	29	0.77
disciples	9	28	0.74
enlightened	11	28	0.74
merely	6	28	0.74
every	5	26	0.69
supreme	7	26	0.69

within	6	25	0.66
mind	4	24	0.64
means	5	23	0.61
declared	8	21	0.56
entity	6	21	0.56
living	6	20	0.53
termed	6	19	0.50
body	4	18	0.48
enquired	8	17	0.45
life	4	17	0.45
personality	11	17	0.45
ideas	5	16	0.42
may	3	16	0.42
man	3	15	0.40
whether	7	15	0.40
charity	7	14	0.37
exercise	8	14	0.37
good	4	14	0.37
therefore	9	14	0.37
obtained	8	13	0.34
perceived	9	13	0.34
sentient	8	13	0.34
woman	5	13	0.34
ages	4	12	0.32
beings	6	12	0.32
doctrine	8	12	0.32
dust	4	12	0.32
faith	5	12	0.32
minds	5	12	0.32
numerous	8	12	0.32
physical	8	12	0.32
arbitrary	9	11	0.29
attained	8	11	0.29
considerable	12	11	0.29
distinctions	12	11	0.29
referred	8	11	0.29
unto	4	11	0.29
bodily	6	10	0.26
buddhist	8	10	0.26
eye	3	10	0.26
form	4	10	0.26
great	5	10	0.26
neither	7	10	0.26
thought	7	10	0.26
entirely	8	9	0.24
ganges	6	9	0.24
idea	4	9	0.24
meaning	7	9	0.24
system	6	9	0.24
thereupon	9	9	0.24
become	6	8	0.21
condition	9	8	0.21
diligently	10	8	0.21
future	6	8	0.21
innumerable	11	8	0.21
material	8	8	0.21
others	6	8	0.21
particles	9	8	0.21
truly	5	8	0.21
arhat	5	7	0.19
bring	5	7	0.19
buddhic	7	7	0.19
continuing	10	7	0.19
dipankara	9	7	0.19
grains	6	7	0.19
greater	7	7	0.19
kingdoms	8	7	0.19
oblivious	9	7	0.19
occasion	8	7	0.19
one	3	7	0.19
paramita	8	7	0.19
proclaimed	10	7	0.19
qualities	9	7	0.19
realise	7	7	0.19
rigorously	10	7	0.19
sand	4	7	0.19
thirty	6	7	0.19

Appendix B. Top 100 Frequent Word List of TT2

Word	Length	Count	Weighted Percentage (%)
subhuti	7	135	5.30
tathagata	9	83	3.26
one	3	75	2.94
buddha	6	66	2.59

dharma	6	58	2.28	true	4	9	0.35
honored	7	50	1.96	two	3	9	0.35
world	5	50	1.96	without	7	9	0.35
called	6	40	1.57	beyond	6	8	0.31
think	5	33	1.30	body	4	8	0.31
person	6	31	1.22	calls	5	8	0.31
merit	5	28	1.10	dust	4	8	0.31
form	4	26	1.02	galaxies	8	8	0.31
samyaksambodhi	14	25	0.98	good	4	8	0.31
teaching	8	25	0.98	know	4	8	0.31
life	4	22	0.86	marks	5	8	0.31
great	5	20	0.79	men	3	8	0.31
anuttara	8	19	0.75	notion	6	8	0.31
beings	6	19	0.75	receive	7	8	0.31
bodhisattva	11	18	0.71	appearance	10	7	0.27
many	4	18	0.71	dipankara	9	7	0.27
mind	4	18	0.71	dwelling	8	7	0.27
attained	8	16	0.63	expounded	9	7	0.27
charity	7	16	0.63	majestic	8	7	0.27
merely	6	16	0.63	means	5	7	0.27
ego	3	15	0.59	must	4	7	0.27
reason	6	15	0.59	neither	7	7	0.27
even	4	14	0.55	paramita	8	7	0.27
personality	11	14	0.55	practice	8	7	0.27
bodhisattvas	12	13	0.51	read	4	7	0.27
sentient	8	13	0.51	realization	11	7	0.27
thought	7	13	0.51	self	4	7	0.27
fortune	7	12	0.47	someone	7	7	0.27
replied	7	12	0.47	thirty	6	7	0.27
sand	4	12	0.47	understand	10	7	0.27
eyes	4	11	0.43	woman	5	7	0.27
minds	5	11	0.43	worlds	6	7	0.27
others	6	11	0.43	anutara	7	6	0.24
virtuous	8	11	0.43	attain	6	6	0.24
ganges	6	10	0.39	buddhas	7	6	0.24
man	3	10	0.39	expound	7	6	0.24
perceived	9	10	0.39	just	4	6	0.24
sutra	5	10	0.39	lands	5	6	0.24
dwell	5	9	0.35	meaning	7	6	0.24
grains	6	9	0.35	notions	7	6	0.24
retain	6	9	0.35	past	4	6	0.24
rupakaya	8	9	0.35	perfect	7	6	0.24
thus	4	9	0.35	real	4	6	0.24

really	6	6	0.24
seven	5	6	0.24
speaks	6	6	0.24
still	5	6	0.24
therefore	9	6	0.24
time	4	6	0.24
treasures	9	6	0.24
universe	8	6	0.24
way	3	6	0.24
women	5	6	0.24

Appendix C. Top 100 Frequent Word List of TT3

Word	Length	Count	Weighted Percentage (%)
subhuti	7	137	5.20
tathagata	9	91	3.45
one	3	71	2.69
world	5	70	2.66
buddha	6	55	2.09
honored	7	53	2.01
spoken	6	48	1.82
marks	5	43	1.63
living	6	42	1.59
beings	6	41	1.56
dharmas	7	38	1.44
called	6	36	1.37
mark	4	36	1.37
good	4	34	1.29
think	5	34	1.29
therefore	9	30	1.14
anuttarasamyaksambodhi	22	29	1.10
sutra	5	29	1.10
others	6	28	1.06
blessings	9	26	0.99
bodhisattva	11	26	0.99
dharma	6	25	0.95
self	4	23	0.87
virtue	6	23	0.87
heart	5	22	0.83
many	4	21	0.80
thought	7	21	0.80
view	4	21	0.80
life	4	20	0.76
thus	4	19	0.72
hold	4	16	0.61
man	3	16	0.61
person	6	16	0.61
receive	7	16	0.61
actually	8	15	0.57
thousand	8	15	0.57
ganges	6	14	0.53
people	6	14	0.53
systems	7	14	0.53
dwell	5	13	0.49
means	5	13	0.49
woman	5	13	0.49
perfection	10	12	0.46
sand	4	12	0.46
someone	7	12	0.46
dust	4	11	0.42
eye	3	11	0.42
grains	6	11	0.42
great	5	11	0.42
paramita	8	11	0.42
reason	6	11	0.42
without	7	11	0.42
across	6	10	0.38
know	4	10	0.38
motes	5	10	0.38
produce	7	10	0.38
two	3	10	0.38
understand	10	10	0.38
body	4	9	0.34
foremost	8	9	0.34
physical	8	9	0.34
seen	4	9	0.34
speak	5	9	0.34
thirty	6	9	0.34
attain	6	8	0.30
attained	8	8	0.30
even	4	8	0.30
extinction	10	8	0.30
four	4	8	0.30
give	4	8	0.30
gives	5	8	0.30

hear	4	8	0.30
merit	5	8	0.30
obtained	8	8	0.30
read	4	8	0.30
recite	6	8	0.30
river	5	8	0.30
three	5	8	0.30
believe	7	7	0.27
big	3	7	0.27
buddhas	7	7	0.27
burning	7	7	0.27
completely	10	7	0.27
dwelling	8	7	0.27
fine	4	7	0.27
form	4	7	0.27
forms	5	7	0.27
future	6	7	0.27
lamp	4	7	0.27
like	4	7	0.27
might	5	7	0.27
see	3	7	0.27
taken	5	7	0.27
told	4	7	0.27
adornment	9	6	0.23
bodhisattvas	12	6	0.23
buddhalands	11	6	0.23
devoid	6	6	0.23
gems	4	6	0.23
gift	4	6	0.23

Appendix D. Top 100 Frequent Word List of TT4

Word	Length	Count	Weighted Percentage (%)
subhuti	7	135	5.10
tathagata	9	82	3.10
one	3	76	2.87
world	5	66	2.49
buddha	6	63	2.38
honored	7	51	1.93
sentient	8	44	1.66
person	6	41	1.55
beings	6	36	1.36
think	5	36	1.36
called	6	35	1.32
therefore	9	30	1.13
complete	8	29	1.10
dharma	6	29	1.10
enlightenment	13	29	1.10
sutra	5	29	1.10
unsurpassed	11	29	1.10
good	4	28	1.06
self	4	22	0.83
life	4	20	0.76
teaches	7	20	0.76
attributes	10	19	0.72
dharmas	7	19	0.72
great	5	19	0.72
physical	8	19	0.72
span	4	19	0.72
attained	8	18	0.68
merit	5	18	0.68
bodhisattvas	12	17	0.64
bodhisattva	11	16	0.60
body	4	16	0.60
means	5	16	0.60
notions	7	16	0.60
charity	7	15	0.57
teaching	8	15	0.57
attain	6	14	0.53
non	3	14	0.53
perfect	7	14	0.53
thought	7	14	0.53
without	7	14	0.53
worlds	6	14	0.53
thoughts	8	13	0.49
even	4	12	0.45
sand	4	12	0.45
actually	8	11	0.42
eye	3	11	0.42
ganges	6	11	0.42
others	6	11	0.42
appearances	11	10	0.38
particles	9	10	0.38
rise	4	10	0.38
tiny	4	10	0.38

yes	3	10	0.38
follow	6	9	0.34
immeasurable	12	9	0.34
know	4	9	0.34
nothing	7	9	0.34
paramita	8	9	0.34
remember	8	9	0.34
teach	5	9	0.34
thus	4	9	0.34
two	3	9	0.34
abide	5	8	0.30
extremely	9	8	0.30
fact	4	8	0.30
form	4	8	0.30
four	4	8	0.30
gives	5	8	0.30
like	4	8	0.30
meaning	7	8	0.30
merits	6	8	0.30
neither	7	8	0.30
practice	8	8	0.30
real	4	8	0.30
resolve	7	8	0.30
thirty	6	8	0.30
also	4	7	0.26
attachment	10	7	0.26
away	4	7	0.26
buddhas	7	7	0.26
comprehend	10	7	0.26
dipankara	9	7	0.26
former	6	7	0.26
give	4	7	0.26
grains	6	7	0.26
men	3	7	0.26
mind	4	7	0.26
reality	7	7	0.26
recite	6	7	0.26
someone	7	7	0.26
verse	5	7	0.26
women	5	7	0.26
come	4	6	0.23
countless	9	6	0.23
faith	5	6	0.23
free	4	6	0.23
jewels	6	6	0.23
line	4	6	0.23
man	3	6	0.23
many	4	6	0.23

Appendix E. Readability Tests Formulae

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$

Automated Readability Index

$$CLI = 0.0588L - 0.296S - 15.8$$

L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

Coleman–Liau Index

$$206.835-1.015\left(\frac{\text{total words}}{\text{total sentences}}\right)-84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Flesch Reading Ease

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Flesch–Kincaid Readability Test

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right)+100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

Gunning Fog Index

$$\text{grade} = 1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

SMOG

Exploiting weak-supervision for classifying Non-Sentential Utterances in Mandarin Conversations

Xin-Yi Chen
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
xysimba.chen@connect.polyu.hk

Laurent Prévot
Aix Marseille Université & CNRS
Laboratoire Parole et Langage
laurent.prevot@univ-amu.fr

Abstract

Non-sentential or fragmentary utterances (NSU) constitute a significant part of the productions in a conversation. Although seemingly incomplete in form, they convey full pragmatic meaning in the context. In the past, their classification had been approached with supervised methods (Fernández et al., 2007; Wong, 2018). Such approaches require relatively large annotated data sets. We explore an approach (Ratner et al., 2017a) that allows the reduce significantly the amount of annotated data needed thanks to strategic use of linguistic knowledge. We explore this method for classifying NSUs in Mandarin conversation corpus. Our evaluation shows that promising results can be obtained with a minimal amount of annotated training data.

1 Introduction

In dialogue, besides well-formed complete sentences, a sizeable amount of utterances are fragments that could be understood without a problem in their context. Traditional grammar attends mainly to written texts and canonical sentence analysis. The oral language has been often regarded as bad, spontaneous, and wrong, in summary not an appropriate research object, as (Blanche-Benveniste, 1997) regrets it. But as interest in oral communication gets more attention, terms like "fragments", "Nonsententials" in (Barton, 1991) or "Non Sentential Utterances" (hereafter NSU) in (Fernández et al., 2007) have also attracted more investigation.

The expressions in example 1 below may sound familiar.

(1) What now?
Not you.
What's for supper? - Ground Beef Tacos.

Even though they are generally short, such utterances constitute an active part of the conversation. They contribute to the efficiency of the conversation flow. The interpretation of NSU is essential for linguistic theories that attempt to get serious about language as it is produced in its most natural and pervasive setting, and also for applications, like dialogue systems. It can be done in different ways, as discussed in (Ginzburg, 2012, p:229). The analysis result can be implemented in human-machine dialogue systems in various domains such as client service or computer aided language teacher.

The percentage of NSU among other utterances in conversation corpus is non negligible, 11.15 % in (Fernández and Ginzburg, 2002), 9% in (Fernández et al., 2007), 10.2 % in (Schlangen and Lascarides, 2003). We think the study of NSUs is useful because of the high frequency mentioned above. What's more, the understanding of NSUs and their classification from their context is not always easy. Even a simple "what" can express various emotions and can have different functions in a context. Apart from the most common function as plain question, it can also express Happiness, Surprise, Sadness, Anger, Disgust or even Fear.

Second, the definition of NSUs can have an impact on the classification of NSU, the inclusion and exclusion of categories can be flexible according to the theories and purpose of classification, the classification criteria could be syntactic leading, semantic

leading or a mix of standards. The treatment of some fragments like *'Greetings'* and *'Filler'* can make a difference in the counts. We will see the detailed discussion in section 2.

The paper is structured as follows. Section 2 presents the related work of utterance classification, including Dialogue Acts and Non-Sentential Utterances. Section 3 introduces the data and methodology. Section 4 provides a qualitative and quantitative description of our corpus and the results of the manual labeling. Section 5 summarizes the labeling functions used in this article. Section 6 talks about the modelling and classification experiment in our work. Section 7 is about the evaluation of the model. Section 8 concludes the article.

2 Related Work

2.1 Non-Sentential Utterances

The NSU taxonomy proposed in (Fernández and Ginzburg, 2002) is supposed to be the first "comprehensive, theoretically grounded classifications of NSU in large-scale corpus". The classification is based on work grounded in British National Corpus (BNC), the classification take into consideration both a relatively complex syntax and the context dynamics. In (Fernández et al., 2007), several machine learning experiments were carried out to get an optimal classification result. The features selected for machine learning in this article is limited in a few "meaningful" ones instead of many arbitrary ones. The features selected for NSU classification came from (i) the utterance itself, (ii) its antecedent, and (iii) their relationship. It results in three sets of features in total: *NSU features, Antecedent features* and *Similarity features*. The NSU features include four aspects, whether it is proposition or question, presence of wh-word, yes/no word, and different lexical items. The antecedent features are similar to those of NSU features, but it also looks at whether it is a finished utterance. The similarity features is a comparison of the utterance and its antecedent, mainly about the repeated words and POS tags and their proportion. Another machine learning experimentation work for classification of NSU is based on the work of (Fernández et al., 2007) with more advanced features in (Dragone, 2015).

The taxonomy can be adapted for languages besides English, following the work of (Fernández et al., 2007), the work of (Wong and Ginzburg, 2013) in classifying NSUs in Chinese adds seven subcategories because of the particular behavior of modal verbs in Chinese. The classification we choose is (Wong, 2018), which is based on the work of (Fernández et al., 2007) in adding some classes considering particular behaviors in Chinese Mandarin with extended discussion of each category compared with (Wong and Ginzburg, 2013).

2.2 Utterance Classification

NSU classification is an utterance classification task, of the same kind as the better known Dialogue Act tagging (Stolcke et al., 2000). Dialogue Act (DA) is about the meaning at the illocutionary level defined in (Austin, 1962), which is the intent or effect produced along with the things being said. In (Stolcke et al., 2000), it is said that DAs can be considered "as a tag set that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria." The DA labels demonstrate the hidden information of the utterance for higher-level processing. It can be used in the interpretation and generation and prediction of utterances and their functions in dialogue systems, as stated in (Stolcke et al., 2000). Therefore DA-tagging is a major applicative task for NLP and Human-Machine Interaction.

Lexical and prosodic cues are both useful for the dialogue act classification. It is observed that some words are symbolic of some DAs. For example, in (Stolcke et al., 2000), "92.4% of the uh-huh's occur in Backchannels, and 88.4% of the trigrams '(start) do you' occur in Yes-No-Questions." For some shared patterns, the differentiation is by pronunciation.

The methodology in DA classification bears similarity with NSU classification. Nevertheless, DA and NSU have differences in their theoretical frameworks and distinctions in aspects such as label uniqueness. NSU is an utterance that is not realized by a full syntactic sentence but produces an effect just like sentential utterances. All utterances can receive a DA label, but only those fragments with incomplete syntactic structure and full semantic value can be labeled as NSU.

By many aspects such as their size as well as their lack of completeness, NSUs can be confused with *disfluencies*. Shriberg (Shriberg, 1996) talked about several types of disfluency: *filled pause, repetition, substitution, insertion, deletion and speech error*. In (Tseng, 1999)'s exploration of modeling the disfluency, there are features found to be useful in the detection of disfluency on the syntactic side: the linguistic length, the syntactic category, the construction types, the location of interruption, the repair onset, and the repair offset. These features could be useful in our examination of disfluency in our corpus. In (Tseng, 2003) 's research about repairs and repetitions in spontaneous Mandarin, the editing term (an indication of speech repair such as "well" "I mean" or filled pauses) is found to be useful in the detection of repetition and repairs.

3 Methodology

A large quantity of training data is necessary for machine learning tasks. But labeled data are not easy to get. Snorkel (Ratner et al., 2017b) provides a solution to this bottleneck by using labeling functions to generate a large amount of labeled data. As stated in (Ratner et al., 2017b), based on theories and experiments, Snorkel has proven effective in training high-accuracy machine learning models, even using potentially lower-accuracy inputs. It has been recently applied to high-level NLP such as discourse parsing in (Badene et al., 2019).

Weakly supervised tools like Snorkel allows for quickly labeling extensive data with minimal but expert manual involvement. The use of Snorkel is to write some labeling functions (LF) to produce some useful training data with labels. A labeling function is a rule that attributes a label for some subset of the training data set. Using Snorkel, it will train a model that combines all the rules defined written to estimate their accuracy, along with the overlaps and conflicts among different labeling functions.

The workflow of Snorkel distinguishes from traditional machine learning approaches; it is based on a data programming paradigm. Briefly, it is composed of two phases, and the first is to produce estimated labels using a generative model, the second is using these labels to train the ultimate model, a discriminative model.

Within this design philosophy, the system design of Snorkel can be divided into three phases: first, pre-processing of the data to have the reorganized data for later use, such as word segmentation and POS tagging. Second, writing labeling functions. Labeling functions do not need to be entirely accurate or exhaustive and can be correlated. Snorkel will automatically estimate their accuracies and correlations in a provably consistent way, as introduced in (Ratner et al., 2016). Third, after the evaluation and calibration of the LFs, we decide on an optimal set of LFs to produce a set of labels to train a model.

4 Data and Manual labelling

Extensive conversational data are limited in numbers. We are interested in the real-time conversational data in talking form transcribed in textual form instead of texts generated in instant-messaging tools. The data we used in this study is from LDC's CALLHOME Mandarin Chinese collection. This is a telephonic conversation corpus, with audio files and transcriptions. The language was in Mandarin even though the participants are from different provinces of China. The corpus includes 120 transcripts in total, and each is a five or ten-minute segment from the telephone speech files. From the description on the website [1], the transcripts are already tokenized automatically using a tool called the Chinese Lexical Analysis System (ICTCLAS). The results were further corrected manually.

In this paper, the corpus concerned is already segmented. In long sentences, an NSU component may appear in the middle, but we won't label it as NSU if it doesn't stand independently. Suppose we deal with a raw corpus not segmented yet. In that case, we will decide the utterance boundary first based on our research question(s) and the conversation context, including syntax, prosody, and pragmatic effect. However, it's also possible to define NSU and describe it first and then extract them or locate them in the corpus.

The original data includes the start time and end time of every segment, the speaker, the textual content of the utterance. In the text transcription, there are also examples of annotation as enrichment of information as illustrated in example 2.

[1] https://catalog.ldc.upenn.edu/LDC2008T17

(2) Examples of annotation [2]
{text}: sound made by the talker. e.g. {laugh} {breath_noise}
//text//: aside (talker addressing someone in background) e.g.// 来 说 ⟨English_Hello,⟩ 您好 . (Come say Hello, hello).//

We processed the data and transformed it into Pandas DataFrame (McKinney, 2010) in order to manipulate it into Jupyter Notebooks (Pérez and Granger, 2007). They are transformed as a table, and the information is divided by columns. The original information is separated into four columns: *Start time*, *End time*, *Speaker*, and *Text transcription*. Based on these, we added other columns (illustrated in figure 1):

- Conversation code: the original code of the file
- Duration : how long the utterance lasts
- Same Speaker: if the utterance is produced by the speaker of the previous utterance (BOOLEAN)
- Latency : a gap between two turns, the Start time minus the previous End time (we only consider the positive value cases)
- Overlap : the duration when more than one person speaks (we only consider the positive value cases)
- Word count: How many units are there in the utterance (depending on the segmentation method, one unit may not necessary correspond to one Chinese character, and the punctuation can be included as well)
- Tagged: the POS tagged text used in this study is attributed by the tool Zpar (Zhang and Clark, 2011)

We have 33485 utterances in total in combining 120 files. Combined with the tagged results, we omit the ones untagged, so we deal with 33431 utterances (229 412 tokens).

[2] http://shachi.org/resources/661

We selected around 5% of the whole data as a sample to tag manually to know the difference between data with NSU tags and the complete data. Only one annotator does the manual annotation for convenience and cost. Then we have another annotator to annotate 7% of the sample data (0.35% of the whole data) to compare with the first annotator's result. We get a kappa score of 0.54 for all the NSU categories and a kappa score of 0.57 for the four most frequent NSU classes((PLAIN ACKNOWLEDGMENT, REPEATED ACKNOWLEDGEMENT, CHECK QUESTION, and INTERJECTION), a kappa score of 0.58 for the four first-level NSU classes (ACKNOWLEDGMENT,QUESTION, ANSWER and COMPLEMENT).

Through a qualitative analysis of the NSU categories in our corpus, we made some adjustments of the classification in (Wong, 2018), the results are shown in table 1.

5 Labelling Functions

When writing labeling functions, there are several strategies: keyword matches, regular expressions, arbitrary heuristics, and third-party models.

In our case, we use the first two strategies combined with three types of cues: the Textual cues, the Timing cues, and the Contextual cues. For each type of signal, we look at the relevant features. There are two variables in (Schlangen, 2005), the structural features and the lexical/utterance-based features. In (Fernández et al., 2007), as mentioned in section 2, there are three sets of features: *NSU features, Antecedent features, and Similarity features*. In (Dragone, 2015), the baseline feature set is the same as in (Fernández et al., 2007), but with extended features at different levels: POS tags, phrase-level, dependency features, turn-taking features, and similarity features. We have chosen these features as presented in the figure 2 based on the characteristics and available information of our corpus.

In our case, the features are used in the writing of labeling functions. Based on the result of LF performance, which is undoubtedly influenced by the majority's classes, the more frequently used features are the keywords, such as feedback/ backchannel word, followed by wh-question word and ques-

Latency	Overlap	Same_speaker	Speaker	Start	Text	Word_count	Tagged
NaN	NaN	False	A	183.47	你 很 爽 哈 你 , 哈 ?	6	你_PN 很_AD 爽_AD 哈_VV 你_PN ,_PU 哈_VV ?_PU
NaN	0.27	False	B	184.93	要 不要 跟 妈 讲话 ?	5	要_VV 不_AD 要_VV 跟_P 妈_NN 讲话 _VV ?_PU
0.08	0.00	False	A	186.03	啊 ?	1	啊_VA ?_PU

Figure 1: Head of the pre-processed corpus dataframe

tion final particles. Features used to detect the Sentential Utterances and Disfluency also have good performance. Some features may be not so effective because of some shared words among different NSUs, thus less frequent due to major classes' existence. For instance, "嗯"(um) is typical in PLAIN ACKNOWLEDGEMENT. Still, it can also appear in INTERJECTION or questions, so that we may need a combination of features such as POS tag features and other corpus-related cues.

Our labeling functions can be divided into three types: Keyword-based LF combined with size-related LF, POS tagging LF, Context-related LF. For the three classification models, we set the size-related limitation such as counted words and we used frequent words for each NSU category in the LF, and also frequent POS tag or tag combination, such as demonstrated in figure 3. We also compare the number or promotion of shared patterns between the utterance and its precedent. For the SU class, we also have LF targeting at disfluency with size-related LF, such as Duration and Word_count, contextual cues (two consecutive utterances produced by the same speaker) and POS tag cues.

6 Modelling and Classification

Our goal is to build a model to classify all the NSU classes, we also build two extra classification models for comparison, one with the four first-level classes ACKNOWLEDGMENT,QUESTION, ANSWER and COMPLEMENT, and another with the four most frequent NSU classes(PLAIN ACKNOWLEDGMENT, REPEATED ACKNOWLEDGEMENT, CHECK QUESTION, and INTERJECTION).

It should be noted the final set for each model only includes the LFs without serious incorrectness. Otherwise, it will only harm the model so that if an LF has more incorrect than the correct cases, we tend to exclude them, especially when the ratio is significant. Based on the result and after the error analysis, this problem could not be solved; we do not have LFs for each NSU. For the main classes model, we didn't get a proper LF for the class REPEATED ACKNOWLEDGEMENT, for the first-level classification model, the ANSWER class, and COMPLEMENT class LF don not enter in the final set. For all-class model, only PLAIN ACKNOWLEDGMENT, CHECK QUESTION, and INTERJECTION) entered in the final set.

The results are presented in table 3. For each model, we run the experiment in three conditions:

- Baseline: with the definite majority class acknowledgment (the most frequent one) with 58% in our sample data frequency;
- System: with all the classes in each model, no use of punctuation (the training label difference in these three conditions can be seen in table 2);

	NSU Class
	A. Acknowledgement
1	Plain Acknowledgement
2	Repeated Acknowledgement
3*	Verbal Acknowledgement
4*	Helpful Acknowledgement
5*	Re-Affirmation
	B. Questions
6	Clarification Ellipsis
7	Sluice
8*	Nominal Predication
9	Check Question
	C. Answers
10	Short Answer
11	Affirmative Answer
12	Repeated Affirmative Answer
13*	Verbal Affirmative Answer
14*	Helpful Affirmative Answer
15	Rejection
16*	Verbal Rejection
17	Helpful Rejection
	D. Complement
18	Filler
19*	Correction
20*	Interjection
21	Propositional Modifier
22	Factive Modifier
23	Bare Modifier Phrase
24	Conjunction + Fragment

Table 1: Classification of NSU in (Wong, 2018)

Feature	Description
NSU feature	Presence of wh-question word
	Presence of question final particle
	Presence of propositional modifier word
	Presence of feedback/backchannel word
	Presence of interjection
	Presene of factual modifier word
	Presence of modal word
	Presence of polar particle
	Presence of heavy tags (noun, verb, ajective and adverb)
	Presence of disfluency
	Presence of repeated pattern(word/tag) in the utterance
	Is the utterance a question or not?
Antecedent features	Presence of wh-question word
	Presence of question final particle
	Presence of disfluency
Structural features	Is the two consecutive utterance produced by the same speaker?
	Common pattern(word/tag) between the utterance and its precedent

Figure 2: Features and description

- Topline: also with all the classes in each model, including punctuation (provided by the transcript) as cues.

Snorkel's Label Model can learn the dependency among the LFs, and its output is an array of single probabilistic training labels. As explained in (Ratner et al., 2016), there are four types of dependency among the LFs: "similar, fixing, reinforcing, and exclusive." A dependency graph will be calculated and established. Overall, the model will give a more data-balanced decision for the data points where there are conflicting LFs.

The Majority Label Voter of Snorkel takes the majority vote for each data point; each LF will cover a portion of data. Its inadequacy is that each vote of the LF are considered of equal efficiency, but this is not the case. Snorkel's Label Model deals with the correlation among LFs when combining all the outputs of the LFs.

As we have mentioned the workflow of Snorkel in section 3, Snorkel's Label Model's output is then used to train the ultimate discriminative model, such as a Scikit-Learn classifier.

The Label Model Accuracy is not always higher than the Majority Vote Accuracy. In (Ratner et al., 2017b), it's explained that for very sparse label matrices (almost no conflicts among LFs) or very dense label matrices (a lot of conflicts among LFs) will probably lead to this result. The F1 score is a Micro average for the multiclass setting, that "calculates metrics globally across classes, by counting the total true positives, false negatives and false positives", as explained in (Sasaki, 2007).

7 Evaluation

So the result of a task to detect just the majority class ACKNOWLEDGMENT from the Sentential Utterace (SU) and the rest of the NSU classes is acceptable, but the abstain votes from the other NSU classes can explain the gaps with the system condition. The small difference among all these three conditions can be attributed to the outcome of the final labeling function sets. Because we omit the LFs with apparent imprecision, we are left with an LF set targeting classes for some major classes like ACKNOWLEDGMENT and a few effective others for the rest.

```
@labeling_function()
def ackplain_pos(x):
    tags = [utt.split('_')[1] for utt in x['Tagged'].split()]
    if x["Word_count"] < 2:
        for tag in tags:
            if tag not in ['NN','AD','VA','PU']
                return ABSTAIN
    return ACKPLAIN
```

Figure 3: Example of LF using unigram POS cues, PLAIN ACKNOWLEDGEMENT

Transcript	Baseline	System	Topline
呃哼/ Uh-huh	SU	Interjection	Interjection
寄出来了/ It's coming out	SU	Repeated Ack.	Repeated Ack.
好不好? / All right ?	SU	Check Question	Check Question

Table 2: Comparison of training labels in baseline, system and topline situations in Majority class classification model

		Baseline	System	Top-line
All-class classification model	Majority Vote Accuracy	72.70%	72.70%	73.70%
	Label Model Accuracy	72%	72%	75.30%
	F1 micro	0.75	0.78	0.82
	Scikit-learn classifier test accuracy	68.70%	74.30%	75.70%
First-level classification model	Majority Vote Accuracy	80.00%	79.70%	84.00%
	Label Model Accuracy	79%	77%	84%
	F1 micro	0.79	0.8	0.83
	Scikit-learn classifier test accuracy	74.30%	74.00%	76.00%
Majority class classification model	Majority Vote Accuracy	67.30%	74.30%	75.00%
	Label Model Accuracy	67%	74%	74.30%
	F1 micro	0.77	0.82	0.77
	Scikit-learn classifier test accuracy	67.70%	73.30%	74.70%

Table 3: Performance of three models in three conditions

The all-class classification model's performance can be attributed to the number of classes and the affiliation relation between them. The 24 tags are mutually exclusive, but some can be grouped under a first-level category. Besides, the SU class is the opposite of all the other classes. With its relatively high frequency, in a binary situation, when we only need to distinguish SU and NSU. Still, in our multi-class setting, one class's negative classification is not yet realized in Snorkel. For some classes, even though we have posed some limits on the counted word number and duration, the LF still targets many SU (including disfluency cases). Consequently, there are many false-positives for some LFs, especially for some minority categories, such as for different sub-categories under ANSWER. Extremely unbalanced data as reference, they do not have a single case present in the labeled data set.

Also, the similarity between ANSWER and ACKNOWLEDGEMENT makes it hard to classify the ANSWER and its sub-classes. They have shared words and sometimes similar scope of counted words; the most credible way is by solving whether the previous utterance is a question. But when we do without the punctuation, the performance is not so good, neither. Some INTERJECTION words are also confused with the ACKNOWLEDGEMENT.

It's exceptionally delicate when dealing with some classed heavily depending on the semantic relationship. For the all-class model, we haven't come

up with LFs for BARE MODIFIER PHRASE and CORRECTION who are hard to capture.

8 Conclusion

In this article, we present our work regarding non-sentential utterances automatic classification. NSUs are utterances partial syntactically but convey integral meaning semantically. We chose one classification for Chinese Mandarin and test it with a telephone conversation corpus, using a weak supervision method to build a model for automatic labeling.

From a broader perspective, the approach adopted shows interesting results. It constitutes an efficient way to combine domain experts (here linguists) with state-of-the art machine learning techniques.

Future development For classes with barely any coverage in the reference data set, such as the subcategories in ANSWER, we can put more data of these classes for the model training and use some data augmentation method so that we can test and find the LF for these classes.

Dealing with classes easily confused with majority class, such as INTERJECTION and ACKNOWLEDGEMENT, we may need audio-related information to distinguish them, such as intensity and energy of utterances. Prosodic information has appeal in separating question from declarative with the rising tone at the end for the Mandarin.

To find the semantic connection for a particular utterance in cases, especially when there are no repeated patterns, we need tools to present the relatedness not only for two consecutive utterances but with a flexible contextual window.

Acknowledgments

We would like to thank Pierre Magistry for helping with the POS-tagging as well as anonymous reviewers for extremely valuable comments. All remaining errors are ours.

References

[Austin1962] John Langshaw Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

[Badene et al.2019] Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Data programming for learning discourse structure. In *Association for Computational LInguistics (ACL)*.

[Barton1991] Ellen Barton. 1991. Nonsentential Constituents and Theories of Phrase Structure. In Katherine Leffel and Denis Bouchard, editors, *Views on Phrase Structure*, pages 193–214. Springer Netherlands, Dordrecht.

[Blanche-Benveniste1997] Claire Blanche-Benveniste. 1997. *Approches de la langue parlée en français*. Collection L'essentiel français. Ophrys, Gap Paris. graph. 21 cm. Bibliogr. p. 149-151. Index.

[Dragone2015] Paolo Dragone. 2015. Non-sentential utterances in dialogue: Experiments in classification and interpretation. *CoRR*, abs/1511.06995.

[Fernández and Ginzburg2002] Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

[Fernández et al.2007] Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach. *Computational Linguistics*, 33(3):397–427, September.

[Ginzburg2012] Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press UK.

[McKinney2010] Wes McKinney. 2010. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

[Pérez and Granger2007] Fernando Pérez and Brian E. Granger. 2007. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May.

[Ratner et al.2016] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3574–3582, Red Hook, NY, USA. Curran Associates Inc.

[Ratner et al.2017a] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017a. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160.

[Ratner et al.2017b] Alexander J. Ratner, Stephen H. Bach, Henry R. Ehrenberg, and Chris Ré. 2017b. Snorkel: Fast Training Set Generation for Information

Extraction. In *Proceedings of the 2017 ACM International Conference on Management of Data - SIGMOD '17*, pages 1683–1686, Chicago, Illinois, USA. ACM Press.

[Sasaki2007] Yutaka Sasaki. 2007. The truth of the F-measure. page 5, October.

[Schlangen and Lascarides2003] David Schlangen and Alex Lascarides. 2003. The interpretation of non-sentential utterances in dialogue. page 10.

[Schlangen2005] David Schlangen. 2005. Towards finding and fixing Fragments—Using ML to identify non-sentential utterances and their antecedents in multi-party dialogue. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 247–254, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[Shriberg1996] Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of the International Conference on Spoken Language Processing*, pages 11–14, Philadelphia, PA, October.

[Stolcke et al.2000] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

[Tseng1999] Shu-Chuan Tseng. 1999. *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. Ph.D. thesis, University of Bielefeld.

[Tseng2003] Shu-Chuan Tseng. 2003. Repairs and repetitions in spontaneous mandarin. In *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*.

[Wong and Ginzburg2013] Kwong-Cheong Wong and Jonathan Ginzburg. 2013. Investigating non-sentential utterances in a spoken chinese corpus.

[Wong2018] Kwong-Cheong Wong. 2018. *Classifying Conversations*. Ph.D. thesis, Université Paris Diderot - Paris 7.

[Zhang and Clark2011] Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Pay Attention to Categories: Syntax-Based Sentence Modeling with Metadata Projection Matrix

Won Ik Cho and **Nam Soo Kim**
Department of Electrical and Computer Engineering and INMC,
Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826
wicho@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Sentence modeling is a vital feature engineering for document classification. Various feature extraction and summarization algorithms have been adopted for efficient classification of a sentence, e.g., dense word vectors and neural network classifiers. Recently, the concept of attention for machine translation has been applied to various natural language processing (NLP) tasks and has shown significant performance. In this paper, we take a look at the syntactic categories of the words, to make up a metadata projection matrix that assigns strong restrictions on determining the attention weight. Unlike conventional attention models, which are considered as a division of location-based approaches, our model adds a selection layer to highlight categorical metadata that may appear more than once. The proposed algorithm shows improved performance compared to the baselines with the tasks in syntax-semantics, suggesting a possibility of extension to other fields such as symbolic music or bitstream analysis.

1 Introduction

Sentence modeling, which incorporates featurization and embedding, has been widely studied from short utterances to large-scale documents. Its usefulness and broad applicability have been proven with various classification and regression tasks. Also, in recent years, attention models have demonstrated the significant performance of such approaches, along with deep learning techniques that have shifted the paradigm of the standard recipes.

In applying the attention models, we noted that the utility of the syntactic properties should be explored in a bit wide point of view. Like the notes in music that have corresponding chords, the observable components of a sentence are assigned syntactic categories after constituency parsing, such as noun, verb, and adjective. They are interpreted as a kind of metadata regarding each token[1], that may appear more than once in the document. We want to claim such information can be exploited in making up the attention weight, not just being adopted as input-level data. For instance, in an oxymoron identification task (Cho et al., 2017), given a sentence like "*This is a sugar-free sweet tea.*", it may be beneficial for the analysis to attend to *sugar-free* and *sweet* with a similar concentration, mainly due to their syntactic property being close to each other.

Although such syntactic properties can be represented in various ways such as tree structure and dependency, we pay attention to part-of-speech (POS), for some practicality. First of all, we already have many computationally efficient tools that can extract syntactic classes from the tokens of the sentence. Next, even though the POS tagger is not entirely accurate, the general tendency may provide sufficient information for classification. This flexibility can be supportive for the proposed model to analyze corpus with non-formal sentences such as tweets.

The proposed model differs from the usual self-attentive models in that it takes into account the information of syntactic categories while maintaining

[1] Henceforth, we interchangeably use (token-wise) categorical data, categorical metadata, and categorical information, all referring to the syntactic classes that each token belongs to.

the original form of classification that uses word vector sequence[2]. Furthermore, the model tells us how much attention we should pay to the components with specific syntactic properties, given the overall summarization of a sentence. The contribution of this study is as follows:

- We suggest a modified version of the conventional location-based attention model by inserting a simple projection layer that contains information on the syntactic categories.
- We verify the utility of the proposed scheme with widely used benchmarks and suggest further usage.

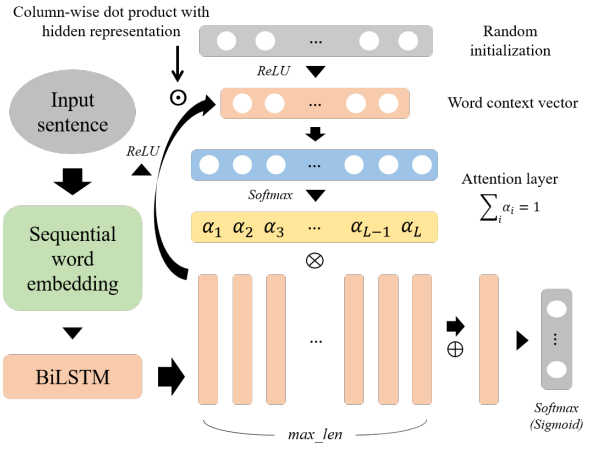


Figure 1: Descriptive diagram of attention model presented in self-attentive sentence embedding (Lin et al., 2017). The arrows in the figure indicate the flow of information. The triangles in the overall system denote the fully connectedness to the dense layer, together with the stated activation functions.

2 Related Work

2.1 Sentence embedding

Embedding a sentence into numerics is an essential process in data-driven sentence classification. Two major types of representation are widely used, namely sparse and dense.

One of the most popular sparse word representations, bag-of-words (BoW) model, is a one-hot encoding of the words in the sentence and is most commonly used for its conceptual clarity. Another well known sparse representation is the term frequency-inverse document frequency (TF-IDF), which conveys the relative importance of the terms in each document.

The main issue of BoW and TF-IDF is that they can hardly give information about the context window of each term in a sentence. Thus, count-based approaches for the local context window of words have been studied, as in Lebret and Collobert (2013). However, it can also be problematic because such approaches can disproportionate weight to words with large counts. They can also cause a dimensional explosion.

To cope with the above, Mikolov et al. (2013) proposed an algorithm that embeds a word into a low dimensional dense vector that involves a local context window. The real-vectorized words facilitate similarity computation between the original words and can be used to represent sentences, e.g., by averaging Le and Mikolov (2014). In Pennington et al. (2014), the advantages of the approaches in Lebret and Collobert (2013) and Mikolov et al. (2013) were combined.

2.2 Modeling techniques in classification

In sentence classification, basic recipes such as naive Bayes, decision trees, and logistic regression models were conventionally used. Among such models, the support vector machine (Cortes and Vapnik, 1995) showed quite a practical accuracy.

However, ever since the computational breakthrough that had taken place in the deep neural network (DNN) system (Hinton et al., 2006), neural architectures have been adopted within the sentence classification tasks, along with the emergence of dense word vectors. Convolutional neural network (CNN), which initially came up for the image classification task (Krizhevsky et al., 2012), was successfully applied to the sentence classification task (Kim, 2014). Recurrent neural networks (RNN) (Schuster and Paliwal, 1997; Graves, 2012), which had been proposed to deal with sequential data processing, also have shown significant performance in sentence classification tasks through various forms such as gated recurrent unit (GRU) (Tang

[2]In other words, here we don't adopt attachment such as 'word/POS'.

et al., 2015) and bi-directional long short-term memory (BiLSTM) (Chen et al., 2017), comprehensively summarizing sentences into dense vectors.

Lately, attention models have been applied to the neural machine translation (Bahdanau et al., 2014) in the way of multiplying the attention vector with the decoder-encoder network matrix to generate a particular target word from the source word. It can be regarded as jointly training a weight vector augmented to a feature or hidden layers to focus on a specific part of the input feature. Driven by its conceptual clarity, it was soon applied to areas such as image captioning (Xu et al., 2015) and natural language interface (Liu et al., 2016).

In Lin et al. (2017), the self-attentive embedding (SA, Figure 1) was applied to the sentence classification, by aggregating essential attributes of the hidden layers into sentence vectors. A word context vector, which is multiplied by the higher-level representation of hidden layers in BiLSTM, is used to create attention (weight) layer with a sum equal to 1.

In detail, for $X = X_1^L$ the input token sequence, $H = H_1^L$ the hidden layers, weight W_t and bias b_t, the BiLSTM hidden layers are defined as:

$$H_t = \tanh\left(W_t\left[X_t, H_{t-1}\right] + b_t\right) \tag{1}$$

As in the right top of Figure 1, each hidden layer is multiplied with word context vector C to yield a softmax-ed attention vector α with $\sum_t \alpha_t = 1$, as:

$$\alpha_1^L = \mathrm{softmax}\left(H_1^L \odot C\right) \tag{2}$$

where $\odot$ denotes a column-wise dot product. α_1^L is further multiplied to H_1^L and is summed to be fed to the final decision layer, as a representative hidden layer output:

$$H_o = \sum_t \alpha_1^L \otimes H_1^L \tag{3}$$

where $\otimes$ denotes a column-wise multiplication of the scalar weights. In the figure, L equals to the maximum sentence length *max_len* and $\oplus$ denotes the weighted sum of the hidden layers.

Note that this basic architecture covers most of the sentence-level attention schemes that precede the contemporary self-attention models (Vaswani et al., 2017; Devlin et al., 2019). In this regard, at this point, we consider this structure suffices as a baseline to implement our scheme on, due to the assignment of attention weight being interpretable and straightforward.

3 Proposed Method

In this section, we demonstrate the concept of *Pay Attention to Categories*, or PAC structure, which can adequately reflect the categorical metadata of each token onto the attention model. It denotes an insertion of a projection matrix that incorporates the information on syntactic classes, which yields the modified attention weight that comes afterward. Materializing it accompanies three main steps, namely (a) constructing word vector sequence, (b) feature extraction for the attention source, and (c) projecting the weight that corresponds with the category of each token (or here, syntactic classes) to the attention layer.

(a) Word vector sequence can be constructed by methodologies used in general. It is briefly depicted at the bottom of Figure 2, especially step (2), where *max_len* denotes the upper limit of the sentence length regarding word count. Summarizers such as CNN and BiLSTM employ this as a feature, using sigmoid (binary case) or softmax (multi-class) as an activation function.

(b) Attention source utilizes various features extracted from the sentence. It can be TF-IDFs, averaged word vectors, or the output layer of a CNN or BiLSTM summarizer. In this paper, (bigram) TF-IDF and BiLSTM hidden layer output were adopted based on the performance. They are fed to PAC structure after passing a single dense layer with rectified linear unit (ReLU) activation, as depicted in the top of Figure 2.

(c) PAC structure consists of a layer carrying the category-wise weight (shortly a weight layer), a projection matrix, and their multiplication (the attention layer). The size of the weight layer (n_p) equals to the number of the categories that appear throughout the document.

In detail, let S be the attention source and $ReLU, hsig$ be activation functions. Then, for given n_p, we get:

$$w_p = \mathrm{hsig}(\mathrm{ReLU}(S)) \tag{4}$$

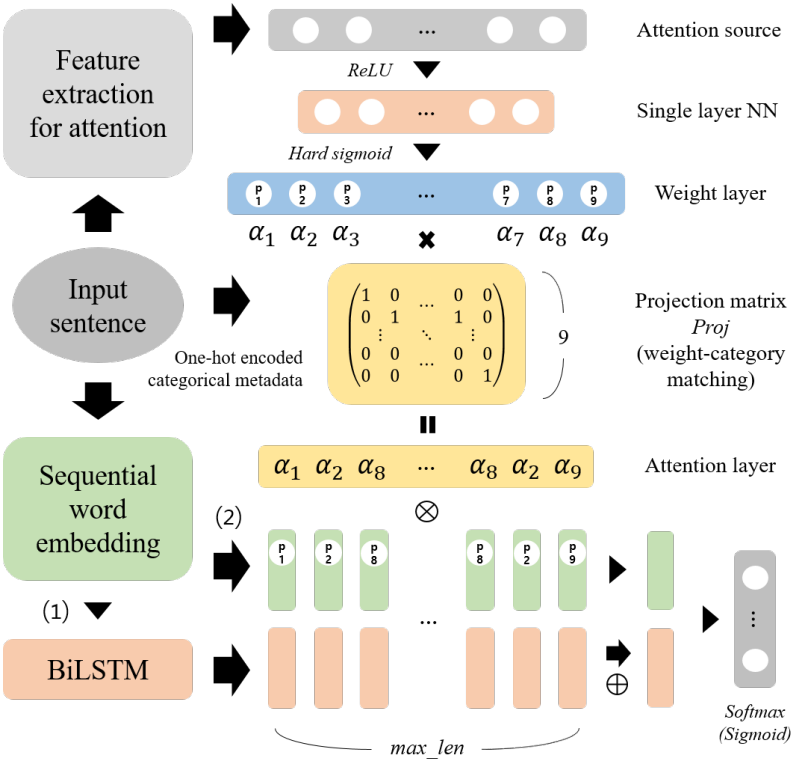


Figure 2: A Descriptive diagram for the proposed system.

On the other hand, we have a fixed projection layer which contains the syntactic information regarding each token. The matrix $Proj$ is of size (n_p, L), and each column tells the syntactic category each token belongs to. In this study, it is represented by POS. We multiply it with the former weight layer to obtain the attention layer of width L:

$$\alpha_1^L = \text{matmul}\,(w_p, \text{Proj}) \tag{5}$$

It consists of the weight corresponding to each word of the sentence and is column-wisely multiplied to either the hidden layers (*PAC-Hidden*) or the word vector sequence (*PAC-Word*). The two strategies are depicted in Figure 2, where $\times$ denotes a matrix multiplication and $\otimes$ denotes a column-wise multiplication of the attention layer to (1) the hidden layer sequence as BiLSTM output (*PAC-Hidden*), or (2) the original word vector sequence (*PAC-Word*). For *PAC-Word*, the weighted word vector sequence becomes an input of BiLSTM again.

More on the figure, to help the readers understand, we specified the number of categories (n_p = 9), as shown in the weight layer w_p. The sequence of one-hot encoded vectors of categorical metadata, $Proj$, expressed in the form of a projection matrix, conveys the weight to the attention layer, concerning the syntactic class that each column (of hidden layers or word vector sequence) incorporates. For instance, if the index regarding a word's syntactic class is 2, as in the case of the second and the second to the last, it is multiplied by the value conveyed from α_2. Note that this setting allows the repetition of the attention weight. It is worth noting that the activation function of the weight layer is set to hard sigmoid as in (4). We surmised that the hidden layer's information should be fully retained even after it is transferred to the projection matrix. Here, hard sigmoid plays a vital role, minimizing information that can be nullified in multiplication with the one-hot encoded matrix.

4 Experiment

In this section, we describe the benchmark datasets, the specific implementation scheme, and the result comparison with baselines.

4.1 Dataset

Five datasets were used in the evaluation. The specification for the datasets is displayed along with the corpus size.

Metalanguage detection (2,393) employs the corpus for English metalanguage detection (Wilson, 2012), which investigates whether a sentence contains explicit mention terms, namely with the lexicons such as 'title' or 'name'. It contains 629 *mentioned* and 1,764 *not-mentioned* instances excerpted from Wikipedia.

Irony detection (4,618) utilizes corpus recently distributed in SemEval 2018 Task 3 for ironic tweet detection (Van Hee et al., 2018). All instances (that includes emoji) in the training set and Gold test data were used. Only the binary label case was taken into account. 2,222 instances contains *irony* and 2,396 does not.

Subjectivity detection (10,000) refers to Pang and Lee (2004), which checks if the movie review contains a subjective judgment, in the view of sentiment polarity. It incorporates equally 5,000 instances for each of the *subjective* and *objective* reviews.

Stance classification (3,835) employs a part of the distributed dataset from SemEval 2016 Task 6 (Mohammad et al., 2016). The original dataset consists of the additional labels corresponding to target, stance, opinion towards and sentiment information. All instances with favor and against stances in the dataset were excerpted. Among instances with none as stance, only those not explicitly expressing opinions were taken into account. There are 1,205, 2,409, and 221 instances for *favor*, *against*, and *none* each.

Sentiment classification (20,632) utilizes the test data released in SemEval 2017 Task 4 (Rosenthal et al., 2017). It consists of 7,059 *positive*, 3,231 *negative* and 10,342 *neutral* tweets, with all instances labeled via crowd-sourcing.

4.2 Implementation

The implementation for the whole network was done with Python libraries, including NLTK (Bird et al., 2009), Scikit-learn (Pedregosa et al., 2011), and Keras (Chollet and others, 2015). In particular, POS tagging and word tokenization process employed the tools included in NLTK. Here, elaborate implementation schemes of baselines and the proposed system are presented.

4.2.1 Baselines

Features Baseline features were chosen from both sparse and dense ones. For sparse features, TF-IDFs and their bigrams were extracted. The dimension (the number of commonly used words) was fixed to 3,000 for a fair comparison, which is the same size as a multiplication of *max_len* (=30) and word vector dimension (=100). The uni/bigrams were obtained via TfidfVectorizer of Scikit-learn.

For dense features, 100-dimensional GloVe (Pennington et al., 2014) pre-trained with 27B token Twitter data[3] was adopted as a word vector dictionary, since the words thereof were expected to cover the lexicons of the task corpora. The dense static features were constructed by aggregating the vectors corresponding to every word in the sentence and normalizing it using the l_2 norm. The dense sequential features were constructed by padding the word vectors with a maximum length of 30.

Basic classifiers All evaluations were conducted using 10% test set. Non-parameter optimized linear-kernel SVM of Scikit-learn was used for sparse features (*TF-IDF-SVM*), and NN classifiers in Keras were used for dense features. NN used for the static dense features (*Averaged GloVe-NN*) consists of a single hidden layer of size hidden dim and is optimized with Adam (Kingma and Ba, 2014) of learning rate 0.0005. The network was trained with mini-batch of size 16, reducing the cross-entropy loss. The implementation toolkit, optimizer, and mini-batch size for all NN classifiers were not changed throughout the experiment. For every model, *hidden_dim* was chosen as the best case after hyperparameter tuning with 32, 64, and 128.

CNN and BiLSTM were used in the baseline sequential feature classification (*GloVe-CNN/BiLSTM*). In CNN, two single-channel convolutional layers (with 32 filters and a window of size 3) were used with a max-pooling layer in between. In BiLSTM, time-distributed hidden

[3] https://nlp.stanford.edu/projects/glove/

layers had an output size of 32*2 = 64 units.

Baseline attention model The attention adopted from Lin et al. (2017) was implemented as depicted in Figure 1. The word context vector of size *hidden_dim*, which is fully connected to a randomly initialized layer, is column-wisely dot-multiplied by the *ReLU*-activated[4] hidden representation of the sequential hidden layers of BiLSTM, also of size *hidden_dim* and length *max_len*. Note that the product layer of size *max_len* undergoes the regularization process using the softmax function (sum to 1), unlike the model proposed in this work.

The attention vector was applied to the word sequence in two different ways: by directly multiplying it to a hidden layer sequence (*SA-Hidden*), or by multiplying it to a word vector sequence (*SA-Word*). In the former case, which was suggested in the original paper, the final decision was made by investigating the weighted sum of the hidden layers. The latter case, which was supplemented to observe the tendency of each strategy, investigates the weighted word vector sequence with BiLSTM.

4.2.2 The Proposed

The proposed system extracts three input features from each sentence: *attention source, projection matrix, and word vector sequence*.

As previously mentioned, two features were adopted as attention source: TF-IDFs and BiLSTM outputs. For TF-IDFs, the sparse vector of dimension 3,000 is itself an attention source[5]. Unlike the case of TF-IDFs where the source is assigned as an input, all parameters of BiLSTM are trained jointly with the entire system.

The attention source is fully connected to the single dense layer of size hidden dim, with *ReLU* activation. Consecutively, this is fully connected to the weight layer with hard-sigmoid activation, as described in the previous section.

The size of the weight layer and the projection matrix depends on the corpus. In a corpus with n_p syntactic classes (the number of categories), a weight matrix of length n_p and a projection matrix of size (n_p , *max_len*) are obtained. Again, the emphasis is that the weight layer is optimized in the training session, but the projection layer is given as input.

Finally, the attention layer of size *max_len* appears as the product of the matrix multiplication of the weight layer and projection layer. All its entries are multiplied as the weight to each column of either the hidden layer sequence of BiLSTM (*PAC-Hidden*) or the word vector sequence (*PAC-Word*).

5 Result and Discussion

Per task characteristics The proposed system surpasses the baseline systems in tasks that are expected to be accompanied by lexical-semantic analysis, such as *META*, *IRONY*, and *SUBJ* (Table 1). Also, it was observed that the systems fit with small datasets as well, considering the significant improvement in *META* and *IRONY*. In tasks where semantics are considered much more important, such as *STANCE* and *SENT*, the proposed system showed a stable and adequate result, not an improvement in performance. This result implies that the proposed system may rather boost the performance of the tasks that utilize the existence and meaning of the lexicons thereof, than the semantic tasks that require more a latent analysis.

Source and assignment of attention We observed that the tendency regarding attention source, namely TF-IDF or BiLSTM output, is opaque and non-consistent, considering that no significant tendency is displayed. On the other hand, the contrast on *Word-* level and *Hidden-* level assignment of attention weight is quite significant per task. Especially for *META*, *IRONY*, and *SUBJ*, where the proposed methods outperform the baselines, we found that *META* highly prefers *Word*-level assignment, while *Hidden*-level assignment works better for the other two. This directly shows that *META* concerns the explicit existence of certain lexical terms, while the other two touch relatively abstract areas of lexical-semantics.

Under context-dependency Specifically, the lower performance and stochastic results in *STANCE* seem to originate in the omission of *target data* in this experiment. It is essential situational

[4] In view of performance and fair comparisons, *tanh* used in the original paper was replaced with *ReLU*.

[5] In case of *META* and *STANCE*, bigram was chosen considering the comparison result (Table 1).

F1 Score	Features	META	IRONY	SUBJ	STANCE	SENT
Sparse features	***TF-IDF***	0.5466	0.6236	0.8953	0.4316	0.5604
	Bigram TF-IDF	0.5489	0.6137	0.8944	0.4334	0.5509
Dense features	***Averaged GloVe-NN***	0.5454	0.6455	0.8845	0.3676	0.6157
	GloVe-CNN	**0.6800**	0.6613	0.9036	0.4141	0.6121
	GloVe-BiLSTM	0.6527	0.6639	0.9159	**0.4763**	0.6304
Attention	***SA-Word***	0.6363	0.6447	0.9152	0.3703	0.6297
	SA-Hidden	0.6478	**0.6771**	**0.9203**	0.4317	**0.6538**
Proposed	***TF-IDF PAC-Word***	0.7105	0.6679	0.9204	**0.4671**	0.6241
	TF-IDF PAC-Hidden	0.6535	**0.7019**	**0.9268**	0.4253	0.6329
	BiLSTM PAC-Word	**0.7261**	0.6585	0.9135	0.4332	0.6353
	BiLSTM PAC-Hidden	0.6400	0.6956	0.9259	0.4475	**0.6529**

Table 1: Performance comparison of the baselines and the proposed system. *META, IRONY, SUBJ, STANCE*, and *SENT* denote the datasets in Section 4.1, respectively. *SA-Word/Hidden* refer to the self-attentive embedding models. TF-IDF and BiLSTM coming before *PAC-Word/Hidden* represent the attention sources. The final decision of the proposed systems was also made through BiLSTM. In the baselines and the proposed models, the best scores were bolded. Underlined cases denote when the proposed system surpasses the baseline.

information in determining a stance towards someone but was not digitized in this experiment. Also, there was a shortage in the number of instances associated with none. On the other hand, for instance, in *IRONY* where situational information is essential as well, the proposed system showed an outperformance. It is assumed that in *IRONY*, hashtagged information plays a critical role (Cho et al., 2018), and accordingly, attention is given to functional parts as well.

Summary We concluded that paying attention to relatively important syntactic classes such as verbs (*META*), nouns (*IRONY*), or adjectives (*SUBJ·IRONY*) is advantageous in some tasks. This inference is also consistent with the consideration for polarity items (Krifka, 1995), which takes into account the relation between words of different syntactic classes. From this point of view, a suitable application of the proposed system would be a case where the categorical metadata plays a significant role in determining the labels of the data, and the pattern is relatively clear, e.g., bitstream analysis.

5.1 Visualization

The normalized attention weight of baseline and the proposed, namely *SA-Hidden* and *TF-IDF PAC-Hidden*, are visualized as Figure 3 with two excerpt sentences from *SUBJ*.

Considering the property of the dataset, it is clear that the attention should be given to the words in the sample sentences that affect the subjectivity. In the top example, the baseline model pays attention to *vile* and *tacky*, which are the subjective modifiers indicating the object *ghost ship*, while the proposed model addresses *best*, the superlative which can directly show the subjectivity of the sentence. Besides, at the bottom, the proposed model pays attention to *comedy*, which reveals the sarcastic tone, while the baseline only attends to *funniest* among the lexical candidate words.

Without a doubt, this kind of advantage in the inference partially benefits from the task being sensitive to specific sentiment items in the sentence. Nonetheless, beyond the examples above, the proposed model can stably give attention to the specific categories that seem to be important in analyzing the document. Given that this kind of consistency is sometimes threatened in the analysis of informal or non-canonical utterances, stable fixation of weight can be advantageous often. Also, we note that each category's weight varies with the content of the sentence, making the proposed model differ from hard attention.

5.2 Further Study

Beyond a simple application that considers syntactic categories as property for words, the proposed system can be extensively utilized to datasets where observable components contain metadata of a type

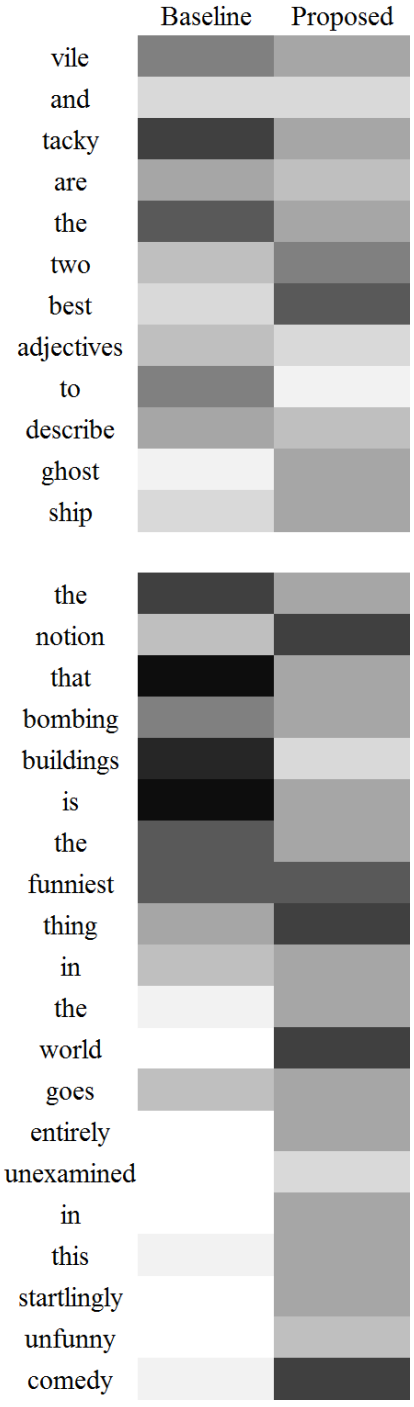


Figure 3: Visualization of the attention weight given to the subjective example sentences in *SUBJ*.

that possibly appears more than once. For example, in a paragraph or large-scale document analysis, a sentence type or document topic can be used as such information. In the field of music information retrieval, chord information can be provided to the attention model to help predict whether the type of the musical phrase (Livingstone et al., 2009) is cadence, semi-cadence, false cadence, or nothing. In acoustic event detection (Choi et al., 2017), event labels can also be used as a property to identify acoustic scenes, even in the multi-label conditions.

6 Conclusions

In this paper, the concept called *Pay Attention to Categories*, or PAC structure, was suggested for efficient sentence classification. The proposed system fully utilizes the syntactic class of each token, which is modeled in terms of POS for words, in making up a special kind of projection matrix, and employ it in building up attention weight. Its conceptual simplicity and flexibility were demonstrated with an intuitive diagram, and the validity was verified via comparison with widely used benchmarks. Beyond utilities in many NLP areas, the system is expected to have a significant role in tasks that require attention to categorical information.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00059, Deep learning multi-speaker prosody and emotion cloning technology based on a high quality end-to-end model using small amount of data).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classi-

fication using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230.

Won Ik Cho, Woo Hyun Kang, Hyun Seung Lee, and Nam Soo Kim. 2017. Detecting oxymoron in a single statement. In *Proceedings of Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 48–52.

Won Ik Cho, Woo Hyun Kang, and Nam Soo Kim. 2018. Hashcount at semeval-2018 task 3: Concatenative featurization of tweet and hashtags for irony detection. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, SemEval-2018, New Orleans, LA, USA, June. Association for Computational Linguistics.

Inkyu Choi, Soo Hyun Bae, Sung Jun Cheon, Won Ik Cho, and Nam Soo Kim. 2017. Weakly labeled acoustic event detection using local detector and global classifier. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1735–1738. IEEE.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Alex Graves. 2012. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Manfred Krifka. 1995. The semantics and pragmatics of polarity items. *Linguistic analysis*, 25(3-4):209–257.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Rémi Lebret and Ronan Collobert. 2013. Word emdeddings through hellinger PCA. *arXiv preprint arXiv:1312.5542*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Steven R Livingstone, Emery Schubert, Janeen Loehr, and Caroline Palmer. 2009. Emotional arousal and the automatic detection of musical phrase boundaries. In *Proceedings of the International Symposium on Performance Science*, pages 445–450.

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.

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 (SemEval-2016)*, pages 31–41.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

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*, pages 1532–1543.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in English Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, SemEval-2018, New Orleans, LA, USA, June. 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.

Shomir Wilson. 2012. The creation of a corpus of English metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 638–646. Association for Computational Linguistics.

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.

Metaphoricity Rating of Chinese KIND Metaphor Expressions

Siaw-Fong Chung
National Chengchi University
sfchung@nccu.edu.tw

Meng-Hsien Shih
National Chung Cheng University
simon.xian@gmail.com

Yu-Hsiang Shen
National Chengchi University
108161001@nccu.edu.tw

Wei-Ting Tseng
National Chengchi University
108161004@nccu.edu.tw

Abstract

Metaphors are ubiquitous in human language, and there has been an increasing interest in metaphor processing and interpretation from the field of computational and cognitive linguistics. Although metaphors have been greatly researched, there remain certain kinds of metaphors that are under-researched. One of them is KIND metaphors. KIND metaphors are less discussed, and less noticed, for their status as metaphors is often less recognized. In the paper, we looked at KIND metaphors by first identifying 245 sentences of KIND metaphors from the Corpus of Contemporary Taiwanese Mandarin, and later rated the metaphoricity of these metaphorical expressions. The evaluation of the rated data reported an inter-rater agreement (Krippendorff's alpha) of 0.65 and a unanimous percentage of 37%, near a tentative agreement, if we used the strictest criterion to keep the quality of our rating.

1 Introduction

The research of metaphors has blossomed in the late 80s and early 90s and has been given great attention for decades. Different theories of metaphors have been proposed since then although the Conceptual Metaphor Theory (CMT) (Lakoff, 1993; Lakoff and Johnson, 1980; Lakoff and Turner, 1989) was the widely-received one in the 1990s. More review of Chinese metaphors can be found in Ahrens and Chung (2019). Although many studies of various languages have tried to validate the CMT, the theory has also received criticisms from many (among many, Vervaeke and Kennedy (1996) was a direct criticism of its example grouping). Recently, the CMT has been revitalized for it is believed to have new contributions in shedding light on language framing or in explaining how "[d]ifferent metaphors frame the same topic in different ways, facilitating different inferences and evaluations" (Potts and Semino, 2019, p. 81). The fact that the CMT has become active in recent development has not changed the types of metaphors being noticed. For example, some marked metaphors, as those mentioned by Goatly (1997), are still under-researched. KIND metaphor is one such type.

Goatly (1997, p. 174) surveyed some markers of metaphors such as (a) superordinate terms (identified by the use of *sort of*, *kind of*); (b) copular similes (*like*, *as*); (c) clausal similes (*as if*, *as though*); (d) perceptual processes (*seemed*, *sounded*, *looked*, *felt*, *tasted*, *+like/as though/as if*), etc. These markers were later used in the development of Metaphor Identification Procedure (MIP) but *sort of*, *kind of* and other "[m]ore general signals of all indirectness" was not included for "it is not always clear that they signal metaphoricity or other aspects of discourse" (Steen, Dorst, Herrmann, Kaal, and Krennmayr, 2010, pp. 40-41). In example such that in (1), the metaphorical expression is marked by *a kind of*,

which serves to mark a Transfer metaphor,[1] but further investigation of this metaphor marker is needed.

(1) *I have no relish for the country;* ***it's a kind of*** *healthy grave.*

This becomes the motivation of the current work, for KIND metaphors (cf. following term used in Shih, Chung, Shen, and Liao, 2020) were not properly researched in the past. It is not clear how metaphorical they are, and their proportions of occurrences in the corpus are also unclear.

For Chinese marked metaphors, they have been discussed in a couple of works, but not qualitatively or extensively. In as early as 1982, 袁暉 provided a list of 'marked metaphors', such as '像'組 *xiang*-group 'like'-group: among which are markers such as 像 *xiang* 'like', 就像 *jiuxiang* 'just like', 很像 *henxiang* 'very much like', etc. (p. 13) and many other types of markers for Chinese metaphors/similes. A follow-up work, by Wang, Lu, Hsu, Lin, and Ai (2019, p. 247), claimed that "[t]o date, MIPVU [MIP developed at the VU University Amsterdam] has not yet gained wide currency in the research field of metaphor in Chinese".[2]

When KIND metaphors are often left out, there has been little or no discussion of their levels of metaphoricity. Skorczynska and Ahrens (2015) investigated "the use of words and phrases that signal metaphors" in the "US presidential addresses, popular science articles, and business periodical articles". In their study, they took "metaphor signals" from Goatly's (1997) work. Among the signals, *sort of* and *kind of* were under the 'Superordinate terms' category. Among the three genres, these two terms appeared most often in the business periodical articles, constituting 0.0312 and 0.0203 frequency per 1,000 words respectively. This is followed by the popular science articles (0.0211 and 0.0130 frequency respectively). The presidential addresses had the least number of these signals. However, different from Skorczynska and Ahrens, our KIND metaphors were judged in terms of their metaphorical meanings, while the uses of signalers *sort of* and *kind of* in both Goatly (1997) and Skorczynska and Ahrens (2015) might not be metaphorical in meanings. This is also the reason why we used the measurement of 'metaphoricity' in our experiments.

Patterson (2017, p. 103) said that despite the advancements in computational research on metaphors, the definition of 'metaphoricity' is still largely ignored: "However, while deriving metaphoric data from corpora is by now well established within the field[…], its premise of focusing on repetitive patterns of use means that some cases of metaphoricity are often ignored." From a computing purpose, Potts and Semino (2019) calculated 'metaphoricity' as the percentages of metaphors found from the total instances. (This definition was derived from the data. It was not implicitly given.) Earlier, Hanks (2006, p. 22) claimed that "some metaphors are more metaphorical than others": "In most metaphorical cases, the secondary subject shares fewest properties with the primary subject. Therefore, the reader or hearer has to work correspondingly harder to create a relevant interpretation. At the other extreme, the more shared properties there are, the weaker the metaphoricity." This shows that when two mapped subjects are highly similar in properties, they carry less metaphorical meaning.

In this paper, two issues will be addressed, namely, the KIND metaphors as a type of marked metaphors, and their patterns of occurrences as well as their level of metaphoricity judged by human raters. Two research questions are postulated for these purposes:

(a) How are the proportions of KIND metaphors found from *kind-of* expressions extracted from corpus?
(b) What are the levels of metaphoricity judged by human raters on the KIND metaphor expressions identified in corpus?

These two research questions will be answered and their results will contribute to our understanding of KIND metaphors in use.

[1] Goatly (1997, p. 18) classified "metaphors as Approximative when the distance between the thought and proposition is small, and as Transfer metaphors when the gap is larger."

[2] In Wang et al.'s short chapter, however, no in-depth discussion was provided except for the list of several direct, indirect, and implicit metaphors with these markers, which will not be discussed in details here. Their markers were similar to 袁暉's (1982) list.

2 Related Works

In the computational field, the issues of metaphor identification and comprehension are no new topics in metaphor research, and the use of corpus in finding patterns of metaphor has also become the trend. Many works have been carried out to annotate metaphorical expressions from various perspectives. The Pragglejaz Group (Pragglejaz Group, 2007; Steen, Dorst, Herrmann, Kaal, and Krennmayr, 2010; Steen, Dorst, Herrmann, Kaal, Krennmayr, et al., 2010) annotated 200,000 words of sentences (from the British National Corpus) with metaphorical meaning. We can observe that their Metaphor Identification Procedure (MIP) focused on the distinction between the basic meaning and metaphorical meaning of metaphor-related words. For example, in their study "all uses of *defend* and *attack* in contexts of argumentation can be analyzed as metaphorical" (Steen, Dorst, Herrmann, Kaal, and Krennmayr, 2010, p. 770).

On the other hand, Dunn (2014) measured the metaphoricity of 60 sentences from the Corpus of Contemporary American English. In his study, 100 unique participants from the Mechanical Turk platform annotated the metaphoricity of the whole sentence with three labels: "Not Metaphoric", "Slightly Metaphoric", and "Very Metaphoric", but the inter-annotator agreement of the results were not shown. With a more fine-grained scale of five points, Shih et al. (2020) rated the metaphoricity of Chinese KIND metaphors and similes by asking annotators to judge the metaphoricity of the two concepts extracted using both parser and by hand. The annotation was based on a five-point Likert scale with 1 being 'least metaphoric' and 5 being 'most metaphoric'. The study, however, encountered several difficulties, which were improved upon in this work. First, with the automatic identification of the two concepts ('A' and 'B' in 'A is a kind of B') in KIND metaphors, it was found that the modifiers were left out and this could affect the results (considering 市場是一種神話 vs. 自由市場是一種神話, the first may mean any kind of markets including the 'traditional market is a kind of legend' while the later means 'free market', a more abstract concept). However, because Shih et al. aimed at comparing the concept distances, it was unavoidable that only the head nouns could be included. Second, the previous work needed a better guideline apart from informing the annotators what was least metaphorical or most metaphorical. In this paper, we tried to improve upon these two issues by providing a two-staged guideline.

Although Shih et al. (2020) had some limitations, it was among the few studies that investigated KIND metaphors as marked metaphors. Much work in English have been conducted to annotate metaphorical expressions, less focus was placed on the nominal concepts in a metaphorical expression, and a Chinese resource with metaphor rating is still under development. In our study, we first manually identified Chinese KIND metaphorical expressions from corpus. From the KIND metaphors we have identified from corpus (Stage one), a Stage-two rating of the metaphoricity of these metaphorical expressions with a three-point Likert scale will be conducted. The agreement among raters will also be examined. The details of our identification and rating guidelines will be given in the next section.

3 Methodology

To identify KIND metaphor expressions and rate the metaphoricity of these expressions, we used a two-staged design to elicit data. In Stage one, we

	Criteria	**Example**
Metaphorical sentences	The domains of A and B (in the pattern of 'A is a kind of B') are different.	**自省** 是 一 種 **防腐劑** *zixing* *shi* *yi zhong* *fangfuji* Self-reflection is a kind of preservative
Non-metaphorical sentences	A is the literal class inclusion of B.	**電視劇** 是 一 種 **通俗** **文化** *dianshiju* *shi yi zhong* *tongsu* *wenhua* TV drama is a kind of popular culture

Table 1: The Identification of KIND metaphors and data description.

Metaphoricity	Description	Examples
3	- This expression has a metaphorical meaning. The concepts A and B have basic sense respectively but they show certain sense in specified context in the KIND metaphor structure. - When A and B are not simple inclusion, their meaning may vary in different contexts.	(2a) **自省 是 一 種 防腐劑** *zixing shi yi zhong fangfuji* 'Self-reflection is a kind of preservative'
2	This expression is less metaphorical than 3 but it is not frozen as in 1.	(2b) **學習 是 一 種 探險** *xuexi she yi zhong tanxian* 'Learning is a kind of adventure'
1	This structure is frozen for B is often used in a formulaic way regardless of what A is.[3]	(2c) **聆聽 是 一 種 藝術** *lingting shi yi zhong yishu* 'Listening is a kind of art'

Table 2: The rating guidelines of KIND metaphor expressions.

collected data from the Corpus of Contemporary Taiwanese Mandarin (COCT), which consists of 319,712,694 words in total. We focused on one pattern of KIND metaphors, 是一種 *shi yi zhong* 'is a kind of'. We used the CQP (Corpus Query Processor) syntax to elicit the data from corpus and downloaded all the data into Excel for further analysis.

For the structure of 是一種 *shi yi zhong* 'is a kind of', a CQP pattern <s> []{0,10} [word="是"] [word="一"] [word="種"] []{0,10} </s> was typed into the query box, and the COCT corpus system returned 9,058 matched sentences. In terms of the first stage of this KIND metaphor identification, a binary classification task was conducted in Excel by recognizing if A and B form a KIND metaphor (see Table 1). For example, the sentence 自省是一種防腐劑 'Self-reflection is a kind of preservative' will be considered as metaphorical, but the sentence 電視劇是一種通俗文化'TV drama is a kind of popular culture' is not. The metaphor and non-metaphor distinction was made based on the distance between the two concepts in the sentences of A and B 'A 是一種 B', and this distance was the main key identifier to conduct the binary identification at Stage one.[4] If simple literal inclusion was found, the sentence would be recorded as non-KIND metaphor and therefore would not be included in the Stage-two metaphoricity rating.

For Stage two (metaphoricity rating), we first used the 245 sentences from Stage one. Three raters were recruited to rate the metaphoricity of these sentences based on the rating guidelines of KIND metaphor expressions (see Table 2).

According to the guidelines, sentences were rated with the metaphoricity of 3 if they feature high metaphorical meanings when A collocates with '是一種 B', which means A and B have a basic sense respectively but they show a combined different meaning in the KIND metaphor structure. For example, in the sentence (2a), 'self-reflection' and 'preservative' both have a respective basic sense, but when they appear as a KIND metaphor, they together form a metaphorical meaning. This principle is more or less similar to the MIPVU but it is also different in the sense that MIPVU considers only one lexical item. In our version, we considered the relation between the concepts A and B in the pattern of 'A is a kind of B'.

On the other hand, the sentences that were rated with the metaphoricity of 1 show to possess frozen meanings although they are still KIND metaphors.

[3] We needed to separate the frozen expressions for their 'B's were high in frequency and we would like to make sure that they did not affect the metaphoric categories of 2 and 3.

[4] For this paper, we defined the nominal concepts as 'domains' but further discussion of their scopes is definitely needed in the future.

Sentences such as (2c) are formulaic because the concept B is too conventionalized that one needs not process further to understand why the concept A is an art, a tool, or a crime (cf. Table 4 for more terms). As for the middle category, 2, their sentences have less metaphorical meanings than those of 3, and are not as frozen as those with the metaphoricity of 1 (e.g., the sentence 2b).[5]

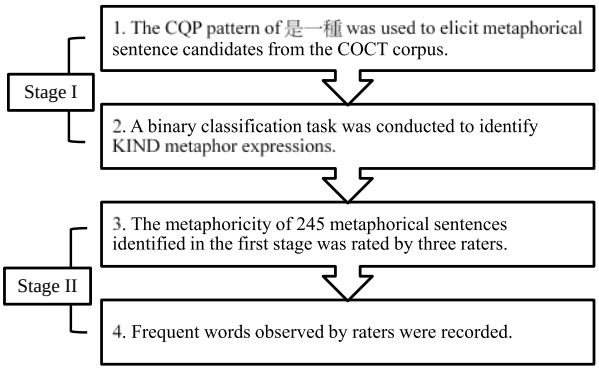


Figure 1: Flow chart of the two-staged design.

In addition, the three raters were requested to record the frequent words which occurred as concept B in the 'A is a kind of B' pattern. By 'frequent', we meant the lexical items in the concept B could occur with several types of concept A. Although separating 2 from 3, and 2 from 1, was not an easy task but we felt that there was a need to do this because some KIND metaphors are less conventionalized but not as frozen. We then asked raters to label the 245 sentences based on our criteria. The flow chart of the two-staged design is illustrated in Figure 1.

4 Results

In the first stage, from the total 9,058 instances, we were only able to analyze 1,780 sentences due to the great amount of data, and among these we found 245 sentences of KIND metaphors (based on the identification criteria in Table 1) for further rating. This indicates that about 14% of the *is-a-kind-of* sentences are metaphorical, a percentage far lower than the 30% metaphors found in previous work (cf. Chung, 2009). This also means that KIND metaphors could be a special type whereby its possibility of carrying a metaphoric meaning is only half of the chances of other metaphors. Using four elements (Topic, Vehicle, Ground, and marker), Goatly (1997, p. 169) once claimed that if a sentence, e.g., (3a), has all the four elements, such expression is "marked out of existence, sometimes to the extent of becoming a literal comparison or simile". Comparatively, sentence (3b), which only consists of two elements, is more metaphorical.

(3) (a) *One or two tupaia species*$_{\text{topic}}$ run along branches$_{\text{ground}}$ ***like*$_{\text{m}}$ *squirrels***$_{\text{vehicle}}$.

vs.

(b) *Housework* $_{\text{topic}}$ *is a* ***treadmill***$_{\text{vehicle}}$.

This helps explains why a marked metaphor such as a KIND metaphor (simile in the case of 3a) is less metaphorical than other metaphors identified through the CMT (cf. Chung, 2009). However, this does not explain whether or not all KIND metaphors are equally more or less metaphorical than one another. This is the second goal of this paper.

In the second stage, the total 245 sentences in the pattern of 是一種 *shi yi zhong* 'is a kind of' from the COCT corpus were rated in terms of their metaphoricity in the three-point Likert scale by three raters. We also provided an option to the raters that if they still considered an expression as a non-metaphor, they could write a remark after the sentence. This was to ensure that they only rated the sentences they all agreed as metaphors. To keep the quality of our rating data, we also removed sentences when either two of the ratings were too diverse (i.e., a difference

Pattern	Number of sentences	Unanimous percentage	Agreement
是一種 *shi yi zhong*	176	37 %	0.65

Table 3: The numbers of rated sentences of KIND metaphors with the corresponding unanimous percentages and inter-rater agreement (Krippendorff's alpha).

[5] According to Goodman (1976, p. 82), "[a] frozen metaphor has lost the vigor of youth, but remains a metaphor."

Metaphoricity	Concept B in the pattern of 'A is a kind of B'
1	途徑、折磨、障礙、革命、財富、威脅、災難(災禍)、過程、習慣(惡習)、負擔、信號、懲罰、折磨、危險、遊戲、訊號(警訊)、***語言、罪(罪惡)***、治療、負擔、知識、信仰(宗教)、修養(修行、修練)、折磨、***陷阱、投資(資本)***、良藥(魔藥)。
2	***語言、罪(罪惡)、陷阱、投資***、浪費、武器、挑戰、***資本***(資源)、***投資***、幻、束縛、罪惡、冒險、機器、享受。
3	–

Table 4: Frequent nouns occurring as the concept B in the pattern of KIND metaphors, and the corresponding metaphoricity.

of two points in the three-point Likert scale). Due to these strict criteria, 69 sentences were removed either because one rater considered that a sentence is not a metaphor, or two raters had diverse ratings.[6]

After removing the 69 sentences, the remaining 176 sentences were evaluated by Krippendorff's alpha inter-rater agreement. We found a 0.65 (near a tentative agreement), as shown in Table 3. From these 176 instances, we obtained a 37% unanimous percentage, meaning that more than one-third of the sentences were given similar rating by the three raters. The remaining 63% were either between 1 and 2, or 2 and 3. As indicated earlier, the separation of 2 from 1, and 2 from 3, was not an easy task. The results that we obtained was more than satisfactory for identification of metaphoricity is never a clear-cut task.

During the process of rating, the raters also found frequent nouns occurring as the concept B in the pattern of KIND metaphors with specific metaphoricity. We collected the frequent concepts in Table 4. Some concepts B were identified as metaphors but with different levels of metaphoricity by the raters. For these concepts, we highlighted them in italic so that we could reconsider them in the future.

5 Discussion and Limitations

This paper sets out to account for Chinese expressions of KIND metaphors in the pattern of 是一種 *shi yi zhong* 'is a kind of' in Mandarin. We first manually identified metaphorical expressions in a binary classification task, and then rated the metaphoricity of these metaphorical expressions with a three-point Likert scale by three raters. After the rating, there were some sentences that at least one rater considered non-metaphorical, or the ratings among two of raters were diverse. We removed these sentences and calculated the inter-rater agreement of each patterns. This section is a more detailed discussion of result of the pattern of A 是一種 B 'A is a kind of B'.

The 69 removed sentences consist of 18 sentences at least one rater considered non-metaphorical, and 51 sentences of which the disagreements between two of the raters were too strong. We found that the raters considered the 18 sentences non-metaphorical because of the following two reasons. First, these concepts of A or B in the 'A is a kind of B' pattern are too domain-specific that the raters found it hard to rate them confidently without first understanding their intended meaning. For example, in sentence (4) 禪 *chan* 'Zen' is a word from Buddhism referring to deep meditation, and 法門 *famen* 'way' is also a word from Buddhism.[7] The raters then decided to

[6] If we took these 69 sentences with diverse rating into consideration, the inter-rater agreement in Krippendorff's alpha is 0.34 with a unanimous percentage of 29%.

[7] Porat and Shen (2015, p. 82) found that the metaphoricity of a class of constructions can be imposed and vary according to either a literal reading or a metaphorical reading of the same sentence (e.g., *this*

regard this example as a non-metaphorical one. The second reason was that sometimes the part-of-speech of A in the 'A is a kind of B' pattern is not a noun, which differs from the traditional form of metaphors, and complex structures are shown in our data. Another possibility was that when both A and B are abstract, raters found it difficult to decide and therefore tended to disregard them.

(4) **禪** ，是 一 種 促進 人類 心智
chan，shi yi zhong cujin renlei xinzhi
Zen ， is a kind boost human mind
甦醒 的 **法門** 。
suxing de famen .
awaken DE way .
"Zen is a kind of way to boost human mind to become awaken."

In the basic structure of KIND metaphors, both A and B are usually nouns, and nouns are easier for raters to rate their metaphoricity. However, in example (5), 這 *zhe* 'this' refers to a previous clause in the previous context, and this complicated the metaphor in this example, leading to the raters' hesitation in labeling their metaphoricity. However, we were not able to remove these examples beforehand because our study was a corpus-based study and we should include all examples that were judged as metaphorical at Stage one.

(5) 佛陀 為什麼 要 拈 花 微笑 呢？
Fotuo weishenme yao nianhua weixiao ne ?
Buddha why will take flower smile NE?
很 明顯 的， **這** 是 一 種 意義
Hen mingxian de ， zhe shi yi zhong yiyi
very obvious DE , this is a kind meaning
豐富 的 象徵式 **語言** 。
fengfu de xianzhenshi yuyan .
rich DE symbolic language .
"Why Buddha smile while taking flowers? Obviously, this is a kind of symbolic language with a rich meaning."

The above were some reasons for raters to analyze the sentences with different outcomes. The problems we encountered in the removed examples were not unexpected because metaphors are creative and even in the less varied form of KIND metaphors, we still found complex sentence patterns.

5.1 Metaphoricity of KIND metaphor expressions

In order to understand the ratings of remaining 176 sentences, we analyzed their different ratings in Table 5. If we take the metaphoricity of 2 as the cutting point, half of the rating fell between 1 and 2 (19%+17%+15%=51%), while a slightly smaller number (9%+16%+14%+10%=49%) fell between 2 and 3 (including 2). This indicates that KIND metaphors were half frozen. Only 10% were considered highly metaphorical. From here, we could answer the question we asked previously – how metaphorical are expressions of KIND metaphors? The answer was 51% close to frozen, and 49% close to metaphorical. Examples that are highly frozen are shown as (6).

Average	Total (%)	Description
3.00	17 (10%)	All raters rated the sentences with the metaphoricity of 3.
2.67	25 (14%)	Two raters rated 3 and one rater rated 2 for the sentences.
2.33	28 (16%)	One rater rated 3 and two raters rated 2 for the sentences.
2.00	16 (9%)	All raters rated 2 for the sentences.
1.67	27 (15%)	Two raters rated 2 and one rater rated 1 for the sentences.
1.33	30 (17%)	One rater rated 2 and two raters rated 1 for the sentences.
1.00	33 (19%)	All raters rated the sentences with the metaphoricity of 1.
1.91	176 (100%)	Total

Table 5: The average metaphoricity of sentences in the pattern of 是 一 種 *shi yi zhong* 'is a kind of'.

book is an encyclopedia). This could become a future work.

(6) **語言　是 一 種　工具** 。
yuyan shi yi zhong gongju .
Language is a kind tool .
“Language is a kind of tool.”

The concept B 工具 *gongju* ‘tool’ can only be interpreted in one way, that is, to indicate that something is useful. Since this kind of interpretation occurs so often, people often do not realize that it is metaphorical anymore. This is what we called fully fixation here. For KIND metaphors, this happen quite frequently.

For KIND metaphors, especially, as time goes by, if the same metaphors are used more frequently, the way of use will gradually be fixed. It might become frozen more easily than other kinds of metaphors. All in all, we have identified several features of KIND metaphors that deserve our in-depth discussion. Through this research, we can provide a preliminary contribution to the metaphoricity and the process of KIND metaphors that have not been previously discovered.

5.2 Marked metaphors and the CMT theory

In the Conceptual Metaphor Theory, most metaphorical expressions are embodied in our daily language. Markers are not necessarily needed. The investigation of the KIND metaphor expressions in this paper, however, implies that marked metaphors (such as KIND metaphors and similes) feature a special distribution of metaphorical expressions in the corpus. The 245 identified KIND metaphor sentences only account for 14% of the 1,780 sentences in the ‘is a kind of’ pattern, with almost only half of the 30% metaphorical sentences found in other research (Chung, 2009, p. 77). In addition, the 1,535 non-metaphorical expressions in the ‘is a kind of’ pattern are worth further analysis.

In this paper, we brought our attention to the mapped concepts marked in the KIND metaphor other expression such as similes, and also provided a window to respond to the embodied concepts in the Conceptual Metaphor Theory. Even with the marker, the metaphoricity of these KIND metaphor expressions varies. On one hand, almost half of KIND metaphor expressions are near frozen, conventionalized, and thus embodied in our language and life. On the other hand, still the other half KIND metaphor expressions are more creative metaphors, which deserves our attention to investigate the creation of metaphors based on the marker and mapped concepts (and the abstractness).

6 Conclusion

This paper addressed the metaphoricity of expressions of a kind of marked metaphors, namely KIND metaphors, by proposing an identification guideline and a rating guideline for KIND metaphors. For the first research question, we have identified 245 KIND metaphor sentences, which account for 14% of the 1,780 sentences in the ‘is a kind of’ pattern, featuring a special distribution in metaphor research. We also discussed how metaphorical the expressions of KIND metaphors are. We reported that the average metaphoricity of the 176 sentences of KIND metaphors was 1.91 in a three-point Likert scale, with almost a half of near frozen metaphors and the other half of more novel metaphors. We also found some frequent nouns (as the concept B in the pattern of ‘A is a kind of B’) that occur with frozen KIND metaphors, which can be further exploited to identify more sentences of frozen metaphors.

In the future study, we will further examine the proportion of KIND metaphor expressions in the whole corpus. We will also investigate other patterns of KIND metaphors and the non-metaphorical expressions of these patterns. We will also extract cross-domain concepts of metaphors by linking the mapping concepts marked by the patterns of KIND metaphors to lexical knowledge resources.

Acknowledgments

This research was supported by the Ministry of Science and Technology (MOST 109-2811-H-004-503 and 109-2410-H-004-163), Taiwan.

References

Ahrens Kathleen & Chung, Siaw-Fong (2019). Metaphors in Chinese. In Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst (Eds.), *The Routledge Handbook of Applied Chinese Linguistics*. (pp. 364-378). New York and London: Routledge.

Chung, S.-F. (2009). *A Corpus-driven Approach to Source Domain Determination*. *Language and Linguistics Monograph Series*. Taipei: Academia Sinica.

Dunn, J. (2014). Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 745–751). Association for Computational Linguistics.

Goatly, A. (1997). *The Language of Metaphors*. Routledge.

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols*. Hackett.

Hanks, P. (2006). Metaphoricity is gradable. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based Approaches to Metaphor and Metonymy* (pp. 17–35). Berlin and New York: De Gruyter.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 202–251). Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G., & Turner, M. (1989). *More than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: University of Chicago Press.

Patterson, K. J. (2017). When Is a Metaphor Not a Metaphor? An Investigation Into Lexical Characteristics of Metaphoricity Among Uncertain Cases. *Metaphor and Symbol*, *32*(2), 103–117.

Porat, R., & Shen, Y. (2015). Imposed Metaphoricity. *Metaphor and Symbol*, *30*(2), 77–94.

Potts, A., & Semino, E. (2019). Cancer as a Metaphor. *Metaphor and Symbol*, *34*(2), 81–95.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*(1), 1–39.

Shih, M.-H., Chung, S.-F., Shen, Y.-H., & Liao, H.-C. (2020). A Study of KIND Metaphor Annotation based on Parsing and ConceptNet. In *the 21st Chinese Lexical Semantics Workshop*. City University of Hong Kong.

Skorczynska Sznajder, Hanna and Kathleen Ahrens. 2015. A corpus-based study of metaphor signaling variations in three genres. *Text and Talk*, 35(3): 359-381.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*, *21*(4), 765–796.

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*. Amsterdam: John Benjamins Publishing Company.

Vervaeke, J., & Kennedy, J. M. (1996). Metaphors in Language and Thought: Falsification and Multiple Meanings. *Metaphor and Symbolic Activity*, *11*(4), 273–284.

Wang, B. P.-Y., Lu, X., Hsu, C.-C., Lin, E. P.-C., & Ai, H. (2019). Linguistic metaphor identification in Chinese. In S. Nacey, A. G. Dorst, T. Krennmayr, & W. G. Reijnierse (Eds.), *Metaphor Identification in Multiple Languages: MIPVU around the world* (pp. 248–265). Amsterdam: John Benjamins Publishing Company.

袁暉. (1982). *比喻 [Metaphors]*. 合肥: 安徽人民出版社.

Latent Topic Refinement based on Distance Metric Learning and Semantics-assisted Non-negative Matrix Factorization

Tran-Binh Dang, Ha-Thanh Nguyen, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
{binhdang, nguyenhathanh, nguyenml}@jaist.ac.jp

Abstract

SeaNMF stands for the Semantics-assisted non-negative factorization. This approach is the current state-of-the-art method for topic modeling. In this study, we propose a new method *(i.e., DML-SeaNMF)* for improving the latent topic by utilizing the distance metric learning. The main idea is to iteratively learn the appropriate term-document and term-term relations based on extracted topics in the previous step. Our experiments show that the DML-SeaNMF outperforms the SeaNMF in evaluating based on the topic coherence and topic-based document classification accuracy on several datasets.

1 Introduction

With the development of the social network, the textual data repository is being enriched by a huge amount of informative posts, comments, and questions from the internet, by which, we can extract latent information and knowledge by using various text mining methods. Among them, topic modeling is a well-known problem. For short text data (posts, comments, and questions), there are some methods like biterm topic model (BTM)(Yan et al., 2013), LeadLDA (Li et al., 2016), etc. These methods used topic modeling variants to reduce the effects of sparsity issues on topic modeling. Miao (2017) proposed an approach that used deep neural network architecture for topic modeling. In another way, Non-negative matrix factorization(NMF) is a solution for topic modeling. Choo (2013) successfully applied this approach. Yan (2013) used factorize a symmetric term correlation matrix for topic model. In WWW 2018, the SeaNMF (Tian Shi et al., 2018) was proposed to learn topics from short texts. The model combines document-word relation and word-context relation as inputs. SeaNMF outperforms state-of-the-art methods for topic modeling such as LDA (Blei et al., 2003), NMF, PTM (Zuo et al., 2016), and GPUDMM (Li et al., 2016).

In this paper, we propose a novel method that incorporates the distance metric learning(DML) for refining latent topics extracted by the SeaNMF, which is called the DML-SeaNMF.

A proper topic is a cluster of words that share a common semantic. We suppose that in the topics extracted by the SeaNMF, there exists a subset of proper topics. Hence, we aim to refine non-proper topics by changing the input matrices of the SeaNMF that is based on the proper ones. To this end, we consider the most likely topic of each word w as the "soft label" of such a word. Besides, each row in the term-document matrix or the term-term matrix (semantic matrix) is a vector representation of each word, denoted by $\overrightarrow{v}_w$. Hence, we aim to learn the new representation of word $\overrightarrow{v'}_w$ based on $\overrightarrow{v}_w$, which satisfies that vectors representing words in the same topic are close in the vector space (they have the small Euclidean distance). That means the distance $d_{euc}(\overrightarrow{v'}_{w_i}, \overrightarrow{v'}_{w_j})$ is small if w_i and w_j have the same topic. We learn $\overrightarrow{v'}_w$ by using the Large Margin Nearest Neighbor. New vectors $\overrightarrow{v'}_w$ form the new term-document and term-term matrices that are input to the SeaNMF to revise current latent topics.

We compare the performance of our approach

with SeaNMF. The experimental results show that our proposed method significantly outperforms the SeaNMF regarding the following points: (i) the coherence of the topics; and (ii) the effectiveness of topic-based representation for document classification.

The outline of the paper is organized as follows. Section 2 presents the basic knowledge about SeaNMF and Distance metric learning. Section 3 presents an approach named DML-SeaNMF. Section 4 shows experiment results which we did. Finally, we conclude the content of the paper in section 5.

2 Background

2.1 Non-negative matrix factorization - NMF

Non-negative matrix factorization(Tian Shi et al., 2018) is a method that divides origin matrix to two sub-matrix, with the property that all three matrices have no negative elements. It is useful when analyzing objects which are high-dimensional data. In topic modeling, NMF is no less than the generative probabilistic model. With a corpus has N documents and the number of work/keyword in vocabulary is M, we will have a word-document matrix A. The column of A represents bag of words a document on vocabulary. Using NMF for this matrix A was created two lower-dim matrices W, H. Multiple of two matrices approximates by matrix A.

$$min_{W,H\geq 0}||A - WH^T||_F^2 \quad (1)$$

More detail, matrix A has size $M \ \ words \times N \ documents$. After factorizing, with K topic, we have two matrices. The matrix W represents the distribution of words in the topic. Each column is the presence of a topic in vocabulary. Size of W is $M \ \ words \times K \ \ topic$. The matrix H shows the distribution topic in documents. Each row is the latent topic space of documents. Size of H is $N \ documents \times K \ topic$.

2.2 Semantics-assisted NMF - SeaNMF

SeaNMF (Tian Shi et al., 2018) is the model based on Non-negative matrix factorization to discover topics from short texts. SeaNMF uses semantics information to implement the information in its learning process. The representation of semantic information explains pointwise mutual information(PMI) (Levy and Goldberg, 2014). SeaNMF uses two matrices as the input of the model: term-document matrix A and semantic correlation matrix S. The matrix S shows the relationship between keyword and their contexts (term - term relation). The objective function of SeaNMF is calculated as follow:

$$min_{W,W_c,H\geq 0}\left|\left|\begin{pmatrix} A^T \\ \sqrt{\alpha}S^T \end{pmatrix} - \begin{pmatrix} H \\ \sqrt{\alpha}W_c \end{pmatrix} W^T\right|\right|_F^2 \quad (2)$$

With input matrices and the number of topics K, the SeaNMF model has factorized to three output matrices W, W_c and H. The matrix W represents the distribution of words in the topic. Each column is the presence of a topic in vocabulary. The matrix H shows the distribution topic in documents. Each row is the latent topic space of documents. In SeaNMF, there is a new output matrix - W_c. The matrix W_c represents the word in context.

2.3 Distance metric learning

The distance metric uses distance function which provides a relationship metric between each element in the dataset. Traditionally, practitioners would choose a standard distance metric (Euclidean, City-Block, Cosine, etc.) using a priori knowledge of the domain. Distance metric learning (or simply, metric learning) is the sub-field of machine learning dedicated to automatically constructing optimal distance metrics.

Weinberger and Saul (2009) proposed Large margin nearest neighbor - LMNN as one of the most widely-used Mahalanobis distance learning methods. The method was designed to work with the nearest neighbor classifiers. It can help to improve the performance of the nearest neighbor classifier. LMNN works based on the proposed: label of samples will be more believed if nearest neighbors have the same labels.

Give a dataset: $X = \{x_1, x_2, x_3, ..., x_n\}$ and their labels: $Y = \{y_1, y_2, y_3, ..., y_n\}$. Consider three samples x_i, x_j, x_k: x_j is target neighbor of x_i, x_k is impostor.

$S = \{(x_i, x_j) : y_i = y_j; \ x_j \ is \ neighbor \ of \ x_i\}$

$R = \{(x_i, x_k) : y_k \neq y_i; \ x_k \ is \ neighbor \ of \ x_i\}$

With a sample x_i, sample x_j is a *target neighbor* of x_i if label of x_j is the same with label of x_i

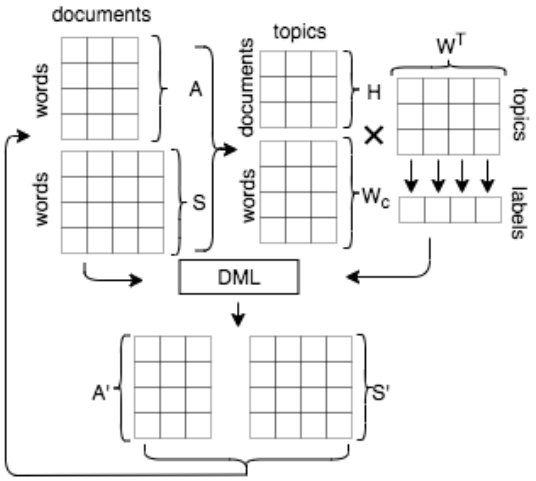


Figure 1: The proposed approach DML-SeaNMF (Example with a corpus: 3 documents, 4 words, 3 topics)

$(y_j = y_i)$ and x_j is k-nearest neighbor of x_i. After establishing the target neighbor, a perimeter is created by the distance of each sample in the dataset X. LMNN tried to learn a distance that no sample difference label in this perimeter. So, a margin is built based on the radius of the perimeter. Any sample of a different class that invades this margin will be called an *impostor*. Now, LMNN brings the target neighbor closer and try to keep impostors as far away as possible.

LMNN uses two penalties in learning process. The first one penalized distant target neighbors (ε_{pull}) and the second one penalized nearby impostors(ε_{push}). Combine two penalties above with parameter t controls the "pull/push" trade-off will create object function of LMNN:

$$\min \{(1-t)\varepsilon_{pull} + t\varepsilon_{push}\} \quad t \in [0,1] \qquad (3)$$

3 The propose method: Distance metric learning for SeaNMF

As mentioned above, the SeaNMF employs an unsupervised approach to effectively learn latent topics in short texts. To improve this method, we aim to iteratively refine topics learned by the SeaNMF as follow.

SeaNMF uses two matrices term-document A and semantic S are input. Term-document matrix A used bag-of-words to show the relationship between word and document. Semantic matrix S was built by the calculation of PMI - a measure of association. The initialization of two matrices is based on the corpus. And, the performance of SeaNMF model depends on the quality of the input matrices. So, we aim to learn better non-negative matrices A and S by a linear transformation f.

To this end, we propose a method that incorporates the SeaNMF with distance metric learning (DML) for topic refinement. The idea behind this method is that: (i) assign the most likely topic, obtained by the SeaNMF, for each word which is called a soft label for such a word; (ii) with the soft label and DML, learns parameter of Mahalanobis distance which is a transformation f. Thus, after refining, new input inherited the essence of latent topics from the previous step. And, new latent topics which learned in the next step is more clear and better.

With a corpus, matrix A and matrix S were built as the input of the SeaNMF model. After the process, three lower-rank matrices were born - W, W_c and H. Based on the result of SeaNMF, we can use the learned topic as a soft label for each word. They depend on SeaNMF's results. Let W be a matrix in which, each row in W represents the probability of a word with K topics in latent topic space. Thus, the soft label of each word was determined that the topic has the max probability in each row of W. Each word in vocabulary has a corresponding label.

Based on these soft labels, "Distance metric learning"(DML) process tries to learn transformation f by Large margin nearest neighbor. LMNN is an approach using nearest neighbor to improve the performance of clustering. In some cases, topic modeling is also considered as a clustering problem. So, LMNN can support the topic modeling method to increase quality. In our approach, dataset X of LMNN is the vocabulary of the corpus, each word is treated as a sample. However, each sample in the dataset has two representations corresponding with two matrices A and S: (i) a vector N-dimensions (N documents) with matrix A; (ii) a vector M-dimensions (size of vocabulary) with matrix S. And their labels Y are soft labels of each word. Through the distance metric learning process, there is a transformation matrix L for each representation. The matrix A' and matrix S' are transformed into metric space by:

$$A' = A \times L_1^T \; and \; S' = S \times L_2^T \qquad (4)$$

After transformation complete, A' and S' are checked non-negative condition and used as input of SeaNMF. The negative values in the matrix were replaced by 0.

This is the end of a time-step in a loop. The process will run with T time-steps. We expect to take three output matrices with their best state. With DML-SeaNMF, we designed a condition to solve the issue as follows: use measure evaluation of topic models - Topic coherence (David et al., 2010). This metric is calculated after the SeaNMF process to check. With time-step t and time-step $t-1$, if topic coherence of t is less than topic coherence of $t-1$, the learning process is stopped. At the time, three lower-rank matrices in time-step $t-1$ is the final output.

4 Experiments

4.1 Datasets

Experiments are conducted with the benchmark short text dataset. SeaNMF is a model fit with short text data. We used three datasets include:

- **TagNews**[1] The dataset is a part of TagMyNews dataset. It is news extracted from RSS feeds of popular newspaper websites. Categories are: Sport, Business, Entertainment, US, World, Health, Sci_tech.
- **Question 2002**[2] This dataset was used in learning question classification experiments of Xin Li, Dan Roth(2002). Data is public dataset.
- **StackOverflow**[3] The dataset used by Jiaming Xu et al.(2015) in VSM-NLP workshop NAACL 2015. It is questioned in StackOverflow through July 31st to August 14 2012.

Table 1 shows the static information of the three datasets, which we used in our experiments.

4.2 Evaluation metrics

Topic coherence (David et al., 2010) is typically an evaluation method for evaluating topic models. With topic k: After the models generates topic consisting of words, this metric is applied on the top n words of the topic. Given a topic k, PMI value is computed on this topic as described in (Tian Shi et al., 2018). Topic coherence is the average value of PMI on all of the topics.

$$TC_k = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le n} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (5)$$

where n is top-n words in topic k, $p(w_i, w_j)$ is the probability of word w_i, w_j co-occurring. $p(w_i)$, $p(w_j)$ is marginal probability of w_i, w_j.

$$Topic\ Coherence = \frac{\sum_{k=1}^{K} TC_k}{K} \quad (6)$$

The higher topic coherence, the better model. In our experiments, the number of the top words n is set to 10, the number of topics K set to 50.

The model runs with the stop condition described in Section 3. Then, the result is compared with the output of the SeaNMF model, which runs separately. We evaluate with two metrics: (i) max topic coherence: the largest value of topic coherence in time-steps and (ii) average topic coherence: the ratio between the sum of topic coherence values and number of time-steps. This experiment compares when we use or not use distance metrics learning.

Besides, we also use document classification performance to measure topic model effectiveness. Latent topics extracted from the models are used as features for a single fully connected layer neural network to perform classification. Training and testing data are randomly split with a ratio of 4:1. The quality is measured by three measures: precision, recall, and F-score.

Dataset	Docs	Terms	Avg doc-len	Labels
TagNews	1000	3505	7.77	7
Question 2002	1000	2837	8.61	6
StackOverflow	1000	2502	8.69	20

Table 1: Basic statistics of datasets in our experiments

		SeaNMF	DML-SeaNMF
TagNews	Max	2.671	3.374
	Avg	2.589	3.073
Question 2002	Max	2.445	3.183
	Avg	2.228	2.782
StackOverflow	Max	2.233	2.545
	Avg	2.187	2.389

Table 2: Topic coherence results with three datasets

[1] https://github.com/isthegeek/News-Classification
[2] https://cogcomp.org/Data/QA/QC/
[3] https://github.com/jacoxu/StackOverflow?

	TagNews	StackOverflow
LDA	2.023	0.675
NMF	2.426	1.035
SeaNMF	2.671	1.919
DML-NMF	2.819	2.009
DML-SeaNMF	3.374	3.11

Table 3: Topic coherence results on 5 methods : LDA, NMF, SeaNMF, DML-NMF, DML-SeaNMF In this experiments, we use StackOverflow dataset with 4000 samples

		SeaNMF	DML-SeaNMF
TagNews	Recall	0.36	0.39
	Precision	0.35	0.36
	F-score	0.35	0.37
Question 2002	Recall	0.55	0.56
	Precision	0.57	0.59
	F-score	0.55	0.56
StackOverflow	Recall	0.42	0.55
	Precision	0.43	0.56
	F-score	0.42	0.54

Table 4: Performance of SeaNMF and DML-SeaNMF in documents classification

4.3 Results and Discussion

In our experiments, we compared the performance of our method with SeaNMF. The topic coherence value is shown in table 2. With two metrics: max topic coherence and average topic coherence, DML-SeaNMF is better than SeaNMF on all of the three datasets. After 2-3 time-steps, DML-SeaNMF could find a better state of input matrices to return higher topic coherence value. The difference between the average performance of the two models is significant.

To analysis overview, we do more an experiment with datasets: TagNews and StackOverflow on 5 methods: LDA, NMF, DML-NMF, SeaNMF, and DML-SeaNMF. The result shown in table 3.

The document classification result is shown in table 4. On the three datasets, DML - SeaNMF all outperformed SeaNMF. The difference between the two methods is about 2-3% on TagNews and Question 2002. This number is the largest on StackOverflow (13%). That happened because StackOverflow samples often contain name entities that are identical with the class labels. Our model extracted this kind of features better than SeaNMF. The refinement of the latent topic helps the topic feature become more descriptive.

After learning latent topics on TagNews and StackOverflow datasets, we find similar topics obtained from DML-SeaNMF and SeaNMF based on the top-10 keywords. The list of the top-10 keywords in the selected topics obtained is shown in Table 5. As we can see, two topics for TagNews are about Sport and Japan news. The topic selected from StackOverflow related to Visual Studio.

In this paper's scope, we conduct experiments with short text datasets. However, our proposal is not limited to short text. Investigating and optimizing the method for long documents is also one of our possible future directions.

5 Conclusion

This paper presents a method to refine latent topics. Our method proposes the combination of Distance metric learning and SeaNMF. Large margin nearest neighbor(LMNN) is chosen to use in the learning distance process. LMNN takes latent topics as labels and sample is the word. This learning process creates a transformation matrix to update the input matrices of the topic model. We compared DML-SeaNMF with one of the state-of-the-art methods(SeaNMF) on three datasets. Experimental results showed that our model is effective when testing the benchmark data. In future works, we want to improve and extend this method especially on long documents.

References

Tian Shi, Kyeongpil Kang, Jaegul Choo and Chandan K. Reddy 2018, *Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations*, In Proceedings of the International Conference on World Wide Web (WWW) Lyon, France

Bac Nguyen, BernardDe Baets 2018, *An approach to supervised distance metric learning based on difference of convex functions programming*, Pattern Recognition Volume 81, pp. 562-574

Weinberger, Kilian Q and John Blitzer and Lawrence K. Saul, 2006, *Distance Metric Learning for Large Mar-*

Table 5: Discovered topic by DML-SeaNMF and SeaNMF

	TagNews				StackOverflow	
Category	Sport		Japan		Visual Studio	
	DML-SeaNMF 6	SeaNMF 34	DML-SeaNMF 42	SeaNMF 10	DML-SeaNMF 2	SeaNMF 2
Top 10 keywords	league	keeps	japan	japan	Visual	Visual
	basketball	winning	nuclear	street	Studio	Studio
	play	semi-finals	trust	wall	Window	project
	global	champions	crisis	nuclear	FreezingTFS	Code
	champions	roundup	government	worries	Might	Using
	semi-finals	nbc	shut	deals	screen	projects
	soccer	basketball	rescue	stocks	IFEnd	Can
	uconn	drought	reactors	dow	Refactoring	Keyboard
	winning	play-off	radioactivity	rescue	Structured	build
	fans	share	quake	quake	IntelliSense	Setup

gin Nearest Neighbor Classification, NIPS 2005, pp. 1473–1480, MIT Press

Shiming Xiang, Feiping Nie, Changshui Zhang 2008, *Learning a Mahalanobis distance metric for data clustering and classification*, Pattern Recognition Volume 41, Issue 12, pp. 3600-3612

Xin Li, Dan Roth 2002, *Learning Question Classifiers* COLING'02

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, Hongwei Hao, 2015, *Short Text Clustering via Convolutional Neural Networks* VSM-NLP workshop, NAACL

David M. Blei, Andrew Y. Ng, and Michael I. Jordan 2003, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 3:993–1022

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng 2013, *A Biterm Topic Model for Short texts*, In 22nd International World Wide Web Conference, WWW 2013 Rio de Janeiro, Brazil, pp. 1445-1456

Jing Li, Ming Liao, Wei Gao, Yulan He, and KamFai Wong, 2016, *Topic Extraction from Microblog Posts Using Conversation Structures*, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume1: Long Papers Berlin, Germany

Yishu Miao, Edward Grefenstette, and Phil Blunsom, 2017, *Discovering Discrete Latent Topics with Neural Variational Inference*, In Proceedings of the 34th International Conference on Machine Learning, ICML 2017 Sydney, NSW, Australia, pp. 2410–2419

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, 2013, *Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization*, IEEE transactions on visualization and computer graphics 19

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, 2015, *Weakly supervised nonnegative matrix factorization for user-driven clustering* Data Mining and Knowledge Discovery 29,pp. 1598–1621

Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang 2013, *Learning topics in short texts by non-negative matrix factorization on term correlation matrix*, In Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM, 749–757

Kilian Q. Weinberger and Lawrence K. Saul, 2009, *Distance Metric Learning for Large Margin Nearest Neighbor Classification*, Journal of Machine Learning Research 10 (2009) pp. 207-244

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong, 2016, *Topic Modeling of Short Texts: A Pseudo-Document View*, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2105–2114

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma, 2016, *Topic Modeling for Short Texts with Auxiliary Word Embeddings*, In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 165–174

Newman David, Lau Jey Han, Grieser Karl and Baldwin Timothy, 2010, *Automatic evaluation of topic coherence*, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108

Jacob Goldberger, Geoffrey Hinton, Sam Roweis, Ruslan Salakhutdinov, 2005, *Neighbourhood Components Analysis*, Advances in Neural Information Processing Systems 17, pp. 513–520, MIT Press

Levy, Omer and Goldberg, Yoav, 2014, *Neural Word Embedding As Implicit Matrix Factorization*, Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14

TDP – A Hybrid Diacritic Restoration with Transformer Decoder

DANG Trung Duc Anh

ducanhhbtt@gmail.com
Hanoi University of Science and Technology
Hanoi, Vietnam

NGUYEN Thi Thu Trang*

trangntt@soict.hust.edu.vn
Hanoi University of Science and Technology
Hanoi, Vietnam

Abstract

Diacritic restoration plays an important role in Natural Language Processing (NLP) for many diacritical languages such as Vietnamese, Czech, Hungarian, etc. With the development of deep neural network, this task could reach a good accuracy, i.e. F1-score is up to 97.7% for the state-of-the-art models. However, the output of these models can include meaningless syllables, the processing time is rather long especially with the sequence-to-sequence method and beam search. This task is a very first step in any text (pre-)processing, which can be a part of another application. Therefore, the processing time is extremely important. To balance both accuracy and time consuming, this paper proposes a novel hybrid method which includes a transformer decoder and a diacritic penalty layer. The transformer decoder is good enough for this problem since an input character only corresponds to exact one output character. The purpose of the penalty layer is to guide the model to produce only possible diacritic letters of the language. The experimental results on Vietnamese corpus show that the proposed model helps the predicting time reduce from about eight to ten times compared to the previous methods. Whereas, the accuracy of the proposed method is better than (i.e. 1%) or equal to the state-of-the-art sequence-to-sequence without or with beam search.

*Corresponding author

1 Introduction

Diacritics are a vitally important component in many diacritical languages such as Vietnamese, Czech, Hungarian, etc. However, a large number of texts without diacritics serve many purposes.

In diacritic languages, typing a diacritic word is far more troubled than typing a non-diacritic one. For instance, in Vietnamese, “đường” can be typed as "dduongwf" (Telex system). The middle or old-aged who do not know the rules for typing or those who want to save time typing diacritics would prefer to type sentences without diacritics, although they can be misleading and incomprehensible to other people. Moreover, many public-originated foreign systems only support non-diacritic characters, leading to a huge amount of this data type to be processed.

There are numerous ways for restoring a sentence to its former full diacritical marks with different meanings. For example, in Vietnamese, the most suitable restoration version for the non-diacritic one “Toi muon mo the tin dung” is "Tôi muốn mở thẻ tín dụng” (I want to open a credit card). However, each syllable in the sentence has multiple ways to become the form with full diacritic marks. "Muon" can be restored as "muốn" (want), "muộn" (late), "muôn" (many) while "the" can be restored as "thẻ" (card), "thè" (put out), "the" (a kind of traditional cloth), "thé" (high pitch), "thế" (and), "thề" (swear), "thê" (wife). This leads to a number of restoration combinations corresponding to the input sentence that we need to disambiguate. Therefore, recovering diacritics is among the most necessary but challeng-

ing problems in natural language processing. It is particularly hard for Vietnamese, whose ratio of diacritical words is highest, i.e. approximately 90%, 80% of which contain ambiguity (Do et al., 2013).

A number of researches on restoring diacritic marks used both machine learning and deep learning approaches. For Vietnamese, three main ones have been proposed, i.e. (i) rule and dictionary-based, (ii) machine learning-based (Nguyen and Ock, 2010) and (iii) deep learning-based approach (Hung, 2018; Náplava et al., 2018; Nga et al., 2019). A typical example of the rule and dictionary-based approach is VietPad[1]. In that work, a dictionary with all Vietnamese syllables was built to restore diacritic marks. However, this tool could not solve a number of ambiguous cases, leading a limited accuracy of about 60% to 85% depending on the domain. The machine learning-based approach (Nguyen and Ock, 2010) achieved an accuracy of 94.7% on their dataset, using a combination of AdaBoost and C4.5 algorithms. Recently, deep learning-based methods with machine translation models have emerged as the state-of-the-art solution to the problem of diacritic restoration. The idea of this method is to treat non-diacritic and diacritic texts as the source and target languages in the machine translation formulation. The best work in this approach used a novel combination of a character-level recurrent neural network-based model and a language model applied to diacritics restoration and reached the highest accuracy of 97.73% on Vietnamese (Náplava et al., 2018).

However, there are several shortcomings of the above state-of-the-art methods, i.e. producing nonexistent outputs and time-consuming for the task. Since the output is generated based on the possibility that the model predicts character by character, the sequence of output text may be nonexistent or meaningless in the language. Moreover, the diacritic restoration is a very-first step of text (pre-)processing for any NLP application. For instance, in question answering or chatbot systems, users sometimes input with non-diacritical marks which should be recovered before many next steps. Diacritic restoration is only a small step in any NLP application. Therefore, the restoration time of this task is hence extremely important in the industry.

In this paper, we propose TDP – a novel hybrid diacritic restoration model which retains the Transformer Decoder at the character-level with Penalty layer. The penalty layer is a restriction mechanism of possible diacritical letters for the output sequence. We have experimented the model for Vietnamese datasets with a promising performance in both accuracy and predicting time. The rest of the paper is organized as follows. Section 2 describes the language orthography and theory of the transformer model. Session 3 presents our proposed hybrid model for restoring diacritical marks. Section 4 discusses on the related works to the model and techniques in our model. In section 5, the experiments on Vietnamese data-sets are described and discussed. Finally, the paper draws some conclusions and perspectives of the work.

2 Background

Since we have experimented with Vietnamese, we provide in this section some backgrounds on Vietnamese orthography with diacritical features. We also present the full transformer model, parts of which are used to construct our model.

2.1 Orthography

In any diacritic language, a limited number of diacritic letters can be restored for a specific non-diacritic one.

Table 1: Possible Vietnamese diacritical letters

Non-diacritical Letter	**Possible Diacritical Letter**
a	a, á, à, ả, ã, ạ, ă, ắ, ằ, ẳ, ẵ, ặ, â, ấ, ầ, ẩ, ẫ, ậ
e	e, é, è, ẻ, ẽ, ẹ, ê, ế, ề, ể, ễ, ệ
i	i, í, ì, ỉ, ĩ, ị
y	y, ý, ỳ, ỷ, ỹ, ỵ
o	o, ó, ò, ỏ, õ, ọ, ô, ố, ồ, ổ, ỗ, ộ, ơ, ớ, ờ, ở, ỡ, ợ
u	u, ú, ù, ủ, ũ, ụ, ư, ứ, ừ, ử, ữ, ự
d	d, đ

In this paper, we describe the orthography of Vietnamese, which has the highest ratio of diacritical

[1] http://vietpad.sourceforge.net/

words among diacritical languages. Based on the Latin alphabet, there are 29 letters in Vietnamese alphabet including 11 vowels and 18 consonants. 22 letters of them are Latin letters (“f”,“j”, “w” and “z” are removed), and the rest are newly created ones (Đoàn, 2016).

Those new ones are the combination of four diacritics and the Roman alphabets (breve, inverted breve, horn, d with stroke) (Đoàn, 2016) . The 5 tone markings (acute, grave, hook, tilde and dot-below) are used to describe the tone of a syllable that can be marked on the vowel. In a word, diacritics in Vietnamese are put on all vowel letters and one consonant letter (d). Therefore, there are 22 input characters without diacritics, from which 89 characters with diacritics are inferred. The rules for converting from non-diacritic to diacritic letters are shown in Table 1. Letters not in the table should be ignored when restoring diacritics, i.e. ’b’, ’c’, ’g’, ’h’, ’k’, ’l’, ’m’, ’n’, ’p’, ’q’, ’r’, ’s’, ’t’, ’v’, ’x’.

2.2 Transformer model

Transformer model (Vaswani et al., 2017) is a type of neural network architecture developed to solve the problem of sequence transduction, or neural machine translation. It is built based on Seq2seq architecture, comprising an encoder and a decoder. The encoder takes the input sequence and maps it into a higher dimensional space using something like an abstract of the input. It is then fed into the decoder, where it is turned into an output sequence.

Before the appearance of transformers, the encoder and the decoder of the Seq2Seq model relied on gated recurrent neural networks (RNNs), such as LSTMs, with added attention mechanisms to handle the input and output sequences without fixed length and avoided gradient vanishing problem. However, the transformer model with only attention-mechanisms without any RNN facilitates more paralleling during training computations, which brings better results with less time for training. Transformers currently have become the state-of-the-art architectures in NLP.

Self-Attention and Multi-Head Attention (Vaswani et al., 2017). Self-attention can be described as mapping a query and a set of key-value pairs to an output. Query, key, and value vector are calculated by multiplying the input by query, key, value, trainable matrix respectively. Then query, key, value vector is fed into Scaled Dot-Product Attention. The detailed description is shown by the following equation:

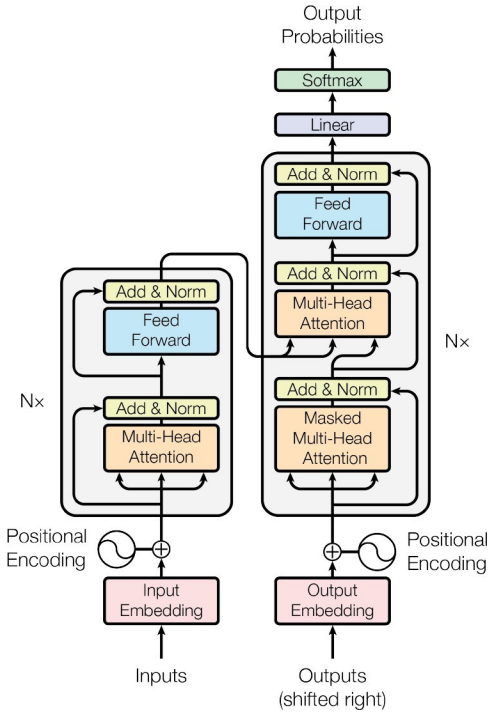


Figure 1: Architecture of transformer model (Vaswani et al., 2017).

Instead of calculating single attention at one time, we can calculate multiple attention in parallel, but each attention uses a different key, value, and query matrices. This technique is called multi-head attention. Each set of 3 key-value-query matrices is a head that pays “attention” to a certain piece of content of the input. The output of all heads will be concatenated together to form a complete vector output. Multi-head attention lets the model jointly attend to information from different representation sub-spaces at different positions.

Position embedding Because it does not use any CNN or RNN classes, the transformer needs to use a different way to handle the order of inputs, i.e. the effect of position encoding. Some information about the relative or absolute position of the tokens are injected into the input. Positional encoding can be learned or fixed. It has the same dimension as the embedding and the two are summed before being fed

into encoder or decoder block.

Model architecture The decoder and encoder are the main components of the transformer model, illustrated in Figure 1. Each part is a stack of the same blocks. Encoder block has two sub-layers: a multi-head self- attention mechanism, and a simple feed-forward network. A residual connection is deployed around each of the two sub-layers, followed by layer normalization. Similar encoder block, decoder block is built based on a multi-head attention and a feed-forward network. However, the decoder block inserts a third sub-layer to perform multi-head attention over the output of the encoder stack. Besides, the self-attention sub-layer is modified to make sure that the predictions for position i only depend on the known outputs at positions less than i.

3 Proposed model

We propose a novel model TDP (Transformer Decoder with Penalty layer) that only includes a transformer decoder at the character level with a penalty layer, whose architecture is illustrated in Figure 2. In this architecture, only the decoder blocks are kept instead of a full transformer model. As the origin full transformer, our model is a stack of 6 decoder blocks. With only the decoder, we can still solve the diacritic restoration problem since the length of the input is the same to that of the output, and an input character only corresponds to exact one output character. The encoder is redundant for this task. Moreover, the predicting time of the only decoder is expected to be much quicker than the full one.

When predicting, an output character corresponds to exact one input character in a position. Hence, we do not need to model the position in a separate layer. In self-attention, the memory keys and values come from the output of the previous decoder layer are used instead of that of the encoder. In the full architecture, the decoder has to be repeated every time step, corresponding to the number of input's characters. However, in the transformer decoder, we do not need to repeat the decode every time step. We can ignore the masking step of the decoder and only run it once. As the result, the predicting time of our model is expected to be reduced about x times compared to the full one, whereas x is the number of input's characters.

As mentioned in Table 1, each input character only has a specific number of output characters. For example, with the input 'i', the output can only be one of the six characters 'í', 'ì', 'ị', 'ĩ', 'ỉ', 'i'. If the input is a consonant like 'g' the output of the model must only be 'g'. Therefore, we propose a penalty layer which restricts the output with only some possible values of the input letter. This layer first looks up from a diacritic conversion dictionary and then calculates a penalty matrix for input characters. The penalty layer is executed in parallel to the decoder model, and then the penalty matrix will be added to the decoder output matrix to force the output character to be one of possible values, illustrated in Equation 1. This mechanism ensures that the output will not produce strange characters for the input. For example, when the user enters the word "co", the model sometimes predict to "ca". This issue can be addressed by the penalty layer.

$$Output = argmax(DecoderOutput + Input * PenaltyMatrix)$$

The penalty matrix works like an embedding matrix, each row of which is a penalty vector corresponding to a non-diacritic character. If the input character can be converted to the output character, the scalar at that corresponding position is 0; otherwise, it is an extremely negative number:

$$PenaltyMatrix_{i,j} = \begin{cases} 0 & \text{if input charater i can convert to j} \\ -\infty & \text{if input charater i can't convert to j} \end{cases}$$

We can feed the input as a sequence of syllables instead of characters with the expectation of reducing the processing time and making the input more meaningful. However, the amount of input and output vocabulary turned to be immense. That makes it complicated to guide the output of the model following the diacritical rules of the language. With the character-level approach, the vocabulary of diacritic conversion is small hence save much more time to construct the penalty matrix from the input. The penalty layer using vectorization makes the calculations much simpler and faster than using directly diacritic rules for the input and output sentences.

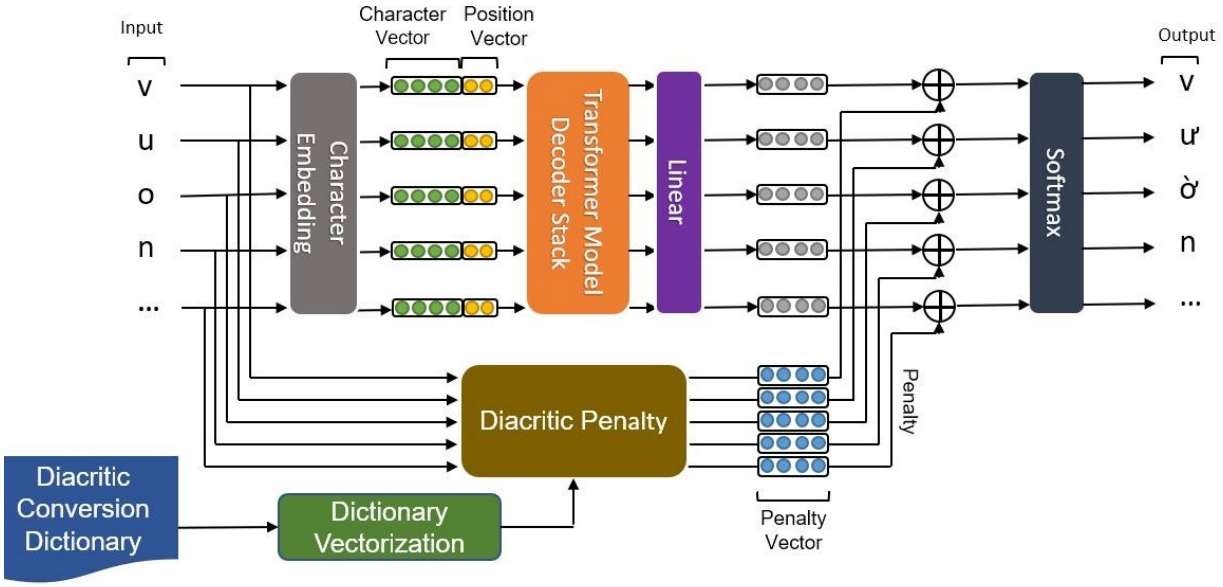


Figure 2: Architecture of the proposed hybrid model for diacritic restoration.

4 Related works

4.1 Transformer model

Since 2017, when the transformer was introduced in (Vaswani et al., 2017) , it has become the basic building block of most state-of-the-art architectures in NLP. In the original paper, transformer models achieve a new state of the art on both WMT 2014 English-to-German and WMT 2014 English-to- French translation tasks with a small fraction of the training costs of the best models from the literature. To solve diacritic restoration problem, the Transformer word-based model is only applied in Yorùbá Language (Orife, 2018) . When compared to the other methods mentioned in the paper (Orife, 2018) , the transformer model outperforms with word accuracy 95.4%, 5.3% higher than the second method on the test set.

4.2 Transformer decoder

In many NLP problems, instead of using the full transformer model, people only use the decoder part. This architecture was first used in generating English Wikipedia articles by summarizing long sequences (Liu et al., 2018). To handle this problem, they propose a two-state method. First, they use extractive summarization to coarsely identify salient information. Then, they use a neural abstractive model to generate the article. Authors affirmed that monolingual text-to-text tasks redundant information is re-learned about language in the encoder and decoder, so they only use the decoder. Their experiment showed that their model with decoder-only for the abstractive stage could handle very long input-output examples, better than using both traditional encoder-decoder architectures and recurrent neural network (RNN).

In addition, the transformer decoder is also used to create pre-trained language models such as GPT model (Radford, Narasimhan, et al., n.d.; Radford, Wu, et al., n.d.). There have been two versions of the GPT model released. GPT is generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine- tuning on each specific task. Both of two versions achieve great results in NLP tasks. GPT-1 improves the state of the art on 9 of the 12 datasets they study. GPT-2 is a direct scale-up of GPT-1, with more than 10X the parameters (1.5B) and trained on more than 10X the amount of data (40 GB text data). Due to author's concerns about malicious applications of the technology, they only release a much smaller model for researchers instead of the trained model.

To the best of our knowledge, this is the first time the transformer decoder is used for the diacritic

restoration task.

4.3 Hybrid method

When solving NLP problems, in order to improve the quality of neural networks, people often combine neural networks with different techniques, such as traditional machine learning models and rule-based models. A hybrid data-model parallel approach (Ono et al., n.d.) was used for reducing training time of sequence-to-sequence machine translation model. In abstractive summarization, to get better performance at content selection of neural network-based method, the work in (Gehrmann et al., 2018) combined a standard neural model with adata-efficient content selector to over-determine phrases in a source document that should be parts of the summary. Furthermore, the hybrid method requires much fewer data to train, which makes it more adaptable to new domains.

In our proposed model, we combine a transformer decoder with a diacritic penalty layer which restricts the output with all possible values corresponding to the input. This guides the model more accurate, reduces training time and gives reasonable outputs.

5 Experiment

In this section, we present some experiments for our proposed model with a Vietnamese corpus.

5.1 Dataset

The test set that we have to work in this paper is from banking domain. It includes 8,000 sentences.

To enhance the training set of this domain (i.e. 25,000 sentences), we retrieve more from Internet newspapers[2] . This corpus contains approximately 29 Gb of Vietnamese text files and approximately 160 millions of Vietnamese sentences. Nonetheless, due to a number of loan words and wrong spelling in that corpus, we only keep about 7% of sentences based on their types (i.e. interrogation, exclamation and affirmation) and constituent words. We use a Vietnamese dictionary VCL[3] as a filter. The final dataset contains about 11 millions of Vietnamese sentences. The valid set consists of 5,000 randomly selected sentences from the banking training data and 25,000 sentences from Internet newspapers. The rest is used for training.

All data in the corpus contain diacritics, which is the output that the model has to predict. The input is sentences after being stripped off all diacritics.

5.2 Evaluation method

Many input characters have only one candidate output, so the high character accuracy does not prove that the model works well. Therefore, although the model is at character level, we use evaluate the model at syllable level:

$$Accuracy = \frac{\#CorrectPredictedSyllable}{\#TotalPredictedSyllable}$$

5.3 Training

For the Transformer decoder, we reuse most of the hyperparameters proposed in (Vaswani et al., 2017). The decoder is composed of a stack of N = 6 identical blocks and each block contains 8 multi- head attention, but the dimension d model is 128 instead of 512 because of a small character vocabulary. The problem of recovering diacritics does not require the use of context too far, so we set the maximum sentence length to be 60 characters to avoid padding too long and save predicting time. Sentences longer than 60 characters will be broken down into sections of 60 characters, between which there will be an overlapping part with offset length = 10 characters.

Our model is trained on the hardware with the configuration as follows: 01 Tesla V100-PCIE-32GB GPU, Intel(R) Xeon(R) Silver 4210 CPU, 120GB RAM. We set the batch size=128 and use the default optimizer proposed in (Vaswani et al., 2017). The model converges after about 5 days. We evaluate the final model obtained by taking the average of the last 5 checkpoints.

5.4 Result

To compare with our model, we retrain the model architecture proposed in (Náplava et al., 2018) on our dataset. We compare this previous seq2seq model with or without beam search.

The result is shown in Table 2. Our TDP model with only transformer decoder and a penalty layer receives a better accuracy (i.e. 1.53%) and 16 times faster than the full transformer one. Compared to the previous seq2seq model, although the results were slightly lower than the model that used beam search

[2]https://github.com/binhvq/news-corpus#full-txt-v2

[3]https://vlsp.hpda.vn/demo/?page=vcl

(0.46%), the predicting time was reduced by approximately 10 times in both cases using CPU or GPU. Our model is 1% better and about 8 times faster than the one without beam search.

Table 2: Experimental results for hybrid diacritic restoration model. The prediction is executed on Tesla V100-PCIE-32GB GPU, Intel(R) Xeon(R) Silver 4210 CPU, 120GB RAM

Model	**Word accuracy (%)**	**Predicting time**	
		GPU(s)	CPU(s)
Seq2Seq re-run (Náplava et al., 2018)	97.52	0.372	0.423
Seq2Seq + Beam search re-run (Náplava et al., 2018)	98.83	0.433	0.533
Transformer model (full)	96.84	0.904	0.896
TDP model (our model)	98.37	0.043	0.055

To further evaluate how the model works in practice, we have performed an error analysis by statistically reporting the cases in which the model predicts incorrectly. The words which are wrongly predicted the most are listed in the table 3 below. The results show that most of mispredicted words are the ones that appear frequently in the banking domain but rarely appear in the others.For example, the word "thẻ" (card), "dùng"(use), "hủy" (cancel), "khóa" (key or stop), "vay" (loan), "lãi"(interest), ect, despite being small in number, are important words for the conversation. Inaccurate diacritic restoration of those words can lead to complete change of sentence meaning. For instance, the sentence "toi muon dung dich vu nay" can be restored to "tôi muốn dùng dịch vụ này" (I want to use this service) or "tôi muốn dừng dịch vụ này"(I want to stop this service), which are of diametrically opposite meanings. Therefore, it is essential that the domain adaptation technique be adopted in the future to bring enhanced efficiency to the industry.

Table 3: The most incorrectly predicted syllables

Expected output	**Number wrong predict**	**Confused with**
Thẻ	120	Thể, thế
dùng	87	Đúng, dụng, dừng, đứng
Hủy	59	huy
khóa	52	khoa
bạn	51	Bán, bàn, bẩn
thế	51	Thẻ, thể
vậy	44	vay
lãi	43	lại

6 Conclusion

In this work, we propose a hybrid diacritic restoration model TDP which includes a transformer decoder model and a diacritic penalty layer. The transformer decoder can solve this problem since an input character only corresponds to exact one output character. The only decoder also helps to decrease much predicting time since it does not need to repeat every time step. The purpose of the penalty layer is to guide the model to produce only possible diacritic letters of the language. The experimental results on a Vietnamese corpus show that our model TDP with only transformer decoders and a penalty layer helps the predicting time reduce from about eight to ten times compared to the state- of-the-art method. Whereas, the accuracy of the proposed method is better than (i.e. 1%) or equal to the sequence-to-sequence without or with beam search. Although the accuracy is quite high, the model wrongly predicts some important words in banking domain, e.g. “thẻ” (card) to “thế” (and), “mượn” (borrow) to “muốn” (want)... In the future, we will work on domain adaptation to solve this problem. In addition, we also consider using language model to improve the quality of the model.

References

Nguyen, K.-H., & Ock, C.-Y. (2010). Diacritics restoration in vietnamese: Letter based vs. syllable based model (B.-T. Zhang & M. A. Orgun, Eds.). In B.-T. Zhang & M. A. Orgun (Eds.), *PRICAI 2010: Trends in artificial*

intelligence, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-642-15246-7_61

Do, T. N. D., Nguyen, D. B., Mac, D. K., & Tran, D. D. (2013, August). Machine translation approach for vietnamese diacritic restoration, In *2013 international conference on asian language processing*. 2013 International Conference on Asian Language Processing. https://doi.org/10.1109/IALP.2013.30

Đoàn, T. T. (2016). *Ngữ âm tiếng việt* [Accepted: 2017-10-09T02:27:47Z]. H. : Đại học Quốc Gia Hà Nội. Retrieved September 12, 2020, from http://repository.vnu.edu.vn/handle/VNU_123/59688

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762 [cs]*, arxiv 1706.03762. Retrieved June 26, 2020, from http://arxiv.org/abs/1706.03762

Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-up abstractive summarization. *arXiv:1808.10792 [cs]*, arxiv 1808.10792. Retrieved July 3, 2020, from http://arxiv.org/abs/1808.10792

Hung, B. T. (2018). Vietnamese diacritics restoration using deep learning approach. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. https://doi.org/10.1109/KSE.2018.8573427

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. *arXiv:1801.10198 [cs]*, arxiv 1801.10198. Retrieved July 1, 2020, from http://arxiv.org/abs/1801.10198

Náplava, J., Straka, M., Straňák, P., & Hajič, J. (2018, May). Diacritics restoration using neural networks, In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan, European Language Resources Association (ELRA). Retrieved June 26, 2020, from https://www.aclweb.org/anthology/L18-1247

Orife, I. (2018). Attentive sequence-to-sequence learning for diacritic restoration of yor\'ub\'a language text. *arXiv:1804.00832 [cs]*, arxiv 1804.00832. Retrieved June 27, 2020, from http://arxiv.org/abs/1804.00832

Nga, C. H., Thinh, N. K., Chang, P.-C., & Wang, J.-C. (2019, December). Deep learning based vietnamese diacritics restoration, In *2019 IEEE international symposium on multimedia (ISM)*. 2019 IEEE International Symposium on Multimedia (ISM). https://doi.org/10.1109/ISM46123.2019.00074

Ono, J., Utiyama, M., & Sumita, E. (n.d.). Hybrid data-model parallel training for sequence-to-sequence recurrent neural network machine translation, 9.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). Improving language understanding by generative pre-training, 12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). Language models are unsupervised multitask learners, 24.

Construction of a VerbNet style lexicon for Vietnamese

HA My Linh
University of Science,
Vietnam National University
Hanoi, Vietnam
hamylinh@hus.edu.vn

LE Van Cuong
University of Social
Sciences and Humanities
Hanoi, Vietnam
cuongle.ussh@gmail.com

NGUYEN Thi Minh Huyen
University of Science,
Vietnam National University
Hanoi, Vietnam
huyenntm@hus.edu.vn

Abstract

Lexical resources like VerbNet (Kipper et al., 2006) or similar lexicons play an important role in the applications involving semantic understanding. For Vietnamese, the currently available computational lexicon (VCL) includes morpho-syntactic information for each lexical entry, subcategorization frames and also some semantic constraints for each verb. However, the information related to verb meaning and behaviors is still very far from complete. In this paper, we present our work on the construction of a VerbNet style lexicon for Vietnamese, called *viVerbNet*, in order to make available a fundamental lexical resource for semantic analysis of Vietnamese language. Each verb entry in viVerbNet is extracted from VCL, and enriched with various information acquired automatically from Vietnamese corpora as well as manually from a comparative investigation of the English VerbNet. At the current stage, we have built semantic components for 50 verb groups resulted from the application of a clustering algorithm on Vietnamese verbs.

1 Introduction

Lexicon are amongst the most important linguistic resources for natural language processing (NLP). Depending on what kind of applications for which the lexicon is developed, the annotation of lexical entries can be one of the most time-consuming and laborious tasks. For an application at a deep level like semantic understanding, a computational lexicon should ideally includes information related to the meanings and the behaviors of each word.

For English, there are several projects carried on lexical resources, e.g. WordNet (Miller et al., 1990), FrameNet (Baker et al., 1998), and more recently VerbNet (Kipper et al., 2006). These resources, together with semantically annotated corpora constitute important linguistic resources for developing language understanding applications. To obtain the semantic representation of a sentence, we usually need to identify the predicate and its arguments in that sentence. A predicate can be a verb, an adjective, or even a noun. The most important and complicated class of predicates is verb. For this reason, we are particularly interested in VerbNet, as it provides us with rich syntactic and semantic patterns of each verb, which proved very useful in semantic role labeling (Shi and Mihalcea, 2005), (Giuglea and Moschitti, 2006), (Loper et al., 2007), or in verb sense disambiguation (Brown et al., 2011), (Kawahara and Palmer, 2014).

In this paper, we present our work on the construction of a VerbNet style lexicon for Vietnamese, called *viVerbNet*, in order to make available a fundamental lexical resource for semantic analysis of Vietnamese language. Each verb entry in viVerbNet is extracted from VCL (Nguyen et al., 2006), and enriched with various information acquired automatically from Vietnamese corpora as well as manually from a comparative investigation of the English VerbNet. A clustering technique is applied to obtain classes of verbs sharing semantic and syntactic behaviors. For each verb class, we investigate their characteristics basing on annotated corpora as well as a comparative study of the corresponding English verb class.

The paper is structured in 4 main sections as follows. Section 2 presents the related works inspiring our project. Section 3 introduces the workflow for the construction of viVerbNet. Section 4 shows the application of a clustering technique for acquiring verb classes. Finally, Section 5 discusses the specifications of each verb.

2 Related Works

As mentioned above, a number of large and meaningful lexical resources have been built for semantic processing of English language.

- Wordnet (Miller et al., 1990) is a large English lexical resource in which words are organized into groups of synonym senses called synsets. Synsets are linked with each other by means of conceptual-semantic and lexical relations. WordNet is a dominant lexicon useful for sense resolution and semantic tagging.

- FrameNet (Baker et al., 1998) is a database containing more than 13,000 lexical units accompanied by their semantic frames. Over 200,000 manually annotated sentences with more than 1,200 semantic frames in FrameNet provide a training dataset for many applications such as machine translation, sentiment analysis, information extraction, etc.

In the following, we will present in more detail VerbNet, another important lexical resource in which verbs are fully syntactically and semantically annotated. VCL (Nguyen et al., 2006), the only Vietnamese large-scale computational lexicon, will be equally introduced as a foundation for building viVerbNet.

2.1 VerbNet

VerbNet (Kipper et al., 2006) is the largest English verb network, linking syntactic and semantic types of more than 5,200 verbs and 237 verb classes. This hierarchical verb vocabulary is mapped directly to other resources such as WordNet, FrameNet, and PropBank (Kingsbury and Palmer, 2002). Verb classes in VerbNet are designed based on Levin's verb classification (Levin, 1993). An example of a verb class is shown in Table 1.

Table 1: A class in VerbNet

Class Put-9.1	
Roles & Restrictions	Agent [+*animate*] Theme [+*concrete*] Destination [+*location & -region*]
Members	arrange, emplace, immerse, implant, lodge, ...
Frames:	
Description	NP V NP PP.destination
Example	I put the book on/under/near the table.
Syntax	Agent V Theme *{{+loc}}* Destination
Semantics	motion(*during(E), Theme*) not(*Prep(start(E), Theme, Destination)*) Prep(*end(E), Theme, Destination*) cause(*Agent, E*)

A verb class in VerbNet is defined by a set of members, the thematic roles and selectional restrictions of the arguments subcategorized by these members, as well as the syntactic and semantic descriptions related to their frames.

2.2 Vietnamese Computational Lexicon (VCL)

The Vietnamese Computational Lexicon (VCL) is the only large-scale lexical resource for fundamental tasks of NLP. VCL (Nguyen et al., 2006) contains about 42000 lexical entries, structured following the Lexical Mark-up Framework (LMF) - an abstract meta model from ISO TC 37/SC 4 (Francopoulo et al., 2006) that provides a framework for the development of NLP oriented lexicons. This lexicon includes all the information (word senses, part-of-speech, definition, examples) from one of the best Vietnamese print dictionaries (Hoàng, 2003). In addition, each entry is described in three aspects: morphology, syntax, and semantics. As Vietnamese words are morphologically invariable, the morphological information in VCL is only related to the word formation: a word can be either single, or compound, or redoubled, otherwise it can be a loan word, or an abbreviation, or a symbol. Table 2 shows an example of an entry in VCL, illustrating the informa-

tion not only at morphology level, but also at syntactic and semantic levels.

Table 2: A meaning of word "yêu" (*love*) in VCL

yêu (love)		
Morp	*simple word*	
Syntactic	Category	*V*
	Subcategory	*Vt*
	FrameSet	*Sub+V+Dob*
	Before	*R: rất (very)*
Semantic	Logical constraint	Categorial Meaning: *Emotion*
		Antonym: *Ghét (hate)*
	Semantic constraint	Sub: *Agt{Person}*
Definition	có tình cảm dễ chịu khi tiếp xúc với một đối tượng nào đó, muốn gần gũi và thường sẵn sàng vì đối tượng đó mà hết lòng	
Example	tôi yêu mẹ (*I love mom*)	

VCL is a very useful lexical resource for the fundamental NLP tasks. Its design allows easy update and extension, as well as a good exchangeability with other languages. Regarding the verbs, VCL still has some limitations in comparison with VerbNet as presented below.

- VCL contains 6652 verbs (8689 senses) and a total of 20 subcategorization frames associated to these verbs. But this information is far from complete: Most verbs are only attached to one frame, and the information about each frame is usually incomplete syntactically and semantically.
- VCL makes use of a set of 16 semantic roles such as: Agent, Experiencer, Possessor, *etc*. This set is quite limited compared to about 30 semantic roles in VerbNet.
- The semantic and logical constraints were manually built, however it remains several cases which have not been covered in the lexicon.

From these observations, we choose to build a VerbNet style lexicon for Vietnamese based on the verb entries available in VCL, in enriching them with other sources of information. This new lexicon is called viVerbNet.

3 Building viVerbNet

In order to acquire an equivalent resource to VerbNet for Vietnamese, we noted the need to revise and gather additional information such as thematic roles, selectional and syntax restrictions, syntactic frame, and semantic predicate for the verbs present in VCL.

As examples, we inspected in detail the thematic roles and components of a transitive verb (viết - *write*), an intransitive verb (đi - *go*), and an emotional verb (yêu - *love*) from the VCL and compared them to their translations in VerbNet. Some observations are made as follows.

- The semantic roles used in VCL are not equivalent to these in VerbNet.
  - VCL uses the semantic role *Content*, which is specialized into more concrete roles in Verbnet: topic, cause, goal, *etc*.
  - In many cases, the semantic roles are not defined in a similar way within the same context. For example, in VCL the subject argument of the verb "yêu (*love*)" is specified as *Agent* {Person}, while in VerbNet it is labeled as *Experiencer*.
- The selectional restrictions for semantic roles in VCL are quite incomplete.

Consequently, from VCL we cannot reconstitute the verb classes and the accompanied descriptions comparable to these in VerbNet. To build viVerbNet, we need to find a way for classifying verbs into groups of verbs having similar behaviors and describe these behaviors of each class. A clustering method applied on a large corpus can be useful for identifying the verb classes. In addition, we should revise the definition of the semantic roles and the selectional restrictions in VCL, insuring their compatibility with the same concepts in VerbNet.

Annotated corpora are equally important resources for extracting the specifications of each verb class:

- Viettreebank (Nguyen et al., 2009) is a constituency treebank with over 10,000 sentences. Subcategorization frames of several verbs can be extracted from this corpus. More detailed information can be equally extracted from a subset of this corpus: the Vietnamese dependency treebank (Nguyen et al., 2013).

- The Vietnamese Propbank (Ha et al., 2015) contains over 5000 sentences from VietTreeBank with labeled semantic roles compatible with the English Propbank. The semantic role labels in this corpus can be used to specify thematic roles for verbs in viVerbNet.

Beside the information acquired from corpora, an investigation of similar verb classes in the English VerbNet helps to determine comparable specifications of Vietnamese verb classes.

Figure 1 shows the summary of our workflow for constructing viVerbNet. In the next sections, we will describe in detail about the clustering of Vietnamese verbs and the specifications of the syntactic and semantic components for each verb group.

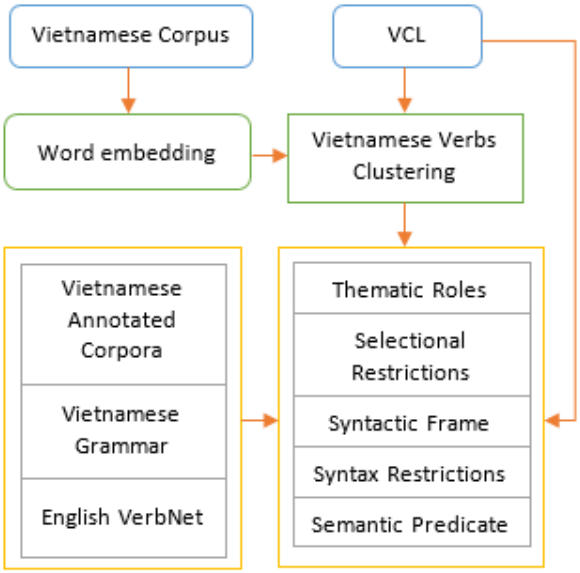


Figure 1: viVerbNet construction

4 Clustering Vietnamese Verbs

4.1 Clustering Method

In order to identify verb classes for Vietnamese, we first extract all the verbs from VCL, then apply an automatic clustering method on these verbs based on a large Vietnamese corpus.

In this paper, we use a hierarchical clustering algorithm (HCA) (Day and Edelsbrunner, 1984) to cluster 6652 verb entries extracted from the VCL dictionary. For any clustering algorithm, the most important questions are the data representation and the similarity measure. In our case, each verb is represented by a vector generated by word embedding models and we use the cosine distance as similarity measure.

Word embedding models. We have experimented with two different word embedding models generated by the word2vec algorithm (Mikolov et al., 2013): one (called *word2vec1*) is trained[1] on our own corpus, the other (called *word2vec2*) is pretrained in (Vu et al., 2019). In addition, we also experimented another pre-trained BERT word embeddings called PhoBERT (Nguyen and Nguyen, 2020), which is currently the best language model for Vietnamese in several problems like POS tagging, dependency parsing, and named-entity recognition.

4.2 Experimentation Results

From VCL, we have extracted 6652 verbs with a total of 8689 senses. The *word2vec1* model, trained on a 379 MB Vietnamese word-segmented corpus containing 99,531 articles collected from two news websites, covers 6039 verbs. The *word2vec2* model, trained on a 7.1 GB Vietnamese news corpus, covers 6388 verbs. PhoBERT, trained on a 20 GB corpus including 1 GB from Wikipedia, and 19 GB from various news websites, has been applied on our corpus to generate the word vectors of 6039 verbs.

Using the HCA algorithm on these three sets of verb embeddings, we get the best number of clusters given by the Silhouettes measure (Rousseeuw, 1987) for our data which is 283. However, for a finer clustering of verbs, we choose 1000 as the number of clusters. As the word embeddings take in account the context of each word, verbs in the same cluster

[1] using Gensim library `https://radimrehurek.com/gensim/models/word2vec.html`

have close meaning and similar behaviors and usage. Consequently these clusters can be served for defining verb classes for viVerbNet.

The experimentation show that two models word2vec1 and word2vec2 give similar results, which is explainable as the two models are trained with the same algorithm and the same style of news corpus, even if our corpus is much more smaller. In the meanwhile, the PhoBERT model gives more different results as it allows to distinguish different senses of a same verb. Some examples resulted from three word embedding models are given in Table 3. These example clusters contain verbs with very close meaning (for PhoBERT we show only the verbs present in the two other models).

Table 3: Some verb cluters

Verbs meaning "give-birth"		
word2vec1	**word2vec2**	**PhoBERT**
đẻ	đẻ	đẻ
sinh_nở	sinh_nở	sinh_sản
sinh_đẻ	sinh_đẻ	sinh
chuyển_dạ	chuyển_dạ	
thai_ngén	thai_ngén	
Verbs meaning "die"		
word2vec1	**word2vec2**	**PhoBERT**
bị_thương	bị_thương	đi
thiệt_mạng	thiệt_mạng	thiệt_mạng
mất_tích	mất_tích	
tử_nạn	tử_nạn	
chết	chết	chết
chết_đuối	chết_đuối	
bỏ_mạng	lâm_nạn	
mắc_kẹt	thương_vong	

A detailed evaluation of the results is being undertaken in investigating the behaviors of each verb in the available annotated corpora mentioned in Section 3. We equally proceed to study syntactic and semantic descriptions of verb classes, in choosing the clusters of verbs having the same meaning.

5 Verb Specification in viVerbNet

In this section, we present the components of VerbNet and focus on the discussion of specifications of Vietnamese language in comparison with English. This comparative study help us to build the components for viVerbNet in a compatible schema with the English VerbNet.

5.1 Thematic Roles

The semantic roles describe the basic semantic relationships between predicates and their arguments. We rely on 24 semantic roles from the Vietnamese Propbank to build thematic roles for viVerbNet. As this PropBank is designed in assuring the compatibility with the English ProbBank, the mapping from these 24 semantic roles to 39 themantic roles of VerbNet is facilitated.

The examples below show different syntactic distributions of the verb "`hoàn_thành`", belonging to a verb class comparable to the VerbNet verb class "*complete*". The contents inside the brackets following a word include the English translation and/or the role of that word. For all examples from now on, V stands for VERB.

- Active voice: `Tôi [`*I*`/Agent] đã [`*temporal marker*`] hoàn_thành [`*finish*`/V] bài_tập [`*exercise*`/Patient]` (*I have finished the exercises*).
- Passive voice: `Bài_tập [Patient] đã được [`*passive marker*`] tôi [Agent] hoàn_thành [V]` (*The exercises have been finished by me*).
- Passive voice without agent: `Bài_tập [Patient] đã hoàn_thành [V]` (*The exercises have been finished*).

We can see that Vietnamese has the same basic syntactic order Subject-Verb-Object as English in active voice. Attention should be paid to the passive voice, where the grammatical calque can produce the passive sentence "`Bài_tập [Patient] được [`*passive marker*`] hoàn_thành[V] bởi [`*by*`] tôi[Agent]`", but it sounds unnatural and is rarely used in good practice. In the case of passive voice

without agent, the passive marker can be absent as shown in the example. For this reason, the category and position of a word in a sentence are not enough for identifying its semantic role. That proves the importance of the selectional restrictions associated to each role.

5.2 Selectional Restrictions

Selectional restrictions determine semantic constraints on semantic roles. These restrictions indicate the existence (+) or absence (-) of semantic attributes such as [concrete], [animate], [organization], *etc*. Logical operators (| (OR) and & (AND)) are used to combine multiple restrictions.

For example, the selectional restrictions of the verb cluster "cấm, đình_chỉ, hoãn, nghiêm_cấm" corresponding to the VerbNet verb class "*forbid-64.4*" are as follows:

```
Agent[+animate|+organization]
Theme[]
Recipient[+animate|+organization].
```

For the sake of interoperability, we mapped 75 semantic classes used for semantic constraints to the set of 37 selectional restrictions in VerbNet.

5.3 Syntactic Frames and Restrictions

Each verb is associated to one or more syntactic frames. A syntactic frame briefly describes the surface structure of sentence constituents. It also specify semantic roles around verbs and syntax restrictions expressing the constraints on sentence constituents associated to these roles, such as plural, sentential, *etc* as illustrated in the following patterns:

1. `Agent V Patient<+plural>`
2. `Pivot V Theme <+np_to_inf>`
3. `Agent V Theme <+sc_ing>`
4. `Pivot V Theme <+ac_ing>`

The first pattern shows the restriction on the number of the patient role (plural), as in the sentence "*The merger associated the two companies*". In Vietnamese, the plural number is expressed by function words for plural markers like `những, các` or by numeral nouns:

company - `công ty`;
companies - `các công ty`;
two companies - `hai công ty`.

The second pattern covers this kind of sentence "*I needed him to go.*", while the third pattern corresponds to the syntactic structure in "*He rehearsed singing the song.*". The sentence "*I need him cooking.*" is an example of the fourth pattern.

Regarding two last patterns related to the gerund construction V_ing in English, it is worth to emphasize the phenomenon of nominalization in Vietnamese. In Vietnamese, we can observe two types of verb nominalization. The first type consists of the verb-noun categorical mutation, where a verb and its verbal noun have exactly the same word form. For example:

- `Tôi đã thỏa_thuận với anh ấy` (*I made deals with him*), where `thỏa_thuận` is a verb meaning *make deals*.
- `Anh ấy và tôi có hai thỏa_thuận` (N) (*He and I have two deals*), where `thỏa_thuận` is a noun.

The second type of verb nominalization consists of adding a function word like "`sự`", "`việc`", meaning "*the fact of*" or a classifier noun like "`cái`", "`kẻ`" in front of that verb. For examples:

- `Kinh_tế nước_nhà phát_triển mạnh` (*The country's economy has developed strongly*), where `phát_triển` is a verb;
- `Sự phát_triển của kinh_tế đã mang lại một bộ_ mặt mới cho đất_nước` (*The development of economy has brought a new face to the country*), where `sự phát_triển` is equivalent to a noun.

In Viettreebank, verb nominalization with classifiers is frequently observed. More than 200 occurences of the pattern "`<classifier> + Verb`" can be found, for example "`cái ăn`" (literal translation <*classifier*> + *to eat*, i.e. *the food*), "`người đọc`" (literal translation <*classifier*> + *to read*, i.e. *reader*), *etc*.

All these specialities of Vietnamese language have been taken in account when we describe the syntactic frames and restrictions in viVerbNet.

While building viVerbNet, we use the same representation format of syntax as VerbNet. Allowed

prepositions in the syntax description are put between curly brackets. The following shows usage example and descriptions of a syntactic frame of the verb "đi" in the sense "*Move from one place to another*" (*) (Hoàng, 2003). This verb entry belongs to the verb cluster "đi, chạy, xuôi" that can be mapped to a subclass of the verb class *"attend"*.

EXAMPLE
Tôi đi chợ (*I go to market*)
DESCRIPTION
NP V Destination
SYNTAX
Agent V Destination
SEMANTICS
Motion(During(E), Theme) Location(End(E), Destination)

The SEMANTICS component is introduced in the next section.

5.4 Semantic Predicate

Each syntactic frame is associated to a conjunction of semantic predicates such as cause, manner, contact, *etc*. Each semantic predicate represents the relationship between participants and events to indicate the core meaning of the sentence.

Several predicates are used for describing different stages in the process of an event: the preparatory (Start (E)), the culmination (During (E)), and the consequent (End (E)) stages of an event. This clear representation helps fully describe the core semantic components as well as changes in complex event structures.

Operators can also be added in semantic predicate construction such as negation (NOT) and the absence (?) of certain roles in the described structure .

For example, here is the semantic predicate for "confine" verb class:

- Not (Location (Start(E), Theme, ?Destination)) Location (End(E), Theme, ?Destination) Confine (Result(E), Theme) Cause(Agent, E)

For the semantic component in viVerbNet, we use the same set of semantic predicates as VerbNet.

6 Conclusions

In this paper, we have presented the ongoing project on the construction of viVerbNet, an English VerbNet style lexicon for Vietnamese. We proposed to implement a clustering algorithm for grouping Vietnamese verbs extracted from the available Vietnamese computational lexicon in similar classes. We focused first on describing a small number of major verb classes, before continuing to for similar verb classes.

At the current stage, we have studied 50 verb classes amongst 1000 obtained clusters, in doing an comparative investigation of these classes with English verb classes with similar meanings. Annotated corpora are equally explored for extracting syntactic and semantic information of each verb entry.

The built viVerbNet is designed in a way to be compatible with the English VerbNet. The resources will be freely available for research purposes. We are developing a platform for a collaborative revision of this verb lexicon. In addition, we plan to develop syntactic frames for adjective and noun predicates as well.

viVerbNet will be an important linguistic resources that can be applied in several problems such as semantic role labeling, deep semantic parsing, or question answering for Vietnamese.

Acknowledgments

HA My Linh was funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2020.TS.20.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Susan Brown, Dmitriy Dligach, and Martha Palmer. 2011. Verbnet class assignment as a wsd task. volume 47, pages 85–94, 01.

William H. E. Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Clau-

dia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936, Sydney, Australia, July. Association for Computational Linguistics.

My Linh Ha, Thi Luong Nguyen, Viet Hung Nguyen, Thi Minh Huyen Nguyen, Hong Phuong Le, and Thi Hue Phan. 2015. Building a semantic role annotated corpus for vietnamese. In *Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology*, pages 409–414.

Phê Hoàng. 2003. *Từ điển tiếng Việt*. Nhà xuất bản Đà Nẵng, Việt Nam.

Daisuke Kawahara and Martha Palmer. 2014. Single classifier approach for verb sense disambiguation based on generalized features. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4210–4213, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, 12.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *arXiv preprint*, arXiv:2003.00744.

Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, and Xuan Luong Vu. 2006. A lexicon for vietnamese language processing. *Language Resources and Evaluation*, 40:291–309, 12.

Phuong Thai Nguyen, Luong Vu Xuan, Thi Minh Huyen Nguyen, Van Hiep Nguyen, and Phuong Le-Hong. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP*, Singapore.

Thi Luong Nguyen, My Linh Ha, Viet Hung Nguyen, Thi Minh Huyen Nguyen, and Hong Phuong Le. 2013. Building a treebank for vietnamese dependency parsing. *Proceedings of the 2013 RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 147–151.

Peter Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, November.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. volume 3406, pages 100–111, 02.

Xuan-Son Vu, Thanh Vu, Son N. Tran, and Lili Jiang. 2019. Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*.

Utilizing BERT for Question Retrieval in Vietnamese E-commerce Sites

Thi-Thanh Ha
HaNoi Uni. of Science and Technology, VietNam
ThaiNguyen Uni. of Information and Communication Technology, VietNam
htthanh@ictu.edu.vn

Van-Nha Nguyen
HaNoi Uni. of Science and Technology
nha282@gmail.com

Kiem-Hieu Nguyen
HaNoi Uni. of Science and Technology
hieunk@soict.hust.vn

Kim-Anh Nguyen
HaNoi Uni. of Science and Technology
anhnk@soict.hust.vn

Tien-Thanh Nguyen
HaNoi Uni. of Science and Technology
20144052@student.hust.edu.vn

Abstract

Question retrieval is an important task in question answering. This task is considered to be challenging due to the lexical gap issue, i.e., similar questions could be expressed in different words or phrases. Although there are numerous researches conducted on question retrieval task in English, the corresponding problem in Vietnamese hasn't been studied much. In this investigation, we highlight our efforts on question retrieval in Vietnamese e-commerce sites majorly in two directions: (1) Building a Vietnamese dataset for question retrieval in e-commerce domain. (2) Conducting experiments using recent deep learning techniques including BERT-based classifiers. Our results provide practical examples of effectively employing these models on Vietnamese e-commerce data. Particularly, we demonstrate that a BERT model trained on e-commerce texts yields significant improvement on question retrieval over BERT trained on general-domain texts.

1 Introduction

Community-based Question Answering (CQA) systems[1][2] have become an increasingly popular online platform. Community websites, where users can post their own questions or answers to other users' questions, provide frameworks for people with dissimilar backgrounds to share their knowledge and experiences. When a user posts a new question on a community website, it usually takes a while for other users to respond. Moreover, in a certain period of time, the number of questions and answers stored in a database gradually becomes enormous and challenging to handle, which means that the possibility of finding duplicated questions increases. As a result, it is time-consuming to retrieve good answers to a given question in an archive of question-answer pairs. In order to reduce latency, CQA systems should automatically find questions which are similar to a given new question. It is hoped that the answers of these related questions could be useful for the new question.

The problem of question retrieval is defined as follows: Given a query question and a set of existing questions, return the most similar questions to the query. Question retrieval has been extensively investigated with the purpose of answering new questions using previous answers in databases [Zhou et al.2013, Zhou et al.2015]. Previous studies delved into the lexical gap challenge in which query question might contain words and phrases different from its similar questions. Figure 1 is a typical pair of similar questions in our Vietnamese dataset.

In order to deal with lexical gap challenge, previous research applied soft alignment technique originated from machine translation or implicitly disambiguated word meaning using topic models [Cai et al.2011]. A huge number of research methods in recent years have focused on end-to-end approaches based on deep neural networks without depending on feature engineering or external knowledge bases [Wu et al.2018, Tay et al.2017]. These approaches leverage pre-trained embeddings and specific-purpose network structures aiming at

[1] https://stackoverflow.com/
[2] https://www.qatarliving.com/

Question 1: Làm ơn chỉ giùm tôi cách tắt phím slide to unclock trên samsung s9 plus
(Can you please show me how to turn off slide to unlock button on samsung s9 plus)
Question 2: Cách tắt màn hình slide to unclock chỉ để màn hình kiểu vuốt để mở khóa máy ss j7 pro
(how to turn off slide to unlock screen on ss j7 pro)

Figure 1: An example of similar question pair

representing syntactic and semantic information in questions. Until recently, BERT, a pre-trained language model, achieves state-of-the-art performance in many natural language processing (NLP) tasks [Devlin et al.2018]. However, to our knowledge, BERT has not been applied to Vietnamese question retrieval.

In the scope of this paper, we advocate: (1) A public CQA Vietnamese dataset in E-commerce domain for question retrieval problem. (2) Experimentation with various deep learning models on this dataset. (3) Empirical findings on tuning and visualizing attention of these models. (4) A pre-trained BERT embedding model for Vietnamese E-commerce texts.

2 Related Work

Over the recent years, numerous methods have been proposed to deal with community question answering tasks and achieved state-of-the-art results.

Traditional methods attempt to deal with CQA problems by transforming the text in questions into Bag-of-Words (BoW) representation with tf-idf weighting scheme, such as BM25 [Robertson et al.1995]. Count-based language models [Cao et al.2009] have also been considered as a popular method to model questions as sequences instead of bags of words. Nonetheless, such models might not be useful when there are a vast number of possible sequences. A sentence should have an exact pattern, such as string or word sequence, matching to a particular part of another sentence. Another popular model based on semantic similarity is Latent Dirichlet Allocation (LDA) [Blei et al.2002], which is a probabilistic model applied in representing questions through a set of latent topics. The learned topic distribution is then applied to retrieve similar historical questions. In another direction, various methods have been developed based on machine translation techniques, such as the monolingual phrase-based translation model, to measure question similarity [Jeon et al.2005] or question-answer similarity.

Top performing systems in SemEval 2017 Task 3 challenge [Nakov et al.2017] use sophisticated feature engineering such as exploiting kernel functions or extracting tree kernel features from parse trees. For instance, the best-performance system [Filice et al.2017], uses similarity features like cosine distance or Euclidean distance and lexical, syntactic, semantic and distributed representations to learn an SVM classifier.

Recent studies in question retrieval and answer selection [Severyn and Moschitti2015, Tan et al.2015] in CQA highlight the effectiveness of neural network models over time-consuming handcrafted feature engineering. These methods learn distributed vector representation of texts and measure question-question or question-answer similarity for question retrieval or answer selection, respectively [Bonadiman et al.2017, Severyn and Moschitti2015].

BERT (Bidirectional Encoder Representations from Transformers) was proposed in [Devlin et al.2018] as a kind of pre-trained transformer network [Vaswani et al.2017], which was applied to various NLP tasks with state-of-the-art performance, including sentence classification, question answering, and sentence pair regression. Several prior studies substantiate that BERT could perform well in many cases [Liu et al.2019, Hao et al.2019]. Particularly, [Liu et al.2019] illustrated that the performance of BERT can be further improved by some small adjustments in the pre-training process. Besides, [Hao et al.2019] focused on the interpretation of self-attention, which is one of the most fundamental components of BERT.

Prior researches were generally conducted on English datasets. In this paper, we explore how well recent deep learning models, especially pre-trained BERT, could possibly perform on Vietnamese. At the same time, we visualize some attention layers to illustrate the effectiveness of BERT models on Viet-

namese.

3 BERT for Vietnamese Question Retrieval

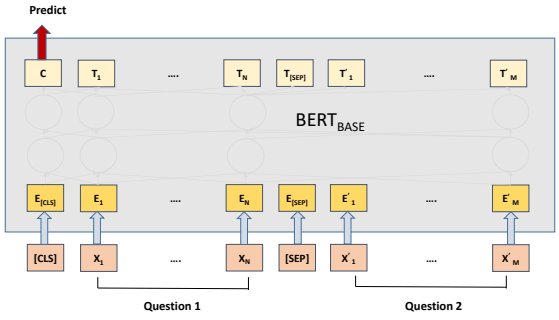


Figure 2: BERT for question retrieval [Devlin el al.2018]

3.1 BERT

BERT is a Bidirectional Encoder Representations achieved from Transformers [Devlin el al.2018] that generates a sentence representation by jointly learning two tasks: masked language modeling and next-sentence prediction. BERT models can be fine-tuned well on both sentence level as well as word level tasks.

BERT has a deep architecture, which has 12 layers of 768 hidden size and 12 self-attention heads. This model begins from the word embeddings layer. In 12 layers, multi-headed attention is calculated using word representations of the previous layer to generate a new intermediate representation. As a result, a token will have 12 intermediate representations with the same size.

In the masked language modeling task, 15% of the tokens are chosen at random to obtain bi-directional pre-trained language model. To avoid mistmatching between pre-training and fine-tuning, in those 15% tokens, a token is replaced with [MASK] 80% of times, 10% of times it is replaced by another random token, and the rest 10% of times it is unchanged.

In the next-sentence prediction task, given a pair of sentences, the aim of this task is to predict whether the second sentence is the true next sentence of the first one.

3.2 BERT for Vietnamese Question Retrieval

In this paper, we apply Multilingual BERT-BASE model (Figure 2), which is considered to be effective on small datasets. It is proved to be good at the ability of cross-lingual generalization by a multilingual representation without being explicitly trained.

Our experiments consist of two parts: Pre-training BERT on unlabeled 1.1M texts of Vietnamese E-commerce (see table 2); and fine-tuning for question retrieval problem on a labeled E-commerce dataset. The parameters of all the layers of our model are fine-tuned at once. A special classification token ([CLS]) and separation token ([SEP]) are added as inputs of our model as followed: $Bert - Input(q_1, q_2) = [CLS]q_1[SEP]q_2[SEP]$, where q_1, q_2 are two questions. The final hidden state corresponding to [CLS] token is applied as an aggregate sequence representation for classification tasks. *Softmax* activation in the last layer is used to predict the label of the considering question.

4 Dataset

We collected questions from users in QA section of The gioi Di dong - an e-commerce website on mobiles, laptops and other electronic devices[3]. An ElasticSearch engine was built from the corpus. We selected a random subset as original questions. Each question was put into ElasticSearch as query. Thereafter, for the first 10 returned questions, human annotators were asked to assess their equivalence to the original question. To increase the difficulty of the task, we removed original questions that could be easily handled by ElasticSearch (i.e questions that have little lexical gap challenge).

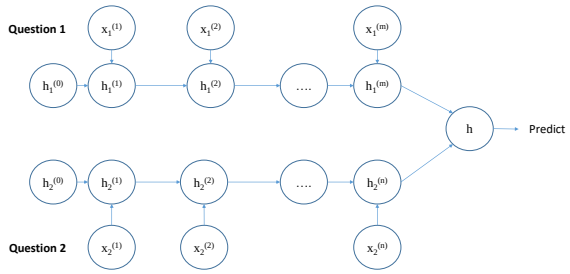
We divided annotated data into three separated sets: training, development, and test (Table 1). In average, 30% of questions were annotated as relevant to the original question.

We also use the large corpus for pre-trained embeddings (Table 2).

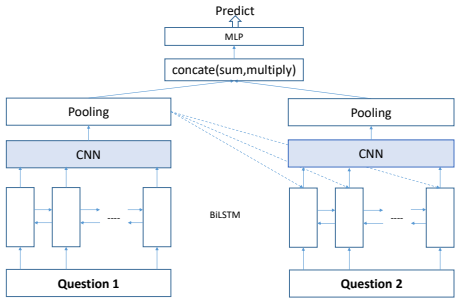
5 Experiments and discussions

Our models were implemented using Tensorflow and all experiments were conducted on GPU Nvidia Tesla p100 16Gb. We used Mean Average Precision

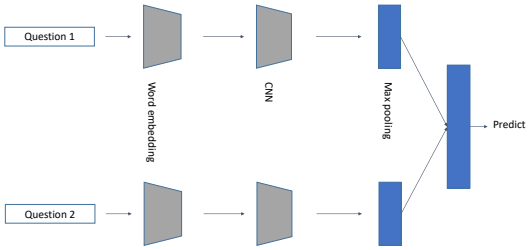
[3]https://www.thegioididong.com/hoi-dap



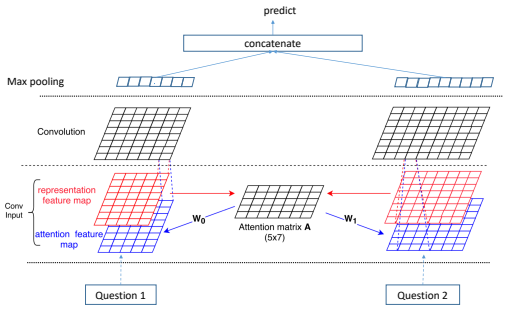
(a) LSTM



(b) LSTM/CNN-attention



(c) CNN



(d) ABCNN

Figure 3: Baseline deep learning models in question retrieval

	Pairs of questions
Train	5,996
Dev	847
Test	1,068
Average length (syllable)	27
Vocabulary (syllable)	5,821

Table 1: Statistics of Thegioididong dataset.

Corpus size	1.1M
Vocabulary size (syllable)	151,735
Average length (syllable)	31

Table 2: Statistics of unlabeled corpus crawled from The gioi Di dong.

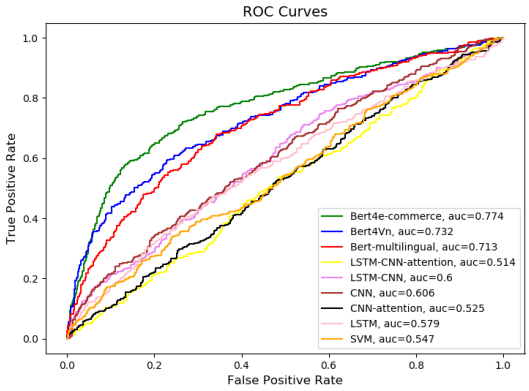


Figure 4: The ROC curves of prediction models.

(MAP) for evaluation. Hyper-parameters were tuned on the development set.

Table 3 presents detailed experimental results on Thegioididong. The results are divided into three parts: vanilla neural networks with LSTM/CNN encoder; BERT pre-trained on different corpora; and baseline bag-of-word models. In all models except PhoBERT, we used syllables as unit input. In PhoBERT, we used its built-in module for word segmentation[4].

Figure 4 illustrates the accuracy of nine models.

[4]https://github.com/VinAIResearch/PhoBERT

Models	MAP
LSTM	52.60
CNN	53.10
ABCNN	51.52
LSTM attention	55.50
BERT-multilingual	61.06
BERT4Vn	63.75
PhoBERT	65.50
BERT4ecommerce	**70.50**
ElasticSearch	52,00
SVM	49.75

Table 3: MAP score of models on Vietnamese dataset.

In general, both Table 3 and Figure 4 show that deep learning approach is better than baseline models; and there was a substantial rise of BERT models, especially when pre-trained on domain data.

5.1 LSTM/CNN Networks

Figure 3 shows the architecture of our models.

- **LSTM**: Both questions are encoded by a shared-weight bi-directional LSTM. The representation of each question is concatenation of the last hidden units of each direction. The representations of two questions are concatenated and is fed into an MLP for prediction.
- **CNN**: Bi-directional LSTM building-block is replaced by CNN.
- **ABCNN** [Yin et al.2015]: This model employs an attention feature matrix to influence convolution. Attention matrix is generated by matching units of the first question representation feature map with units of the second question representation feature map. It can be viewed as a new feature map of two questions to put into next layer.
- **LSTM/CNN-attention**: In this model, outputs from all words of both questions are passed through a word-wise dot product to create a word-by-word attention alike matrix. Updated hidden vector of both question from attention serve as inputs of CNN structure. A global max pooling is then applied to collect important features before prediction. This model is close to LSTM siamese networks as in [Tan et al.2016].

We pre-train syllable embeddings using word2vec on the unlabeled e-commerce corpus. Embedding layers were initialized by pre-train vectors. Adam [Kingma and Ba2014] is used as optimization function. Hyper-parameters used in each experiment are shown in Table 4.

As shown in Table 3, simple concatenation of output from LSTM/CNN and using MLP for prediction slightly outperform baseline models. Learning attention weights as in ABCNN even hurts the performance. In LSTM/CNN-attention, directly calculating word-by-word attention using dot product results in significant improvement.

5.2 Pre-training and Fine-tuning Bert

BERT experiments are performed using Multiligual BERT-BASE model[5]. We first pre-trained BERT on unlabeled E-commerce Vietnamese with maximum length of 200, batch size of 32, and learning rate of $2e^{-5}$ with 20000 steps. We call this model BERT4ecommerce. After pre-training, our model was fine-tuned on question retrieval using Thegioididong dataset.

We also compare our in-domain pre-trained model with other general-domain pre-train BERTs:

- **BERT-multilingual** [Pires et al.2019]: 110K wordpiece vocab, pre-trained on Vietnamese Wikipedia corpus
- **BERT4Vn**[6]: Pre-trained on 500M words of Vietnamese news.

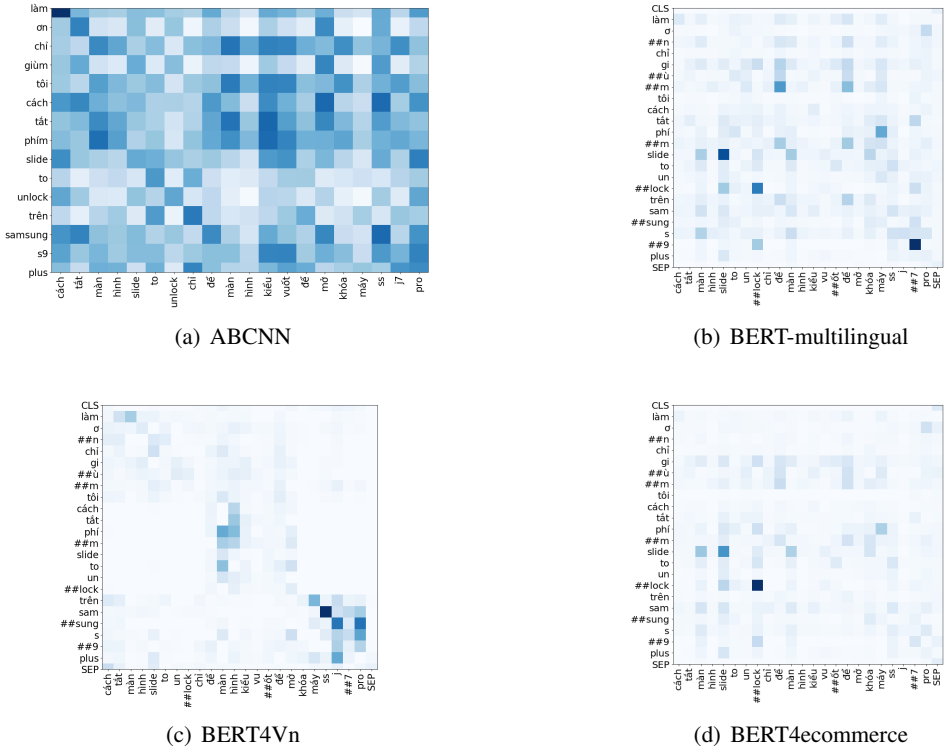
As both shown in Table 3 and Figure 4, significant improvement was obtained by using BERT. Especially, Bert4E-commerce achieved the highest performance (70.50% in MAP, 77.4% in AUC). These experiments advance the idea that when source domain used in pre-training model and target domain are the same, it could have good impact on the final result. E-commerce vocabulary consists of a wide range of words used for technological devices such as Iphone, Samsung S9, "mua-tra-gop" (pay by installments) and so on. Moreover, E-commerce data or social data in general has no guarantee in spelling, grammar and word usage. For instance, numerous spelling mistakes and abbreviations such as

[5]https://github.com/google-research/bert

[6]https://github.com/lampts/bert4vn

	Emb-size	Hid-size/filter-size	L-rate	P_{drop}	Batch size	epochs	Params (x10^5)
LSTM	300	300	0.0001	0.2	64	25	21
LSTM/CNN-attention	300	300	0.0001	0.2	64	25	27
CNN	300	3	0.003	0.5	64	25	33
ABCNN	300	3	0.001	0.2	32	25	34

Table 4: The hyper-parameters set of LSTM/CNN models



(a) ABCNN (b) BERT-multilingual

(c) BERT4Vn (d) BERT4ecommerce

Figure 5: Visualization of BERT and ABCNN

"thoong bao" (notification), "mk" (password) "ss" (Samsung), "f" (keyboard) were found in our dataset. Thus, retraining word embedding on E-commerce domain is required and much more effective than using pre-trained model on news source data such as Wiki and news in this situation.

5.3 Word-based BERT

So far, all our models were based on syllables. In this section, we use a word-based BERT model and apply it to segmented questions. We chose PhoBERT [Nguyen and Nguyen2020], a pre-trained model on 3B segmented texts from Wikipedia and news.

Results show that PhoBERT performs better than BERT-multilingual and BERT4Vn which indicates that word segmentation is helpful for question retrieval in in-domain social texts. Nevertheless, without word segmentation, BERT pre-trained on in-domain texts still outperforms PhoBERT in a large margin. This result is encouraging as word segmentation in in-domain texts suffers from unknown words and spelling mistakes that could propagate errors to downstream tasks.

5.4 Attention visualization

It is argued in [Wiegreffe and Pinter2019] that attention can be use to explain model prediction. In this section, we visualize attention of BERT4 and

	max-length	learning rate	steps reach max
BERT-multilingual	200	$2e^{-5}$	650
BERT4Vn	200	$2e^{-5}$	1600
PhoBERT	200	$2e^{-5}$	1000
BERT4ecommerce	200	$2e^{-5}$	900

Table 5: The hyper-parameters set of fine-turning BERT models

ABCNN to point out that self attention of BERT could learn semantic relationship in questions better than some commonly known attention mechanism such as ABCNN. An attention matrix of Bert was extracted from the first attention layer.

Figure 5 visualizes word-by-word attention between query question (Y-axis) and candidate question (X-axis). This visualization presents alignment weights between two questions, where darker color correlates with larger value.

The attention distribution of BERT is sparser than that of ABCNN. This helps to strengthen interaction between important words such as 'slide' with 'màn hình' (screen), 'lock', 'tắt phím' with 'khóa máy' as seen in the example. The research in [Cui et al.2019] shows that sparse attention matrix achieved from BERT leads to a more interpretable representation of inputs.

6 Conclusion

We carried out a range of experiments with LSTM, LSTM attention, CNN, ABCNN and fine-tuning BERT for question retrieval on a Vietnamese dataset. In particular, our BERT model pre-trained on an ecommerce corpus could be useful for related research.

We hope our work can give a boost to applications related to CQA on Vietnamese Ecommerce data. In the future, we are going to investigate the effect of word segmentation to question answering in ecommerce domain.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2002. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 601–608. MIT Press.

Daniele Bonadiman, Antonio E. Uva, and Alessandro Moschitti. 2017. Multitask learning with deep neural networks for community question answering. *CoRR*, abs/1702.03706.

Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. 2011. Learning the latent topics for question retrieval in community QA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 273–281, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 265–274, New York, NY, USA. Association for Computing Machinery.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2019. Fine-tune BERT with sparse self-attention mechanism. 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 3548–3553, Hong Kong, China, November. Association for Computational Linguistics.

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.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2017. KeLP at SemEval-2017 task 3: Learning pairwise patterns in community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 326–333, Vancouver, Canada, August. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of bert. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and an-

swer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, page 84–90, New York, NY, USA. Association for Computing Machinery.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

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.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada, August. Association for Computational Linguistics.

Dat Quoc Nguyen and A. Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *ArXiv*, abs/2003.00744.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, January.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 373–382, New York, NY, USA. Association for Computing Machinery.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473, Berlin, Germany, August. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2017. Enabling efficient question answer retrieval via hyperbolic neural networks. *CoRR*, abs/1707.07847.

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. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China, November. Association for Computational Linguistics.

Wei Wu, Xu Sun, and Houfeng Wang. 2018. Question condensing networks for answer selection in community question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1755, Melbourne, Australia, July. Association for Computational Linguistics.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR*, abs/1512.05193.

Guangyou Zhou, Yubo Chen, Daojian Zeng, and Jun Zhao. 2013. Towards faster and better retrieval models for question search. In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, CIKM13, page 2139–2148, New York, NY, USA. Association for Computing Machinery.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. pages 250–259, 01.

Language change in Report on the Work of the Government by Premiers of the People's Republic of China

Renkui Hou
Guangzhou University, Guangzhou, China
hourk0917@163.com

Chu-Ren Huang
The Hong Kong Polytechnic University, Hong Kong
churen.huang@polyu.edu.hk

Kathleen Ahrens
The Hong Kong Polytechnic University, Hong Kong
Kathleen.ahrens@polyu.edu.cn

Abstract

The present paper explored the focusing topics change and language change in Report on the Work of the Government by Premiers of the People's Republic of China (hereinafter Report texts). The text clustering and correspondence analysis showed the focusing topics change in selected three periods Report texts. The Report texts were represented by the clause length distribution and clustered. The clustering result showed the differences of clause length usages in the Report texts. The relationship between clause length and word length was studied. The average word length decreases with clause length and were fitted using the function, $y = ax^b$ based on the Menzerath-Altmann Law. The relationship between the three periods Report texts represented by the fitted parameters, *a* and *b*, were explored.

1 Introduction

Language change has been main concern of linguists from centuries in many parts of the world and is the topic of research for classical linguistic studies. However, these early works tended to focus on sound and sound changes, tracing back to work by Pāṇini in 4th century BCE. Biber (2012) argued strongly that reference works that describe different linguistics levels, i.e., lexical, grammatical, and lexico-grammatical, should consider register difference.

Language is the mode of political discourse. For example, the president candidates will demonstrate their ideas and policies in the debates. Van Dijk (1997) defined the political discourse based on three dimensions: the actors, the political scope of the discourse and the context of communication. From the above definition, a discourse is considered as 'political' when it is produced by a political actor carrying out a political action (e.g. to govern, legislate, protest or vote) in an institutional context of communication. Randour et al. (2020) conducted a systematic literature review of 164 scientific articles from the Scopus database and confirmed that political discourse is generally limited to the discourses of (institutionalized) political elites and most specifically to oral monological speeches.

The Menzerath-Altmann law originates from the fact that the length of a construct influences the lengths of its immediate constituents in different language domains. It is summarized as "the greater the whole, the smaller its parts" by Paul Menzerath after he detected the dependency of syllable length on word length (Menzerath 1954). Altmann generalized this hypothesis to all the language levels, formulating it as "The longer a language construct, the shorter its components" (Altmann 1980). Hřebíček (1992, 1995, 1997) showed that the whole hierarchy of textual levels is based on this dependency, and called this the Menzerath-Altmann law.

Altmann (1980) gave the theoretical derivation and the corresponding differential equation of the MA law, as shown in Equation (1).

$$\frac{y'}{y} = -c + \frac{b}{x} \quad \text{Equation (1)}$$

The solution to this differential equation is shown in the Formula (1):

$$y = ax^b e^{-cx} \quad \text{Formula (1)}$$

where *y* is the mean size of the immediate constituents, *x* is the size of the construct, and parameters *a*, *b*, and *c* depend mainly on the levels of the units under investigation.

A large number of observations have shown that parameter *c* is close to zero for higher levels of language whereas lower levels lead to very small values of

parameter *b*; only for intermediate levels is the full formula needed (Köhler, 2012).

The two simplified formulas were obtained when higher and lower levels were studied respectively. Formula (2a) has become the most commonly used "standard form" for linguistic purposes (Grzybek, 2007).

$$y = ax^b \qquad \text{Formula (2a)}$$
$$y = ae^{-cx} \qquad \text{Formula (2b)}$$

This paper explores the language change in Chinese political discourse, Report on Work of the Government by Premiers of the People's Republic of China, based on the content development and Menzerath-Altmann law from the perspective of quantitative linguistics.

1.1 Literature review

Millar and Trask (2015) demonstrated that languages change (even their spelling rules) throughout their history. Previous studies about language change focused on sound and sound changes, word and word changes mostly. Lieberman et al. (2007) studied the regularization of English verbs over the last 1200 years and how the rate of regularization depends on the frequency of word usage. Lexicostatistics was used to calculate the evolutionary history of a set of related languages and varieties (Bakker et al. 2009, Barbancon et al. 2013). Baker (2011) focused on words that have changed their frequency and meaning in the study of change in British English over the twentieth century. Degaetano-Ortlieb and Teich (2018) have used relative entropy for detection and analysis of periods of diachronic linguistic change. Campos et al. (2020) set a corpus-driven methodology to quantify automatically diachronic language distance between chronological periods of several languages. The results showed that a diachronic language distance based on perplexity detects the linguistic evolution that had already been explained by the historians of the three languages.

There is a long tradition of linguistic research on political discourse. Van Dijk's (1997) definition of political discourse brings together studies focusing on discourse produced by political elites in an institutional context with the aim of carrying out a political action (Randour et al. 2020). There are some studies aims at studying political issues, events and actors, such as ideology or identity construction (Wang 2007, Wodak and Boukala 2015). Other studies concentrated on specific linguistics characteristics of the discourses, such as Wang and Liu's (2018) analysis of Trump discourse, or Roitman's (2014) study of the use of pronouns by French presidential candidates. Lu and Ahrens (2008) studied the metaphor usage in political discourse. Savoy (2018) examines the verbal style and rhetoric of the candidates of the 2016 US presidential primary elections.

Yu (2008) demonstrated that machine learning methods can be trained to classify congressional speeches according to political parties. Better performance levels can be achieved when the training examples are extracted from the same time period as the test set. This means that the congressional speeches have different stylistic features in different periods. Yu (2013) explored the correlation between language usage and gender, and reveals that (political) feminine figures tend to use emotional words more frequently and employ more personal pronouns than men.

Previous research has validated the MA law at different language levels. Tuldava (1995) examined the dependence of average word length on clause length, finding a statistically highly significant interdependence between average word length and clause length, indicating that there are other factors that influence average word length. Motalová et al. (2014) and Ščigulinská and Schusterová (2014) verified the validity of the MA law applied to contemporary written and spoken Chinese respectively. Benešová (2016) tested the potential validity of the MA law on samples in different languages and attempted to test the concept of this language universal. Hou et al. (2017) studied the relationship between sentence length and clause length in Chinese language and concluded that the relationship in formal written register can be fitted by the MA law.

There are some researches for studying the theory and formula per se. Köhler (1984) interpreted the parameters in Formula (2a) and assumed that *a* represents a quantity depending on the language and language levels, *b* might represent a shortening tendency and might describe the range of structural information that has to be stored for each language component. Cramer (2005) showed that parameters *a* and *b* might be linked by a systematic connection through a correlation analysis. Hou et al. (2019) fitted the relationship between clause length and word length in different Chinese registers and concluded that the relationship between the fitted parameters, *a* and *b*, in each register can be fitted by the linear regression. The result of linear regression is different in different Chinese register texts.

There are few studies on Chinese language change and on linguistic characteristics of political discourse. This paper will explore the language change in Chinese political discourse based on topic words and the MA law.

1.2 Data and methodology

The Report texts on Works of Government of China in three different periods were selected to establish the corpus. These three different periods are 1978-1982, 1997-2001 and 2016-2020 respectively. The first five years, 1978-1982, are the initial stage of the reforming and opening up in China. Hong Kong was returned to China

in 1997. The last five years, 2016-2020, is the 13th five-year plan was initiated and finished.

The texts of Report were segmented and Parts of Speech tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS).

The text vectors were established based on Vector Space Model and bag of words. Text clustering was used to validate whether the selected linguistic characteristics can differentiate the Report texts in different periods. The differences can be demonstrated and interpreted. They reflect the language change and the focusing topic transition in three periods Reports. Correspondence analysis was used to analyze the correlation between these characteristics and different periods report texts.

The Formula (2a) was used to fit the average word length distribution in the clause with certain lengths (i.e., the relationship between clause and words). Then, the texts can be represented by these fitted parameters, *a* and *b*, and were displayed in a 2-dimensional space. Thus, the relationship between Report texts from these periods can be explored.

2 The change of the thematic words

The Report texts, as political discourses, speak about sharply defined topics and concentrate more or less on the core of the information. Usually the topics are linguistically represented by a particular number of nouns (or even proper nouns) and first order predicates, namely verbs and adjectives. The nouns represent the concrete and abstract concepts in the Report. The verbs represent the action from the subjects or on the objects which are all represented by nouns. The nouns can be modified by the adjective(s) and one noun phrase can be established. The nouns, adjectives and verbs are summarized as thematic words (Popescu et al. 2009).

The nouns, verbs and adjectives, occurring more than 50 times in all texts, were selected to represent the Report texts in three periods. Each report text is represented as one vector using the relative occurrence frequencies of these words.

The hierarchical clustering analysis was used to cluster the texts. Kullback-Leibler Divergence (Relative Entropy) were adopted to compute the distance between text vectors. The sum of squares of the deviations was used to calculate the distance between two clusters. The hierarchical clustering result, dendrogram, is shown in Figure 1.

From Figure 1, we can see that the Report texts from one period were clustered into one cluster. This means that there are systematical similarities of the content words usages in the same period texts and there are systematical dissimilarities between different periods Report texts. The internal differences of content words usages in one period texts, 1997-2001 and 2016-2020 respectively, are small compared with that in 1978-1982 texts. This may be caused by the drastic social changes in that period. The height of common ancestors of these three periods Report texts showed the relationship between the content words usages.

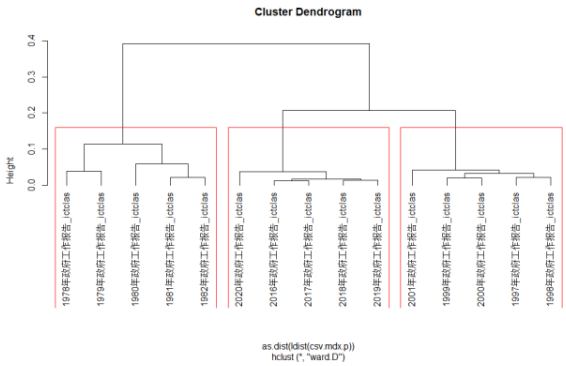


Figure 1: The result of hierarchal clustering of texts represented by the content words

Correspondence analysis is a summary technique which outputs a correspondence plot. A 2D correspondence plot is the most useful depiction of complex reality because it reduces the number of dimensions of variation to the manageable two dimensions represented by the x-axis and the y-axis. Its unique feature is the fact that it captures both the column (content words) and row (periods of report) categories of the cross-tabulation table in the same space.

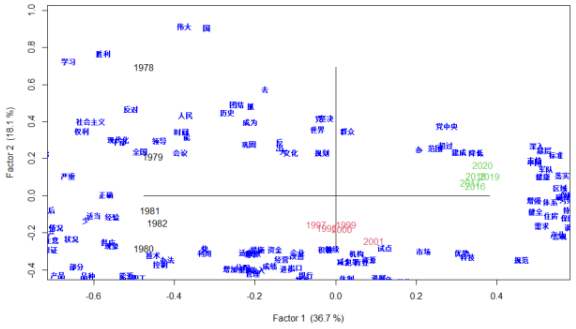


Figure 2: The correspondence analysis result of Report texts in three periods

The result of correspondence analysis is shown in Figure 2. The Figure 2 shows three periods Reports texts (1978-1982, 1997-2001, 2016-2020) were clustered according to their usages of different content words (nouns, verbs and adjectives). Both the reports texts and content words are displayed in the plot. From Figure 2, we can see that some words co-occurred with some Reports from different periods. For example, "健康/住房/增强 (health/housing/enhance)" overlapped with the Reports from 2016-2020. It means these issues are the concerning

focus of Government of these five years. “经验/学习/现代化 (experience/learn/modernization)” were the concerning focus of Government in the initial stage of the reform and opening. “企业/机构/试点/改善 (enterprise/organization/experimental unit/improvement” were focused by Government in 1997-2001.

3 The change of relationship between clause length and word length

The sentence, as the maximal grammatical unit and minimal statement unit, is considered to be a basic linguistic unit in all languages. Chinese sentences are often defined in terms of characteristics of speech (Huang and Shi 2016; Lu 1993). Chao (1968) and Zhu (1982) defined a sentence as an utterance with pauses and intonation changes at its boundaries.

A common approach for identifying sentences in syntactically annotated corpora (e.g., Chen et al., 1996; Chen et al., 2003; Huang and Chen, 2017 for Sinica TreeBank) is to mark all segments between punctuation marks that indicate utterance pauses as sentences. Such punctuation marks include commas, semicolons, colon, periods, exclamation marks, and question marks. Wang and Qin (2014) and Chen (1994) also adopted this operational definition and called such units *sentence segments*. Wang and Qin (2014) considered the lengths of *sentence segments* to be relevant to language use in Chinese. Sentences (as defined by Chen et al., 2003; Huang and Chen, 2017) and sentence *segments* (as defined by Chen, 1994; Wang and Qin, 2014) are roughly equivalent to clauses.

3.1 The distribution of clause length

Clause length is defined as the number of words included. Words is considered to be the segments delineated by blank spaces in the texts segmented by a Chinese lexical analysis system. The occurrence frequencies of clauses with certain lengths were calculated in three periods texts and the relative frequency distributions of the clauses in three periods texts are shown in Figure 3.

Figure 3 shows that the relative frequency distributions of clause length in three periods Report texts are similar. The relative occurrence frequencies of clauses increase firstly and then decrease with the increasing of lengths. The occurrence frequencies of one- and two-word clauses are highest in 1978-1982 and are lowest in 2016-2020. The clauses lengths concentrate on the 3-10 words. The percentages of clauses with 1-15 words lengths are more than 95% in three periods texts.

Each text is represented by the relative frequency of clause lengths. Correspondence analysis was used to analyze the texts. The correspondence analysis result, correspondence plot, is shown in Figure 4.

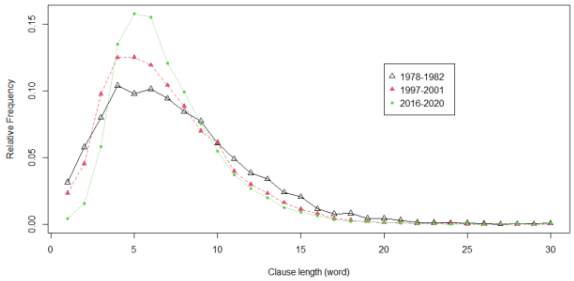


Figure 3: The frequency distribution of clause length in terms of words

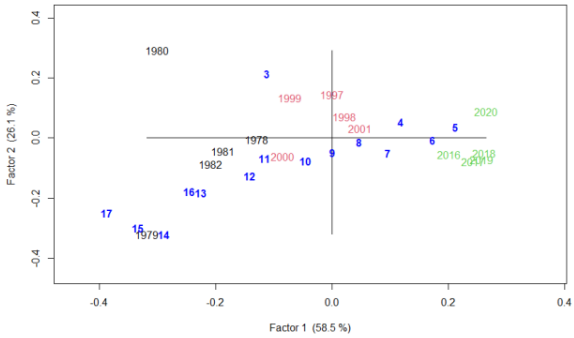


Figure 4: The result of correspondence analysis of texts

From Figure 4, there are differences of clause length usages between Report texts from three different periods. Combined with Figure 3, the short clauses are frequently used in 2016-2020 Report texts and the long clauses are frequently used in 1978-1982.

3.2 Fitted results of average word length in clauses

The average word length in clauses with certain lengths was calculated as the number of Chinese characters in the given clauses divided by the number of words in those clauses. We calculated the average word length distribution in each period texts and fit them using the Formula (2a). The fitted result is shown in Table 1 and Figure 5.

Table 1: The fitted result of the average word length in clauses with certain length

	a	*b*	R^2	*Adjusted* R^2
1978-1982	2.452	-0.144	86.49%	85.45%
1997-2001	2.457	-0.132	88.02%	87.1%
2016-2020	2.654	-0.180	88%	87.08%

From Figure 5, we can see that the average word length decreases with the clause length. The average word

length distribution can be fitted by the Formula (2a) in each period text. The determination coefficient, R^2, in Table 1 showed that the fitted result is good. Based on the fitted result, we conclude that the relationship between clause and word abides by the MA law.

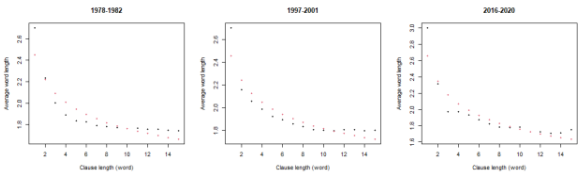


Figure 5: The fitted result of average word lengths in clauses

The average word lengths in 15 texts from three periods were fitted by the MA law. The fitted results were shown in Appendix 1. The values of R^2 demonstrate that the relationships between clause and word lengths in all Report texts abide by the MA law.

The Report texts from three periods are represented by the two fitted parameters, *a* and *b*, of the average word length in clause with certain length. They are displayed in a two-dimensional space, as shown in Figure 6.

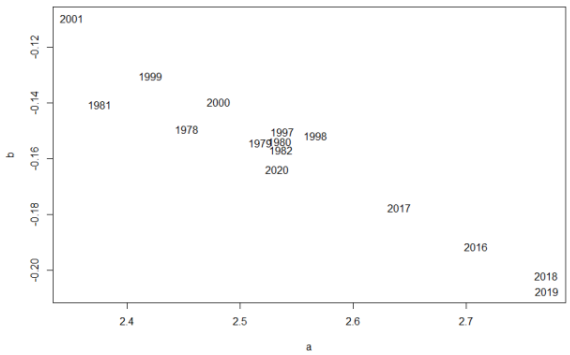


Figure 6: The relative position of Report texts from each period

From Figure 6, the two parameters, *a* and *b*, correlated negatively, which means *b* decreases with *a*. The *b* values are smallest in Report texts in 2016-2020, which means the extent of the decreasing of the average word length in clauses is maximum. The big *a* values in 2016-2020 Report texts mean that average word length in short clauses are larger than the other two periods texts. The ranges of fitted parameters are similar, but the relationship between these two parameters is not similar in texts from 1978-1982 and 1997-2001.

4 Conclusion

This paper studied the changes of language and focusing topics of Chinese political discourse represented by Reports on Work of the Government by Premiers of the People's Republic of China. We selected the Report texts in three periods, 1978-1982, 1997-2001 and 2016-2020, to establish the corpus of political discourse. Text clustering result showed that the thematic words in Report texts of these three periods are changing. Correspondence analysis showed that the correlation between Report texts and thematic words in correspondence plot. From correspondence plot, it can be seen that the changes of thematic words usages in these three periods Report texts.

Then the Report texts were represented by the clause length distribution and analyzed using correspondence analysis. The result of correspondence analysis showed that the more short clauses were used and the less long clauses were used with time from 1978-1982, 1997-2001 to 2016-2020. The average word length in clauses with certain lengths were calculated, which decreases with the clause length. Formula 2a was used to fit the average word length. The fitted result showed that the relationship between clause and word lengths abide by the MA law.

The 15 Report texts were represented by the fitted parameters of average word length. The parameters were used to represent the Report texts. The two-dimensional space was used to show the relationship between the texts. The result showed that the parameters *b* in 2016-2020 Report texts are smaller than that in Report texts from 1978-1982 and 1997-2001. It needs to be explored for the relationship between the fitted parameters and the development of Chinese language.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful and helpful comments.

Funding support. Research on this paper was funded by National Social Science Fund in China (Grant No. 16BYY110), the Hong Kong Polytechnic University Grant 4-ZZFE, National Natural Science Fund in China (Grant No. 61866035).

References

Altmann, G. (1980). Prolegomena to Menzerath´s law. Glottometrika 2, 1-10.

Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C.H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E.W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. Linguistic Typology 13(1), 169–181.

Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. Journal of English Linguistics, 39(1), 65–88.

Barbançon, F., Evans, S., Nakhleh, L., Ringe, D. and Warnow, T. (2013). An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30, 143–170.

Benešová, M. (2016). Text segmentation for Menzerath-Altmann law testing. Palacký University, Faculty of Arts.

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8(1), 9-37.

Campos, J., Otero, P., & Loinaz, I. (2020). Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. Natural Language Engineering, 26(4), 433-454. doi:10.1017/S1351324919000378.

Chao, Y. R. (1968). A Grammar of Spoken Chinese. Berkeley and Los Angeles: University of California Press.

Chen, H. H. (1994). The contextual analysis of Chinese sentences with punctuation marks. Literary and linguistic computing, 9(4): 281-289.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp. 167-176.

Chen, Keh-Jiann, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. (2003). Sinica Treebank: Design Criteria, Representational Issues and Implementation. In Anne Abeillé (Ed.), Treebanks: Building and Using Parsed Corpora (pp. 231-248). Dordrecht; Boston: Kluwer Academic Publishers.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. Journal of Quantitative Linguistics, 12, 41–52.

Degaetano-Ortlieb, S. and Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 22–33.

Grzybek, P. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In P. Grzybek and E. Stadlober. Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, 62, 205.

Hou, R, Chu-Ren Huang, Hue San Do & Hongchao Liu (2017): A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, Journal of Quantitative Linguistics. 24(4): 350-366.

Hou, R., C.-R. Huang, M. Zhou & M. Jiang. (2019). Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis. Glottometrics. 45: 24-56.

Huang, Chu-Ren and Shi, D. (2016). A Reference Grammar of Chinese. Cambridge: Cambridge University Press.

Huang, C.-R. & K.-J. Chen. (2017). Sinica Treebank. In N. Ide and J. Pustejovsky (eds), Handbook of Linguistic Annotation. Berlin & Heidelberg: Springer.

Hřebíček, L. (1992). Text in communication: Supra-sentence structure. Bochum, Brockmeyer.

Hřebíček, L. (1995). Text levels: Language constructs, constituents and Menzerath-Altmann law. Trier: WVT.

Hřebíček, L. (1997). Lectures on text theory. Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In W. Lehfeldt & U. Straus (Eds.), Glottometrika 6, 177-183. Bochum: Brockmeyer.

Köhler, R. (2012). Quantitative syntax analysis (Vol. 65). Berlin: Walter de Gruyter.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. Nature 449(7163): 713-716.

Lu, J. (1993). The features of Chinese sentences. Chinese Language Learning. No. 1:1-6.

Lu, LW-L & K. Ahrens. (2008). Ideological influence on BUILDING metaphors in Taiwanese presidential speeches. Discourse & Society 19 (3): 383-408.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes (Vol. 3). F. Dümmler.

Millar, R.M. and Trask, L. (2015). Trask’s Historical Linguistics. Abingdon-on-Thames: Routledge.

Motalová, T., Spáčilová, L., Benešová, B., Kučera, O. (2014). An application of Menzerath-Altmann law to contemporary written Chinese. Křížkovského, Olomouc: Univerzita Palackého v Olomouci.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram,B.D., Köhler,R., Krupa,V.,Maˇcutek, J., Pustet, R., Uhlíˇrová, L., Vidya, M.N. (2009). Word frequency studies. Berlin-New York: Mouton de Gruyter.

Randour, F., Perrez, J., & Reuchamps, M. (2020). Twenty years of research on political discourse: A systematic review and directions for future research: Discourse & Society, 31(4), 428–443.

Roitman, M. (2014). Presidential candidates’ ethos of credibility: The case of the Presidential Pronoun I in the 2012 Hollande-Sarkozy Debate. Discourse & Society 25(6): 741-765.

Ščigulinská, J. & Schusterová, D. (2014). An Application of the Menzerath-Altmann Law to Contemporary Spoken Chinese. Palacký University in Olomouc. First Published 2014.

Tuldava, J. (1995). Informational measures of causality. Journal of Quantitative Linguistics, 2(1), 11-14.
Van Dijk TA. (1997). What is political discourse analysis? Belgian Journal of Linguistics 11: 11–52.
Wang, J. (2017). Representing Chinese Nationalism/Patriotism through President Xi Jinping's "Chinese Dream" discourse. Journal of Language and Politics. 16(6): 830-848.
Wang, Y and H. Liu. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during 2016 election. Discourse & Society 29(3):299-323.
Wang, K., & H. Qin. (2014). What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. Corpus Linguistics and Linguistic Theory, 10(1), 57-77.
Wodak, R. & S. Boukala. (2015). European identities and the revival of nationalism in the European Union. Journal of Language and Politics 14(1): 87-109.
Yu, B. (2008). Classifying party affiliation from political speech. Journal of Information Technology & Politics. 5(1): 33–48.
Yu, B. (2013). Language and gender in congressional speech. Literary and Linguistic Computing. 29(1): 118–132.
Zhu, D. (1982). Lectures on Grammar. Beijing, China: Commercial Press.

Appendix 1:

Table：The fitted result of average word lengths in 15 Report texts from three periods

	a	*b*	R^2
1978	2.452905	-0.14949	0.792514
1979	2.517411	-0.15424	0.817301
1980	2.534983	-0.15377	0.901367
1981	2.375607	-0.14064	0.822611
1982	2.536153	-0.15683	0.953379
1997	2.53746	-0.15027	0.80805
1998	2.566591	-0.15172	0.862896
1999	2.420279	-0.13037	0.865351
2000	2.480595	-0.13977	0.91811
2001	2.351106	-0.10954	0.925222
2016	2.708322	-0.19148	0.871852
2017	2.640937	-0.17755	0.860396
2018	2.770532	-0.20207	0.809142
2019	2.771229	-0.2077	0.938813
2020	2.532724	-0.16393	0.891272

From Sense to Action: A Word-Action Disambiguation Task in NLP

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Yu-Hsiang Tseng
Graduate Institute of Linguistics
National Taiwan University
seantyh@gmail.com

Chiung-Yu Chiang
Graduate Institute of Linguistics
National Taiwan University
cychiang@ntu.edu.tw

Richard Lian
Graduate Institute of Networking and Multimedia
National Taiwan University
dclian@nlg.csie.ntu.edu.tw

Yong-fu Liao
Graduate Institute of Linguistics
National Taiwan University
mcku1115@gmail.com

Mao-Chang Ku
Graduate Institute of Linguistics
National Taiwan University
mcku1115@gmail.com

Ching-Fang Shih
Graduate Institute of Linguistics
National Taiwan University
r08142004@ntu.edu.tw

Abstract

Words are conventionalized symbols that present the function by which meaning is attached to form. The Word Sense Disambiguation, which has been taken as one of the core semantic processing tasks in the pipe-lined NLP architecture, aims to assign proper word sense to lemma form in varied contexts based on a word-sense inventory such as WordNet. However, there are some theoretical assumptions unattested from a functional linguistic point of view. This paper proposes an alternative by introducing a novel task called word action disambiguation task (WAD) concentrated on the observable pairs between words and actions. The accompanying dataset, which was manually edited and compiled, is composed of 419 multiple-choice questions. We further verified the dataset through item evaluation with human rating data, and the semantic relations among the dataset were annotated automatically. A baseline performance with an accuracy of 38.64% was also provided with BERT models and 43.18% after incorporating paradigmatic knowledge with semantic graph. We expect the proposal of the WAD task and dataset would motivate computational models to incorporate more complex aspects of human language.

1 Introduction

Due to its polysemous behavior, selecting the most appropriate sense for a word in a text has been one of the most important yet challenging NLP tasks over the years. Given a pre-defined sense inventory, computationally assigning each word in target texts with proper sense (thus Word Sense Disambiguation) is assumed to be crucial for MT, IR, QA, and other systems (Navigli, 2009). Although the sense inventory such as WordNet has been continuously maintained and implemented cross-linguistically, the issue regarding the extent to which the sense granularity (i.e., levels of semantic specificity) in the sense inventory would be sufficient for downstream NLP tasks remains less explored.

Three tacit and intertwined assumptions underlying the conventional WSD task are (1) word senses can be operationalized as discrete and distinguishable ones, (2) word senses (as included in the sense inventory) can be shared by the entire language community, and (3) WSD with the fine-grained sense specification can be successfully applied to actual language data, and facilitate a wide range of downstream NLP tasks. However, the reported poor inter-annotator agreement (IAA) and low reliability of sense distinction/annotation in the

task seem to falsify these assumptions and thus motivate projects like OntoNotes (Hovy et al., 2006; Cinková et al., 2012).

This paper aims to serve as a first attempt to propose an alternative to the underlying assumptions from the functional and granular linguistic perspective. First, the notion wordhood of as assumed in the WSD task is not self-evident, particularly for languages whose writing systems do not provide the delimiter of a word boundary. In this aspect, word segmentation or determination is rather theory-laden and would be best regarded as the wordhood annotation rather than the preprocessing task with ground truth as conventionally taken. Second, word-meaning pairs are fluid in nature, whose granularity (in terms of the length of the word and the functions it carries) is influenced by its underlying ontology (paradigmatic dimension), surrounding context (syntagmatic dimension) and real-world application (pragmatic force). Under this view, it is hard to get a common, static, or solid 'feel of sense' among native speakers. Finally, it is still unclear regarding the relation between WSD and Natural Language Understanding (NLU). For instance, what levels of granularity of sense (from fine-grained to coarse-grained) do we need for the machine comprehension, or in what sense can we justify that WSD is a sine qua non for NLU?

There has been a huge amount of related work trying to grapple with the WSD-related issues by exploiting various machine learning models (Navigli, 2009). On the resource side, in order to achieve better efficiency and performance, sense granularity in the sense inventory such as WordNet was explored and annotated in OntoNotes project (Weischedel et al., 2011; Palmer, Dang & Fellbaum, 2005). However, the paradigm underlying the WSD task has also been questioned since (Kilgariff, 1997), and sense discretization and enumerative view of word senses inventory that is implicitly/explicitly presume is strongly criticized as well (Pustejovsky, 1995). Consequently, we adopt a functional linguistic approach to the linguistic units and introduce the design of a novel task, which can be regarded as an in vivo evaluation of the WSD system.

In terms of language understanding, we see language as a communication device used to ask, demand, raise questions. The utterance, either in spoken or written forms, is an observable word sequence which encodes the speaker's illocutionary force, the "combination of the illocutionary point of an utterance, and the particular presuppositions and attitudes that must accompany that point" (Searle and Vanderveken, 1985). In pragmatics, illocutionary force further distinguishes the following types of acts: inquiring, promising, asserting, ordering, etc. As the words which serve as the building block in the sequence are mostly polysemous, it is thus commonly and naively assumed that one core part of our NLU competence depends on the identification of the correct word sense for each word in the utterance, and the understanding is accomplished in a compositional manner. That's the basic underlying philosophy of the current WSD task. However, these senses are unobservable theoretical constructs. In the communicative context, as long as listeners can react with proper observable responses, the mechanism underlying the word sense disambiguation inside the listener's mind is only latent constructs. That is, the listeners understand the utterances when they react with proper actions against them. This leads to Davidsonian notion of action (Davidson, 1985), that an action is something an agent does that was 'intentional under some description'. The relations among words, senses, and the Davidsonian actions with a framework is depicted in Figure 1.

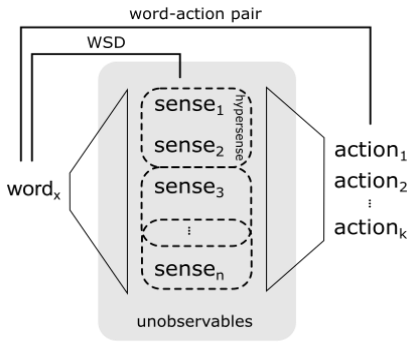


Figure 1: A schematic explanation of the relationships between words, word senses, and proposed *actions*.

To illustrate the relationship between words and actions, we develop a novel task called word action disambiguation task (WAD) to highlight the communication and dynamic aspect of word and action. The action inevitably reduced to textual descriptions in the task in order to be efficiently processed by machines. However, the proposed task underlines the interactions between words and actions by emphasizing the pragmatic and context-dependent

nature among them, and the relationship between them cannot be solely determined by lexical semantics.

2 Word Action Disambiguation Task and its Dataset

This section explains the proposed task and its corresponding dataset to alleviate the issues when splitting fine-grained, continuous word senses as assumed in previous WSD studies. The new task concentrates on the observable words and actions elicited. The task is implemented in the form like multiple-choice decision. A dataset with 439 items was also compiled to accompany the proposed task by 9 annotators. In each item, the question states a scenario, situation, or dialogue, in which a critical word is embedded. The critical words are polysemous single-character words selected from CWN. Resulting from the word's polysemy, four possible descriptions of actions are listed as options. An agent's (models or computer agents) task is to select the most proper action based on the understanding of the critical word's sense.

The critical words are selected from Chinese Wordnet (CWN). Followed by rigorous lexical-semantic theories, CWN distinguishes fine-grained differences between word senses. In the WAD dataset, we selected 400 single-character verbs with more than 3 verbal senses. Among these senses, we defined 4 critical senses of each word where proper action would be impossible if the word senses are conflated. For example, 叫 "jiao4" has 13 senses listed in CWN. In the sentence: “餓了就叫水餃來吃 (Order some dumplings if you are hungry.)”, the sense of the critical word 叫 (jiao4) refers to “order something”. If the agent misunderstands it as calling someone over, 水餃 (shui3jiao3) would be a human, not a kind of food, the resulting actions would be improper.

A complete WAD task item is as follows. We first identify 4 critical word senses and created multiple-choice questions and options (the critical word is marked with angle brackets):

我昨天<吃>了公館夜市的臭豆腐，真棒
I <had> stinky tofu in Gongguan Night Market yesterday. That was great!
A. 難怪假日的時候人潮都很多
No wonder it is so crowded on weekends.
B. 做這事真耗體力，不划算
It is not worthy of doing such labor-consuming work.
C. 這機器太爛了吧，卡插進去就拔不出來
This machine sucks. You can’t get the card back after you insert it.
D. 貨物這麼重喔，難怪船無法停泊在這港口
The cargo must be heavy. No wonder the cargo ship cannot anchor here.

The critical word, 吃 (chi1), has 28 senses in CWN. The question states a scenario in a night market, using the sense of 吃 (chi1) which refers to “eat something”. Options followed are 4 other possible responses toward based on other critical senses: (A) to eat something in; (B) to consume lots of resources; (C) to indicate that a card is captured by a cash machine, and (D) to displace the water while the boat is immersed in the sea. The correct answer to the question is option A.

These 4 options refer to the respective sense by the frame semantics, pragmatics, context, or common-sense knowledge. Importantly, the options are designed not to relate to the question with lexical semantics alone. That is, the questions and options are designed so the mapping relations between words and actions cannot be easily learned by models based on current syntagmatic vector semantics.

The proposed WAD dataset is aimed to be pragmatically, contextually, real-world relevant word action pairs, and these pairs cannot be determined by a model trained only on syntagmatic relations. Therefore, we verify the dataset with two approaches. (1) Item evaluation: we collect human raters' responses on these items and select the most appropriate items to include in the final dataset (Section 3). (2) Dataset Evaluation: we attempted a current deep learning model; the resulting performance is a tentative baseline on the proposed dataset (Section 4).

3 Item Evaluation

We evaluated items in the dataset with human ratings. Results of rating data were used to select the most appropriate items to include in the final dataset. We first describe methods of collecting rating data and item selection results (Section 3.2).

3.1 Item Rating Study

Five Mandarin native speakers, aged from 19 to 24, were recruited in the rating study. After researchers gave instructions, raters were asked to evaluate how well each option matches the question stem. We used a 5-point scale Likert scale on each rating item: from definitely not the correct answer (point 1), not likely to be the correct answer (point 2), possibly incorrect or possibly correct (point 3), likely to be the correct answer (point 4), and definitely the correct answer (point 5). Each rater went through all 1756 question-option pairs. They responded with independent spreadsheets so that ratings data would not be seen by other raters.

There were 8,780 rating scores collected. The mean and the standard deviation of each question-option pairing were shown in Figure 2. The rating means of each pair are bimodally distributed, where modes occurred in point 1 and point 5. The pattern was expected as it indicated the raters tend to agree on which option should or should not be the appropriate choice. The fact that the frequency of ratings with higher scores (above 4) is lower than the frequency of ratings with lower scores (below 2) also aligns with this expectation since only a quarter of the options were designed to be the appropriate choices in the sense-action dataset.

The standard deviation of the ratings for each option signified inter-rater agreements. If raters did not agree on a pair, the rating scores would differ widely, resulting in a large standard deviation; on the contrary, if raters all agree on a pair and gave it the same scores, the standard deviation would be 0. As shown in Figure 2, the distribution of the standard deviations is right-skewed, with most of the standard deviations (64%) having values below 1.0. This indicates a high agreement on the ratings among the raters.

3.2 Item Selection

We devised a two-phase selection scheme each employing a criterion to select appropriate items respectively: agreement criterion and contrast criterion. Two indices were calculated for each criterion: (1) Agreement between correct (as designated by the question authors) and maximally rated options and (2) the ratio between the highest and second-highest rating.

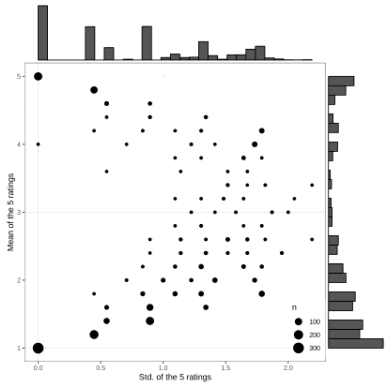


Figure 2: Distributions of the mean and standard deviation of the 5 raters' ratings for the options in the dataset

The agreement between correct and maximally rated options indicated the appropriateness of the answer created by the question authors. If the correct answer is rated lower than other options, the question was clearly not suitable in the dataset and therefore dropped. There were 10 questions omitted in this phase. This process filters out ten sense-action pairs and yields 429 remaining pairs (98%). In the second phase, we remove those pairs where the ratio between the highest and second highest below 1.15. This index indicated the ambiguity of the correct answer among other candid options. If the correct options were rated close to other options, the questions may involve complicated pragmatic or context considerations that cannot be resolved clearly even by human raters. There were further 10 items dropped in this phase. After two phases of item selection, there were 5% of dropped items and resulted in 419 items included in the final dataset[1].

4 Dataset Evaluation

WAD task involves learning the relations between words and actions, where pragmatic, semantic, and common-sense knowledge interact with each other. To evaluate the extent how current machine learning models perform on the WAD task, we compare two different models with two different feature representational approaches as baseline models of the dataset.

[1] Dataset is available at https://github.com/lopentu/WAD

4.1 Feature Representation

Two feature representation approaches are explored in this study. The first approach takes advantage of recent development of contextualized embedding models, specifically BERT (Devlin et al., 2018), to train a multiple-choice model on the proposed WAD task. Past studies showed that, as a transformer based model, a pre-trained BERT model is learned to represent lexical semantics of words and their syntactic relations within the sentences (Manning, Clark, Hewitt, Khandelwal, & Levy, 2020). This approach models the syntagmatic aspects of the linguistic inputs.

However, WAD items are designed to involve more than words' syntagmatic behaviors. Therefore, we devise a second approach to represent the information in items, which is more aimed to capture the paradigmatic relations among the stem and options in an item. Lexical resources, such as ConceptNet (Speer, Chin, & Havasi, 2017), is incorporated into the model through constructing a semantic graph. The graph has all the words in the dataset as nodes and relations (as defined in lexical resources) as edges. The resulting graph consists of 15,600 nodes and 807,426 edges. The graph contains 633 components (groups of nodes connected with each other), 608 of which are single node components. The largest component is composed of 12,469 nodes. An example of the semantic annotation on an item is shown in Figure 3. The semantic graph is further encoded into vectors with node embeddings (Grover & Leskovec, 2016). The hypothesis is that, equipped with paradigmatic and syntagmatic knowledge, the agent performs better in the WAD task.

4.2 Model Results

The dataset is split into a training set and a validation set with 80% and 20% proportions, respectively. A training example is composed of each of the four options concatenated with the question stem, resulting in a vector of four vectors, each one representing a question-option pair. The model needs to learn the indices of the correct answers.

Two models are trained and compared. The first model only uses BERT embedding as input, and a standard multiple-choice readout head, which is composed of a fully-connected layer of 768 hidden units, is stacked upon the output embeddings. The model finally predicts the index of the correct

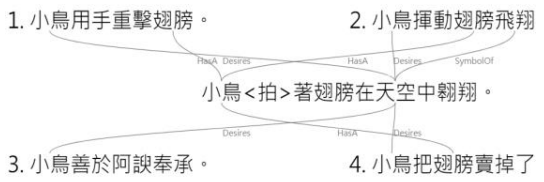


Figure 3: An example of an annotated item. The question stem is in the middle, surrounded by the four candidate options. The links among them are the semantic relations.

question-option pair. The second model's input includes BERT embeddings and node embeddings derived from the semantic graph. The input sequence of node embeddings is fed into a GRU layer, in which the hidden size is 100. The last hidden states of GRU are transformed with a fully-connected layer and concatenated with the BERT output as generated in the first model. We trained these two models on the WAD dataset with a batch size of 8, and the parameters are optimized with Adam optimizer with a learning rate of 5e-5 for 3 epochs.

The first model, with BERT embeddings only, achieved 38.64% accuracy, which was above randomly choosing (25% in a 4-option multiple-choice problem). The model with BERT and semantic graph embeddings achieves better performance with an accuracy of 43.18%. The pattern suggests paradigmatic information is helpful in learning the WAD task.

The current model with contextualized embeddings and semantic graph node embeddings can be considered as a tentative baseline performance for the WAD task. Distinctive from the traditional word sense disambiguation task, where word senses are mostly determined by its syntagmatic context, the WAD task deals further with pragmatics and real-world knowledge. These contextual knowledges are only implied in the text. The common-sense knowledge extracted from ConceptNet is a tentative approach that paves the way for a more comprehensive scheme. Such a scheme may involve annotating the common sense or real-world knowledge suggesting relations underlying the question stem and candidate options. Therefore, the connections between the question stem and correct options would be more accessible for a machine learner.

5 Conclusion

The enumerative and discretization of word senses impose profound limitations, both theoretically and computationally, on fine-grained sense inventories. In addition, the relationship between WSD and NLU remains unclear. Even given the success of WSD/sense tagger, how does that WSD process can logically entail the proper understanding of response in context? In this paper, we bring a 'meaning-in-action' philosophy into the WSD field. We identified the relations between words, senses, and actions and emphasize the observable pairs among them, i.e. word-action pairs. We then proposed a new task called "word-action disambiguation" (WAD), and its accompanying dataset which consisted of 419 multiple-choice questions. The task is designed to incorporate the semantic, pragmatic, real-world aspects of linguistic uses, and the relations between question and option pairs cannot be reduced to merely lexical semantics. We further evaluate each item with human rating data, to ensure the correctness and clearness of each item. A deep learning model, based on BERT, was trained on the WAD dataset to serve as a baseline performance. We expect the proposal of the WAD task and dataset would shed new light to the current architecture of WSD and motivate computational models to incorporate more complex aspects of human language.

Acknowledgement

This work was supported by Ministry of Science and Technology (MOST), Taiwan. Grant Number MOST. 108-2634-F-001-006.

References

Bojanowski, P., Grave, E, Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. Available on arXiv preprint arXiv:1607.04606.

Chang, L. L., Chen, K. J., & Huang, C. R. (2000). A Lexical-semantic Analysis of Mandarin Chinese Verbs: Representation and Methodology. *Computational Linguistics and Chinese Language Processing*, 5, 1-18.

Cinková, S., Martin Holub, Vincent Kríž (2012). Optimizing Semantic Granularity for NLP-report on a Lexicographic Experiment. In: *Proceedings of the 15th EURALEX International Congress*.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-Training with Whole Word Masking for Chinese BERT. arXiv preprint arXiv:1906.08101.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

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.

Grover, A. & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Hovy, E., M. Mitchell, M. Palmer, L. Ramshaw, and R. Weischedel (2006). 'OntoNotes: The 90% Solution'. In: *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, NAACL-Short '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 57–60.

Kilgarriff, A. (1997). 'I Don't Believe in Word Senses.' *Computers and the Humanities* 31: 91– 113.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, Vol. 41, No. 2.

Palmer, Martha, Dang, Hoa & Fellbaum, Christiane. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* (13): 137-163.

Pustejovsky, J. (1995). The Generative Lexicon. MIT press.

Searle, J.R., Vanderveken D. (1985). Speech Acts and Illocutionary Logic. In: Vanderveken D. (eds) *Logic, Thought and Action. Logic, Epistemology, and the Unity of Science*, vol 2. Springer, Dordrecht.

Searle, J.R. (1990). Collective Intentions and Actions. In P. Cohen, J. Morgan, and M. Pollak (eds.). *Intentions in Communication*, Cambridge, MA: MIT Press.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of AAAI 31.

Tripodi, R. & Pelillo, M. (2017). A Game-theoretic Approach to Word Sense Disambiguation. *Computational Linguistics*, 43(1):31-70.

Weischedel, R., et al. (2011). OntoNotes Release 4.0. Philadelphia: Linguistic Data Consortium.

On the syntax of negative *wh*-constructions in Korean

Okgi Kim
University of Wisconsin-Milwaukee
Department of Linguistics
P.O. Box 413
Milwaukee, WI 53201-0413
okgikim@uwm.edu

Abstract

This paper investigates the syntax of Negative WH-Constructions (NWHCs) in Korean and argues, under Coniglio and Zegrean's (2012) split-ForceP framework, that NWH-phrases like *mwe-ka* and *ettehkey*, which are base-generated above or at the edge of IP, undergo covert movement to the split-Force domain to reflect their sensitivity to clause type and turn the original information-seeking force into the speaker-oriented rhetorical force.

1 Introduction

This paper examines so-called Negative WH-Constructions (henceforth, NWHCs) in Korean, which are exemplified by (1) (Cheung, 2008; 2009) (throughout the paper, small capital letters are used in glossing NWH-items to distinguish them from ordinary *wh*-items).[1][2]

(1) a. pi-ka **mwe-ka** o-ni?!
rain-NOM WHAT-NOM come-QUE
'No way is it raining. (It isn't raining.)'

b. ku-ka **ettehkey** i pangpep-ulo
he-NOM HOW this way-in
sihem-ul thongkwaha-l
exam-ACC pass-CONN
swu.iss-keyss-ni?!
can-FUT-QUE
'No way could he pass the exam in this way. (He couldn't pass the exam in this way.)'

c. Mary-ka **eti** Seoul-ul
Mary-NOM WHERE Seoul-ACC
ttena-l swu.iss-keyss-ni?!
leave-CONN can-FUT-QUE
'No way would Mary be able to leave Seoul. (Mary wouldn't be able to leave Seoul.)'

d. ku-ka **encey** sip-nyen ceney chayk-ul
he-NOM WHEN 10-year ago book-ACC
ss-ess-ni?!
write-PST-QUE
'No way did he write the book ten years ago. (He didn't write the book ten years ago.)'

As seen from the English translations, NWHCs are used to express the speaker-oriented rhetorical/refutatory force and not the information-seeking force typically conveyed by ordinary *wh* or yes/no-questions (Cheung, 2008; 2009; Saruwatari, 2015; Yang, 2015). That is, positive NWHCs have the illocutionary force of a negative assertion, as in (1), and negative NWHCs have the illocutionary force of a positive assertion, as in (2).

[1] The abbreviations used for glossing Korean data include NOM: nominative, ACC: accusative, QUE: question, DECL: declarative, COP: copular, TOP: topic, CONN: connective, PST: past, IMP: imperative, EXCL: exclamative, MOD: modifier, FUT: future, and PROG: progressive.

[2] Cheung (2008) notes that there are only three NWH-items in Korean: *ettehkey* 'HOW', *encey* 'WHEN' and *eti* 'WHERE'. But, as given in (1a), the *wh*-phrase *mwe-ka* 'WHAT-NOM', where *mwe* is the contracted form of *mwues*, can also be used as an NWH-item. See Saruwatari (2015) for Japanese NWHCs using *nani-ga* 'WHAT-NOM and *doko-ga* 'WHERE-NOM'.

(2) {**ettehkey/mwe-ka**} John-i
HOW/WHAT-NOM John-NOM
tayhakwensayng-i ani-ni?!
graduate.student-NOM not-QUE
'No way is John not a graduate student. (He is a graduate student.)'

1.1 Differences from information-seeking and rhetorical *wh*-questions

NWHCs behave differently from both information-seeking and rhetorical *wh*-questions in some respects. First, while an ordinary *wh*-adjunct cannot cooccur with another adjunct of the same kind in the same clause, as in (3), such adjunct doubling is allowed in NWHCs, as in (1b-d) (Cheung, 2008; 2009).

(3) a. *Mary-ka eti Seoul-ey ka-ss-ni?
Mary-NOM where Seoul-to go-PST-QUE
'Where did Mary go to Seoul?'

b. *Mary-ka encey ocen hansi-ey Seoul-ey
Mary-NOM when a.m. 1-at Seoul-to
ka-ss-ni?
go-PST-QUE
'When did Mary go to Seoul at 1 a.m.?'

Second, the NWH-item *mwe-ka* 'WHAT-NOM' functions as an adverbial, just like the other NWH-items, though it is isomorphic to the ordinary *wh*-argument *mwe-ka* 'what-NOM'. Evidence supporting this idea is that the NWH-item WHAT can occur with a subject in an intransitive construction, as in (1a). In a similar vein, Yang (2015) takes Chinese NWH-items *shenme* 'WHAT' and *nali* 'WHERE', exemplified in (4), as *wh*-adverbials which are highly grammaticalized and have nothing to do with interrogativity (cf. Cheung, 2009).

(4) zhe-ci huiyi, {**nali/shenme**} ta hui
this-Cl meeting WHERE/WHAT he will
lai?!
come
'This meeting, it is not the case that he will come.'
(adapted from Yang (2015))

Third, NWH-adverbials have lost their lexical meanings. For example, the NWH-phrases *mwe-ka* 'WHAT-NOM' and *ettehkey* 'HOW' do not quantify over things/entities and manners/methods, respectively, but contribute only to the negative/positive assertion (Cheung, 2008; Yang, 2015).

Finally, NWHCs must be uttered after the interlocutor's statement as a way to express disapproval toward the interlocutor. That is, they cannot be uttered discourse-initially or out of the blue (Cheung, 2009; Yang 2015).

1.2 Research questions

This paper aims to address the following two research questions:

- Where is the base position of NWH-adverbials?
- Do they undergo LF-movement from their base position to a higher functional projection? If so, why?

As for the first question, the paper argues that NWH-adverbials are base-generated above or at the edge of IP (Cheung, 2008). As to the second question, the paper proposes that under Coniglio and Zegrean's (2012) split-ForceP hypothesis where ForceP is split up into two projections, namely C(lause) T(ype) and ILL(ocutionary Force), the NWH-phrase moves covertly from its base position to [Spec,CTP] to reflect its sensitivity to clause type and then moves to [Spec,ILLP] to derive the speaker-oriented rhetorical force.

1.3 Roadmap of the paper

In Section 2, I argue that NWH-adverbials originate above or at the edge of IP. In Section 3, I argue that NWH-phrases undergo LF-movement from their base position to the Force domain in the left periphery. In Section 4, I propose a novel two-step movement approach to NWHCs from the split-ForceP perspective. In Section 5, I summarize the main arguments of the paper.

2 Base-generation above or at the edge of IP

Through investigating how NWH-adverbials behave with respect to negative island effects and scopal interactions with quantifiers, I argue here that NWH-adverbials originate above or at the edge of IP.

• **Negative island effects:** The examples in (5) illustrate the *how-why* asymmetry with regard to

a Negative Island Effect (NIE), a phenomenon in which negation blocks extraction of certain (*wh-*)phrases (Rizzi, 1990; Shlonsky and Soare, 2011):

(5) a. Why didn't Geraldine fix her bike?

b. *How didn't Geraldine fix her bike? (Shlonsky and Soare 2011: (14))

The asymmetry receives a natural account if we follow Rizzi (2001) and Tsai (2008) in assuming that unlike manner/instrumental *how* base-generated below negation, reason *why* is directly merged in the CP region. On this view, *why* is immune to the NIE since it originates above negation, as illustrated in (6a), whereas *how* violates the NIE as it undergoes LF-movement to its scope position in the CP domain, as illustrated in (6b).

(6) a. $[_{CP}$ why $[_{IP}$ ... NegP ...]]

b. $[_{CP}$ how $[_{IP}$ ... NegP t_{how} ...]]

Note that the *how-why* asymmetry in NIEs also holds for ordinary *wh*-questions in Korean:

(7) a. Mary-nun **way** cha-lul kochi-ci
Mary-TOP why car-ACC fix-CONN
anh-ass-ni?
not-PST-QUE
'Why didn't Mary fix the car?'

b. *Mary-nun **ettehkey** cha-lul kochi-ci
Mary-TOP how car-ACC fix-CONN
anh-ass-ni?
not-PST-QUE
'How didn't Mary fix the car?'

As observed in (7a), *way* 'why' does not exhibit the NIE, just like English *why*, indicating that *way* is base-generated above negation (Ko, 2005; 2006). On the other hand, the ill-formedness of (7b) suggests that manner/instrumental *ettehkey*, which corresponds to English *how*, originates below negation.

With the *ettehkey-way* asymmetry described above in mind, let us consider the following NWHC examples:

(8) a. salam-i **ettehkey** cwuk-ci
human.being-NOM HOW die-CONN
anh-ni?!
not-QUE
'No way do human beings not die. (Human beings die.)'

b. John-i **mwe-ka** maykcwu-lul
John-NOM WHAT-NOM beer-ACC
masi-ci anh-ass-ni?!
drink-CONN not-PST-QUE
'It is not true that John didn't drink beer. (John drank beer.)'

As observed here, the NWH-adverbials *ettehkey* and *mwe-ka* are not sensitive to negation in the clause with which they are construed, indicating that they are base-generated above negation. Meantime, one may point out here that the insensitivity of NWH-adverbials to the NIE would be due to their non-movement at LF from their base position below NegP. However, as we will see below in Section 3, NWH-adverbials are taken to move at LF.

• **Scopal interactions with quantifiers:** The example in (9) illustrates that the negation evoked by NWH-adverbials always takes scope over the subject Quantifier Phrase (QP) (Cheung, 2008).

(9) (context: there are only three people in the group: John, Mary, and Mimi.)

{**mwe-ka/ettehkey**} motwu-ka
WHAT-NOM/HOW everyone-NOM
haksayng-i-ni?!
student-COP-QUE

(i) It is not the case that everyone is a student. (NEG > everyone)
(ii) For each person *x*, *x* is not a student. (*everyone > NEG)

(9i) is compatible with a situation where the speaker believes that some members of the group are not students (e.g. only John is a student). (9ii) is compatible with a situation where nobody in the group is a student. However, the second reading is unavailable. This scopal pattern may follow from the assumption that NWH-adverbials are base-generated above IP (or at the edge of IP as argued by Cheung (2008)). Since the NWH-adverbial is initially merged above

IP, it is impossible to interpret the NWH-item under the (raised) subject QP.[3]

3 LF-movement into ForceP

3.1 Intervention effects

Korean exhibits another asymmetry between *way* and other *wh*-operators, in that unlike the former, the latter cannot be preceded by a Scope Bearing Element (SBE) like *amwuto* 'anyone'. This phenomenon has been known as an intervention effect (Beck and Kim, 1997; Beck, 2006; among others).[4] Consider the following relevant examples:

(10) a. *amwuto **mwues-ul** mek-ci
anyone what-ACC eat-CONN
anh-ass-ni?
not-PST-QUE
'What did no one eat?'

b. **mwues-ul** amwuto mek-ci
what-ACC anyone eat-CONN
anh-ass-ni?
not-PST-QUE
'What did no one eat?'

(11) a. amwuto **way** sakwa-lul mek-ci
anyone why apple-ACC eat-CONN
anh-ass-ni?
not-PST-QUE
'Why did no one eat an apple?'

b. **way** amwuto sakwa-lul mek-ci
why anyone apple-ACC eat-CONN
anh-ass-ni?
not-PST-QUE
'Why did no one eat an apple?'

(10) shows that the *wh*-argument *mwues-ul* 'what-ACC' must precede the SBE *amwuto* 'anyone'. On the other hand, (11) illustrates that the *wh*-adjunct *way* can precede or follow the corresponding SBE.

To account for such an asymmetry in intervention effects, Ko (2005), adapting a proposal of Beck and Kim (1997), proposes the following intervention effect constraint on *wh*-movement at LF:

(12) *Intervention Effect* (Ko, 2005: 871):
At LF, a *wh*-phrase cannot be attracted to its checking (scope) position across an SBE.

Let us examine how the constraint captures the asymmetry, particularly under Ko's (2006) split-CP analysis of *wh*-licensing, according to which *way* in an interrogative clause is directly merged into its checking position [Spec,Int(errogative)P], while other *wh*-phrases covertly move to [Spec,Foc(us)P], higher than IntP, for feature checking.[5] In (10a), the *wh*-argument *mwues-ul* must undergo LF-movement to [Spec,FocP] to be licensed. However, the SBE *amwuto* induces the intervention effect by blocking the LF-movement, resulting in a derivational crash. This is why (10a) is ruled out. The well-formedness of (10b) is because the overt scrambling of the *wh*-argument over the SBE avoids the intervention configuration. In (11a), unlike the *wh*-argument, the *wh*-adjunct *way* can be preceded by the SBE. This is because *way* does not move at LF as it is initially licensed in its base position, i.e. [Spec,IntP], before the overt scrambling of the SBE over it.[6] The well-formedness of (11b) is simply because *way* is not located in the intervention configuration.

Now let us take a look at the following NWHCs regarding intervention effects:

(13) A: Nobody is a student here.

B: {**mwe-ka/ettehkey**} amwuto
WHAT-NOM/HOW anyone
haksayng-i ani-ni?!
student-NOM not-QUE
'It is not the case that nobody is a student here. (Some of the members are students.)'

B′: amwuto {?***mwe-ka**/??**ettehkey**}
anyone WHAT-NOM/HOW

[3] I leave further investigation of the exact base position of NWH-adverbials to future work.

[4] SBEs also include *man* 'only', *anh* 'not', *pakkey* 'only' (NPI), *to* 'also', *nwukwunka* '(non-specific) someone', and *nwukwuna* 'everyone' (Ko, 2005).

[5] For the split CP domain, Ko (2006) suggests only two functional heads, Int and Foc, for licensing ordinary *wh*-phrases and uses the terms C_{Int} and C_{Foc} to avoid unnecessary confusion with Rizzi's (1999, 2001) split-CP system in Italian in (i), where Int is configured higher than Foc.

(i) Force (Top) Int (Top) Foc (Top) Fin IP ... (Rizzi, 1999)

[6] If *way* occurs in an embedded declarative clause, it is required to move covertly to the matrix IntP[+Q] to take scope (Ko, 2005; 2006).

haksayng-i ani-ni?!
student-NOM not-QUE
'(int.) It is not the case that nobody is a student here. (Some of the members are students.)'

As shown in (13B′), the NWH-adverbials are not allowed to follow the SBE. If the intervention effect constraint in (12) is on the right track, the contrast between (13B) and (13B′) suggests that NWH-adverbials undergo LF-movement.

The sensitivity of NWH-adverbials to intervention effects induced by quantificational adverbs further supports the argument that NWH-phrases move at LF. To illustrate such an intervention effect, let us first look at the Hungarian data in (14).

(14) a. *Mindig **kit** hitá meg?
always who-ACC invited PV
'Who did you invite all the time?'

b. **kit** hitá meg mindig?
who-acc invited PV always
'Who did you invite all the time?' (adapted from den Dikken (2003))

The examples here illustrate that the *wh*-phrase *kit* 'who-ACC' cannot follow but must precede the adverb of quantification *mindig* 'always'. To explain this paradigm, Lipták (2001) suggests that the ill-formedness of sentences like (14a) is attributed to intervention effects: the quantificational adverb harmfully intervenes between the *wh*-phrase and the interrogative $C_{[+wh]}$, as roughly represented below.

(15) *$[_{CP}$ $C_{[+wh]}$ $[_{DistP}$ mindig $[_{FocP}$ $kit_{[+wh]}$ $[_{Foc}$ hitá [...]]]]]

To be more specific, the quantificational phrase, which occupies [Spec,Dist(ributive)P] higher than FocP, blocks the feature movement of the *wh*-phrase from [Spec,FocP] to $C_{[+wh]}$, resulting in a derivational crash.

Yang (2007; 2015) discusses Chinese NWHCs (in his term, refutatory *wh*-questions) in terms of the aforementioned intervention effect so as to suggest that NWH-items merged at FocP undergo covert movement to ForceP to derive the speaker's refutatory force. To illustrate, consider the following contrast:

(16) a. {*meitian/*changchang} {**nail/shenme**}
everyday/often WHERE/WHAT
ta hui lai?!
he will come
'Everyday/often it is not the case that he will come.'

b. {**nail/shenme**} ta {meitian/changchang}
WHERE/WHAT he everyday/often
hui lai?!
will come
'Everyday/often it is not the case that he will come.' (adapted from Yang (2007))

He argues that the deviance of (16a) is because the quantificational phrase like *meitian* 'everyday' and *changchang* 'often' blocks LF-movement of the NWH-phrase into ForceP, giving rise to the intervention effect within the CP field (Cheung, 2008). Meantime, there is no such intervention effect in (16b) since the NWH-phrase is located in a higher position than the SBE and thus freely moves to ForceP at LF.

When it comes to Korean NWHCs, the following examples illustrate that they exhibit the same intervention effect as Chinese counterparts:

(17) a. *hangsang {**mwe-ka/ettehkey**}
always WHAT-NOM/HOW
John-i sinmwun-ul ilk-ni?!
John-NOM newspaper-ACC read-QUE
'(int.) No way does John always read a newspaper.'

b. {**mwe-ka/ettehkey**} hangsang John-i
WHAT-NOM/HOW always John-NOM
sinmwun-ul ilk-ni?!
newspaper-ACC read-QUE
'No way does John always read a newspaper.'

Assuming that the quantificational phrase like *hansang* 'always' is sitting in [Spec,DistP] higher than FocP as argued by Lipták (2001), the contrast in (17) suggests that the NWH-phrase undergoes LF-movement from its base position to a higher functional projection above DistP in the CP region.[7]

[7] Yang (2015) takes Top(ic)P as the functional projection hosting quantificational adverbs.

3.2 The interaction with illocutionary force and clause type

It has been proposed that NWH-phrases move at LF to a higher functional projection. In this respect, then, two important questions arise as to (i) what is the functional projection to which NWH-phrases move at LF and (ii) why they undergo LF-movement to the assumed functional projection. In addressing the first issue, I argue here that NWH-phrases move covertly to ForceP, given that they closely interact with both clause type and illocutionary force encoded in ForceP (Rizzi, 1997; cf. Coniglio and Zegrean, 2012). In what follows, let us look at some evidence for the argument.[8]

The interaction of NWH-adverbials with illocutionary force is evidenced by their inability to occur in embedded clauses, as in (18): pragmatically, elements conveying the expressive force (i.e. the speaker's subjective opinion and attitude) can only be carried out by direct speech (Pan, 2015).

(18) *motun salam-i John-i
every person-NOM John-NOM
{**mwe-ka/ettehkey**} haksayng-i-nci
WHAT-NOM/HOW student-COP-QUE
a-ni?!
know-QUE
'(int.) Does every person know that John is not a student?'

If the NWH-phrase in (18) occurs in the matrix clause instead of the embedded one, then the resulting sentence becomes well-formed, as in (19). In this case, as one can expect, the NWH-phrase is only associated with the matrix clause, as seen from the English translation, since it cannot originate within the embedded clause.

(19) {**mwe-ka/ettehkey**} motun salam-i
WHAT-NOM/HOW every person-NOM
John-i haksayng-i-nci a-ni?!
John-NOM student-COP-QUE know-QUE
'It is not the case that every person knows whether John is a student or not.'

[8]Tsai (2008) argues that while Chinese causal *zenme* 'how' is placed at Int, denial *zenme* originates at the head of ForceP to reflect the change of illocutionary force, i.e. from eliciting information to denial.

NWH-adverbials' interaction with clause type can be verified by the fact that they can occur only in yes/no questions, as in (20a), but not in *wh*-questions, as in (20b), declaratives, as in (20c), imperatives, as in (20d), or exclamatives, as in (20e).[9]

(20) a. {**mwe-ka/ettehkey**} Mary-ka
WHAT-NOM/HOW Mary-NOM
haksayng-i-ni!?
student-COP-QUE
'It is not true that Mary is a student.'

b. *{**mwe-ka/ettehkey**} nwu-ka
WHAT-NOM/HOW who-NOM
haksayng-i-ni!?
student-COP-QUE
'(int.) It is not true that Mary is a student.'

c. *{**mwe-ka/ettehkey**} Mary-ka
WHAT-NOM/HOW Mary-NOM
haksayng-i-ta.
student-COP-DECL
'(int.) It is not true that Mary is a student.'

d. *{**mwe-ka/ettehkey**} Mary-ka
WHAT-NOM/HOW Mary-NOM
ttena-la!
leave-IMP
'(int.) It is not true that Mary left.'

e. *{**mwe-ka/ettehkey**} Mary-ka
WHAT-NOM/HOW Mary-NOM
yeyppu-kwuna!
pretty-EXCL
'(int.) It is not true that Mary is pretty.'

This distributional constraint may indicate that NWH-adverbials undergo covert movement to ForceP to reflect their sensitivity to clause type.

In what follows, I will address the remaining issue of why NWH-adverbials undergo LF-movement to ForceP, within Coniglio and Zegrean's (2012) split-ForceP framework.

[9]It is possible for the NWH-word to occur in a yes/no question with a *wh*-indefinite like *mwe* (the contracted form of *mwues*), as shown in (i).

(i) **mwe-ka** John-i mwe-lul
WHAT-NOM John-NOM something-ACC
mek-ess-ni?!
eat-PST-QUE
'No way did John eat something.'

4 Proposal

4.1 Similarities with adverb-based discourse particles

The close interaction of NWH-adverbials with both illocutionary force and clause type is reminiscent of adverb-based discourse particles like Italian *tanto*. Dohi (2020) suggests that sentence-initial *tanto* interacts with clause type, given that it occurs only in *wh*-questions, as in (21a), or declaratives, as in (21b), but not in other clause types like imperatives, as in (21c).

(21) a. *Tanto* cosa ci stai a fare qua?
Prt what there you.stay to do here
'What are you going to do here anyway? (You have nothing to do here.)'

b. *Tanto* non succederà mai.
Prt not will.happen never
'It will never happen in any case.'

c. **Tanto* lascialo sul tavolo.
Prt leave.it on.the table
(Dohi, 2020)

In addition, he suggests that *tanto* also interacts with illocutionary force, in that it pragmatically functions to modify the original illocutionary force of the utterance where it occurs. To illustrate this, let us consider (22).

(22) a. cosa ci stai a fare que?
what there you.stay to do here
'What are you going to do here?'

b. *Tanto* cosa ci stai a fare que?
Prt what there you.stay to do here
'What are you going to do here anyway? (You have noting to do here.)'
(Dohi, 2020: 5)

(22a) can be interpreted as an information-seeking question (or a rhetorical one), but if *tanto* is inserted into the utterance, the result in (22b) is interpreted only as a rhetorical question, which has been derived by the discourse particle modifying the original information-seeking force on Dohi's view.

To account for the peculiar properties of *tanto*, Dohi modifies Zimmermann's (2004) analysis of the German discourse particle *wohl*, within Coniglio and Zegrean's (2012) split-ForceP hypothesis where ForceP is split up into two different projections, namely C(lause) T(ype) and ILL(ocutionary Force). By so doing, he argues that the adverb-based discourse particle *tanto* is base-generated in [Spec,CTP] and enters into a Spec-Head agreement relationship with the CT head codified as a clause-type operator such as *decl* for declaratives and *int* for interrogatives. This agreement relationship captures the discourse particle's sensitivity to clause type. He further argues that *tanto* merged in [Spec,CTP] moves at LF to [Spec,ILLP] to derive the rhetorical force through modifying the default illocutionary force codified as a privative operator like *assert(ion)* for declaratives and *?* for interrogatives. On this split-ForceP analysis, for example, (22b) is derived as follows:

(23) $[_{\text{ILLP}}\ \text{Tanto}_i\ ?\ [_{\text{CTP}}\ t_i\ \textit{int}\ [_{\text{FocP}}\ \text{cosa}\ [_{\text{VP}}$ ci stai a fare que]]]]?

4.2 A split-ForceP approach to NWHCs

Given the similarities between NWH-adverbials and adverb-based discourse particles like *tanto* in closely interacting with both clause type and illocutionary force, it would be reasonable to apply Dohi's (2020) split-ForceP analysis to NWHCs.[10] Therefore, from the split-ForceP perspective, I propose the following two-step movement approach to licensing NWH-adverbials with no interrogativity:

- **Step 1**: The NWH-adverbial first moves covertly from its base position to [Spec,CTP] to agree with a question morpheme like *ni* with [+Q, -WH], in a Spec-Head relationship, to reflect its sensitivity to clause type, i.e., obligatory occurrence in yes/no questions.

- **Step 2**: The NWH-adverbial then moves to [Spec,ILLP] to derive the speaker-oriented rhetorical force through modifying the original information-seeking force codified as the privative operator *?* in ILL^0.[11]

[10] Here I avoid discussing whether NWH-adverbials are adverb-based discourse particles. I leave the issue to future research.

[11] Yang (2015) notes that the speaker-oriented rhetorical force is strong enough to override the original interpretation of an interrogative *wh*-question.

On this split-ForceP analysis, for instance, the NWHC in (20a), repeated below in (24a), is assumed to be derived like (24b):

(24) a. {**mwe-ka/ettehkey**} Mary-ka
WHAT-NOM/HOW Mary-NOM
haksayng-i-ni!?
student-COP-QUE
'It is not true that Mary is a student.'

b. [$_{ILLP}$ mwe-ka$_i$/etthekey$_i$ [$_{CTP}$ t$_i$ [$_{TP}$ Mary-ka haksayng-i]-ni$_{[+Q, -WH]}$]*?*]

In the meantime, *wh*-phrases used in ordinary information-seeking questions do not need to undergo covert movement into the split-ForceP region, since, unlike NWH-adverbials, they do not modify the original interrogative force and are insensitive to clause type, occurring in (embedded) declaratives as in (25a), (embedded) imperatives as in (25b), and exclamatives as in (25c).

(25) a. ne-nun [Mary-ka **mwues-ul**
you-TOP Mary-NOM what-ACC
mek-ess-ta-ko] sayngkakha-ni?
eat-PST-DECL-COMP think-QUE
'What do you think Mary ate?'

b. ne-nun [Mary-eykey **mwues-ul**
you-TOP Mary-to what-ACC
mek-ula-ko] malhayss-ni?
eat-IMP-COMP said-QUE
'What did you order Mary to eat?'

c. nay yecachinkwu-ka **elmana**
my girlfriend-NOM how
yeyppu-tako!
be.pretty-EXCL
'My girlfriend is really pretty!'

4.3 The assumed left peripheral map

Based on the observations so far, we can postulate the following left periphery for ordinary *wh*-phrases (Ko, 2006) and NWH-adverbials at LF:

(26) [$_{ILLP}$ **NWH$_i$** [$_{CTP}$ t$_i$ [$_{DistP}$ [$_{FocP}$ **wh** [$_{IntP}$ **way** [$_{IP}$ ...]]]]]]

According to the proposed LF structure, we can predict that different from NWH-phrases, ordinary *wh*-phrases may not be sensitive to the intervention effect induced by quantificational adverbs, since they are assumed not to move covertly to the split-Force domain and DistP is located higher than both FocP and IntP where ordinary *wh*-phrases are licensed. This prediction is borne out by the following attested examples:

(27) a. hangsang **way** John-un sinmwun-ul
always why John-TOP newspaper-ACC
ilk-ni?
read-QUE
'Why does John always read a newspaper?'

b. hangsang **mwues-ul** way mek-ko
always what-ACC why eat-PROG
iss-ni?
be-QUE
'Why are you always eating what?'

In (27a), the SBE *hangsang* can precede *way* 'why' without inducing the intervention effect since *way*, directly merged in [Spec,IntP], does not move across the SBE at LF. In (27b), the *wh*-argument *mwues-ul* has scrambled over *way*, indicating that it is located in the CP region in overt syntax. In this case, the *wh*-argument can be preceded by the SBE, simply because DistP is configured higher than FocP. That is, the SBE in [Spec,DistP] does not have an effect on LF-movement of the *wh*-argument to its checking position [Spec,FocP].

5 Summary

This paper has investigated the syntax of negative *wh*-constructions in Korean, which, to my knowledge, has not been much discussed in the literature. Under the split-ForceP hypothesis, it has been argued that NWH-adverbials like *mwe-ka* 'WHAT-NOM' and *ettehkey* 'HOW', which are base-generated above or at the edge of IP, covertly move to [Spec,CTP] to reflect their sensitivity to clause type and then move to [Spec,ILLP] to turn the original information-seeking force into the speaker-oriented rhetorical force. I hope the discussion presented in this paper contributes to a better understanding of the left periphery of the clause in Korean.

Acknowledgments

I would like to thank the three anonymous reviewers for their comments and suggestions.

References

Beck, Sigrid. 2006. Intervention Effects Follow from Focus Interpretation. *Natural Language Semantics*, 14:1-56. DOI: https://doi.org/10.1007/sl 1050-005-4532-y

Beck, Sigrid and Shin-Sook Kim. 1997. On wh-and operator scope in Korean. *Journal of East Asian Linguistics*, 6(4):339–384. DOI: https://doi.org/10.1023/A:1008280026102

Cheung, Lawrence Yam-Leung. 2008. *The negative wh-construction*. PhD dissertation, UCLA.

Cheung, Lawrence Yam-Leung. 2009. Negative *wh*-construction and its semantic properties. *Journal of East Asian Linguistics*, 18:297-321. DOI: https://doi.org/10.1007/s10831-009-9051-2

Coniglio, Marco and Iulia Zegrean. 2012. Splitting up Force: Evidence from discourse particles. In Lobke Aelbrecht, Liliane Haegeman, and Rachel Nye (eds.), *Main clause phenomena: New horizons*, 229-255. Amsterdam: John Benjamins.

den Dikken, Marcel. 2003. On the morphosyntax of *wh*-movement. In Cedric Boeckx and Kleanthes Grohmann (eds.), *Multiple wh-fronting*, 77-98. Amsterdam: John Benjamins.

Dohi, Atsushi. 2020. CP-internal discourse particles and the split ForceP hypothesis. *Lingua* 233. DOI: https://doi.org/10.1016/j.lingua.2019.102757

Ko, Heejeong. 2005. Syntax of why-in-situ: Merge into [Spec, CP] in the overt syntax. *Natural Language & Linguistic Theory*, 23(4):867-916. DOI: https://doi.org/10.1007/s11049-004-5923-3

Ko, Heejeong. 2006. On the Structural Height of Reason *Wh*-Adverbials: Acquisition and consequences. In N. C. Lisa Lai-Shen Cheng and Norbert Corver (eds.), *Wh-Movement: Moving on*, 319-349. MIT Press, Cambridge, MA.

Lipták, Anikó. 2001. *On the syntax of wh-items in Hungarian*. Doctoral dissertation, University of Leiden.

Pan, Victor Junnan. 2015. Mandarin peripheral construals at the syntax-discourse interface. *The Linguistic Review*, 32(4):819-868. DOI: https://doi.org/10.1515/tlr-2016-1005

Rizzi, Luigi. 1990. *Relativized Minimality*. MIT Press, Cambridge, MA.

Rizzi, Luigi. 1997. The Fine Structure of the Left Periphery. In L. Haegeman (ed.), *Elements of Grammar*, 281-337, Kluwer, Dordrecht.

Rizzi, Luigi. 1999. On the Position of "Int(errogative)" in the Left Periphery of the Clause. Ms, Universitá di Siena.

Rizzi, Luigi. 2001. On the position of 'Int(errogative)' in the left periphery of the clause. In Guglielmo Cinque and Giampaolo Salvi (eds.) *Current studies in Italian syntax*, 267-296. Amsterdam: Elsevier.

Saruwatari, Asuka. 2015. Wh-NP rhetorical questions in Japanese and Chinese. *Shizen Gengo-e-no Rironteki Apurooti* [Theoretical Approaches to Natural Languages]. 21-30. Osaka University.

Shlonsky, Ur and Gabriela Soare. 2011. Where's 'Why'? *Linguistic Inquiry*, 42:651-669. DOI: https://doi.org/10.1162/LING$_{\alpha 0}$0064

Tsai, Wei-Tien Dylan. 2008 Left periphery and *how-why* alternations. *J East Asian Linguist* ,17(83):83-115. DOI: https://doi.org/10.1007/s10831-008-9021-0

Yang, Barry Chung-Yu. 2007. Rhetoric/Disapproving *wh* and Intervention Effect. unpublished manuscript. Harvard.

Yang, Barry Chung-Yu. 2015. Locating *Wh*-Intervention Effects at CP. In Wei-Tien Dylan Tsai (ed.), *The Cartography of Chinese Syntax*, 153-186. Oxford: Oxford University Press.

Zimmermann, Malte. 2004. Discourse particles in the left periphery. *ZAS Papers in Linguistics*, 35:543-566.

Generation and Evaluation of Concept Embeddings Via Fine-Tuning Using Automatically Tagged Corpus

Kanako Komiya Daiki Yaginuma Masayuki Asahara Hiroyuki Shinnou

Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
{kanako.komiya.nlp, 18nm740n, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo, Japan
masayu-a@ninjal.ac.jp

Abstract

Word embeddings are used in various fields of natural language processing. The use of word embeddings and concept or word sense embeddings demonstrated effectiveness in many tasks, such as machine translation and text summarization. However, it is difficult to obtain a sufficiently large concept-tagged corpus, as the annotation of concept-tags is time-consuming. Therefore, in this paper, we propose a method for generating concept embeddings of Word List by Semantic Principles, a Japanese thesaurus, using both a corpus tagged by an all-words word sense disambiguation (WSD) system and a manually tagged corpus. We generated concept embeddings via fine-tuning using both an automatically tagged corpus and a small manually tagged corpus. In this paper, we propose a novel method of evaluating concept embeddings using the tree structure of Word List by Semantic Principles. Experiments revealed the effectiveness of fine-tuning. The best performance was achieved when the concept embeddings were initially trained with a corpus tagged by an all-words WSD system and retrained with a manually tagged corpus.

1 Introduction

In this paper, we propose a technique for generating concept embeddings using fine-tuning and two types of corpora. In recent years, word embeddings, which are distributed representations of words with low-dimensional vectors, and concept [1] (or word sense) embeddings demonstrated their effectiveness in a number of tasks, such as machine translation and text summarization.

Word embeddings are usually generated using text corpora. It is possible to generate concept embeddings by the same method used to generate word embeddings if the word sequence (i.e., text corpus) is replaced with a concept sequence constructed from a concept-tagged corpus. However, it is difficult to obtain a sufficiently large concept-tagged corpus because the annotation of concept tags is time-consuming. There have been several studies that assigned word senses using the all-words word sense disambiguation (WSD) method (Edmonds and Cotton, 2001), (Snyder and Palmer, 2004), (Navigli et al., 2007), (Iacobacci et al., 2016), (Raganato et al., 2017a), (Raganato et al., 2017b), (Suzuki et al., 2018), (Shinnou et al., 2018). As a result, it is possible to create a concept-tagged corpus using the methods proposed in these studies. However, the results of all-words WSD systems are not always correct; therefore, an automatically tagged corpus created via all-words WSD may not be suitable for generating concept embeddings.

In this paper, we generate concept embeddings of Word List by Semantic Principles (WLSP) (National Institute for Japanese Language and Linguistics, 1964), a Japanese thesaurus, from manually and automatically tagged corpora. First, concept embeddings are generated from a concept-tagged corpus tagged by an all-words WSD system and are fine-tuned using a small, highly accurate corpus in which the concept tags are manually annotated. For comparison, we also generate the following concept em-

[1] Concept refers to a meaning unit of Word List by Semantic Principles.

beddings: (1) concept embeddings generated from only a small, highly accurate corpus in which the concept tags are manually annotated, (2) concept embeddings generated from only a concept-tagged corpus tagged by an all-words WSD system, and (3) concept embeddings initially trained with a small, highly accurate corpus in which the concept tags are manually annotated and fine-tuned using a concept-tagged corpus tagged by an all-words WSD system. The obtained concept embeddings are evaluated by rankings measured by the distances between the concept embeddings based on the tree structure of WLSP, which is a proposed evaluation method in this paper.

2 Related Work

In recent years, word embeddings have been widely used in various fields of natural language processing. In addition, there have been a number of studies on the generation of concept (or word sense) embeddings.

For example, a study by Ouchi et al. (2016), to construct distributed representations of word senses, the authors utilized the distributed representations of synonyms of each word sense. In addition, Yamaki et al. (2017) proposed a method for constructing sense embeddings using training data with sense tags and the multi-sense skip-gram (MSSG) model, which considers the frequency of each word sense. However, these studies did not use a sense-tagged corpus, but rather, a regular text corpus and word embeddings.

Word embeddings are usually generated using a text corpus that is a word sequence. Concept or word sense embeddings can be generated using the same tools as for a sense-tagged corpus, that is, a word sense sequence or concept sequence instead of a text corpus. However, it is generally difficult to obtain a sufficiently large sense-tagged corpus, as only several are available and most are small.

If there are insufficient tagged corpora, automatic generation of tagged corpora may be helpful. A concept-tagged corpus can be automatically created with the all-words WSD system. There are several studies on all-words WSD systems. For example, in studies by Raganato et al. (2017a) and Shinnou et al. (2018), all-words WSD is considered a labeling problem in which every word is assigned a concept tag. Using an automatic tagger, it is possible to create a concept-tagged corpus. However, an automatic tagger does not always produce correct results. For example, there may be cases in which concept tags are not assigned to new words. In these cases, the concept-tagged corpus would not be suitable for generating concept embeddings.

Therefore, in this study, we generate concept embeddings of WLSP using two types of corpora: a large corpus in which the concept tags are assigned using the all-words WSD method and a manually tagged corpus.

3 Generation of Concept Embeddings

We generated four types of vectors using two corpora tagged with concepts from WLSP.

3.1 WLSP

WLSP is a Japanese thesaurus in which a word is classified and ordered according to its meaning. A WLSP record is composed of the record ID number, lemma number, record type, class, division, section, article, concept number, paragraph number, small paragraph number, word number, lemma with explanatory note, lemma without explanatory note, reading and reverse reading. The concept number consists of a category, medium item, and classification item. In WLSP, some words are polysemous; for example,“子供 (child or children)” is a polyseme, and two concepts are registered in WLSP: 1.2050 and 1.2130 (Table 1).

The tree structure of WLSP is illustrated in Figure 1.

3.2 Corpora

In this study, we used two concept-tagged corpora based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). The first corpus is a large corpus in which concept tags were automatically assigned using the all-words WSD method. We used an all-words WSD tagger proposed by (Shinnou et al., 2018). Hereinafter, this corpus is referred to as the all-words WSD corpus. The second corpus is a small corpus in which concept tags were manually assigned. We used the annotation data of WLSP by the National Institute of Japanese Language and Linguistics (Kato et al.,

Concept number	Class	Division	Section	Article
1.2050	Nominal words	Agent	Human	Young or old
1.2130	Nominal words	Agent	Family	Child or descendant

Table 1: Concept tags and their corresponding class, division, section, and article of "子供 (child or children)" from Word List by Semantic Principles

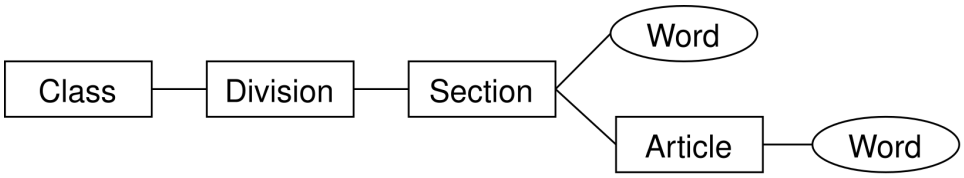


Figure 1: Tree structure of Word List by Semantic Principles

2018). This corpus is in its infancy. Hereinafter, this corpus is referred to as the manual corpus. There are two types of BCCWJ: the core and non-core data. For the core data, the word tokenization is manually conducted, but for the non-core data, word tokenizer, MeCab with Unidic dictionary is used for the word tokenization. The core data includes approximately 1,300,000 words and the non-core data includes approximately 25,800,000,000 words. The core data is included in the non-core data. We used the non-core data including the core data for the all-words WSD corpus, with the concept tag annotation via the all-words WSD system. The manual corpus is the part of the core data with manual annotation of the concept tags, which includes approximately 340,000 words.

Examples of the text corpus and a generated concept sequence are presented in Table 2. In the table, an original Japanese text, its English translation and concept sequence are shown. The concepts of "なく" and "ない" are both 3.1200 because they are the same words after lemmatization. Table 3 presents the number of words, vocabulary, and concepts in each corpus.

3.3 Vectors

In this study, word2vec[2] (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c) was used to generate concept embeddings. Then, fine-tuning was performed. Fine-tuning is a method in which generated distributed representations are given as initial values and retrained with a new corpus. The following four types of concept embeddings were created:

- All-words WSD vector: concept embeddings were trained with the all-words WSD corpus.
- All-words WSD-fine vector: concept embeddings were trained with the all-words WSD corpus and retrained with a manual corpus.
- Manual vector: concept embeddings were trained with a manual corpus.
- Manual-fine vector: concept embeddings were trained with a manual corpus and retrained with the all-words WSD corpus.

When fine-tuning the embeddings, vectors of the new words in the new corpus were generated if the number of occurrences of the new words exceeded the threshold value.

4 Evaluation of Concept Embeddings

We evaluated the concept embeddings using WLSP. Because WLSP has a tree structure, we assume that concepts that belong to the same node are similar to each other. Figure 2 presents an example of leaves of WLSP. In this figure, we assume that the concept of *wolf* is closer to that of *hyena* than that of *cat* or *dog*. Based on this assumption, evaluation of the generated concept embeddings was performed.

[2] https://code.google.com/archive/p/word2vec/

Text	モノでなく心ではないのか
English translation	It is not a thing but a heart, isn't it?
Concept sequence	1.4000 で 3.1200 1.3000 では 3.1200 の か

Table 2: Example of concept-tagged corpus

Concept Embeddings	Words	Vocabulary	Concepts
All-words WSD corpus	23,968,826	75,028	851
Manual corpus	347,094	3,164	916

Table 3: Number of words, vocabulary, and concepts in each corpus

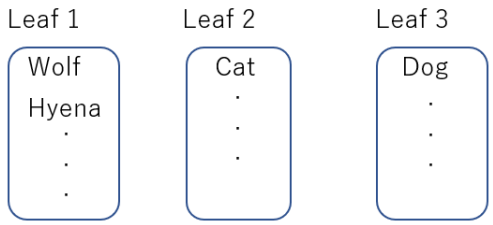


Figure 2: Example of leaves of Word List by Semantic Principles

4.1 Evaluation Procedure

The evaluation procedures were as follows.

1. For each concept c of the concept embeddings e, identify a corresponding leaf node n in WLSP.

   For example, if c is the concept of *wolf*, the corresponding node n includes concepts such as *hyena*. In Figure 2, n is Leaf 1. In this method, we assume that every concept has at least two words so that the distance between them can be calculated.

2. Obtain a sibling leaf node set N of n.

   A sibling leaf node set N includes a node that contains a concept such as *cat* and another node that contains a concept such as *dog*. In Figure 2, N includes Leaves 2 and 3.

3. Calculate d_c, the average distance between e and the concept embeddings of all concepts in n except for c.

   For this step, we calculated d_c, the average distance between the concept embeddings of *wolf* and the concept embeddings of *hyena* and other concepts in n (Leaf 1). We used the arithmetic mean to average the distance.

4. Calculate the average distances $d_1 ... d_{|N|}$ between e and the concept embeddings of all concepts in each leaf node in N.

   We calculated the average distance between concept embeddings of *wolf* and the concept embeddings of all concepts from the node containing *cat*, and obtained d_1. Likewise, we calculated the average distance between the concept embeddings of *wolf* and the concept embeddings of all concepts from the node containing *dog*, and obtained d_2. Following this step, we obtained the averaged distances $d_1 ... d_{|N|}$.

5. Obtain the ranking of n compared with all nodes in N based on the average distance from e_i.

   We compared $d_1 ... d_{|N|}$ and d_c, and obtained the ranking of d_c. For example, if d_c was the second shortest in $d_1 ... d_{|N|}$ and d_c, n was in second place.

6. Obtain the closest distance d_{close} and the closest leaf node to e based on the average distance.

   We obtained the closet leaf node to e. For example, if the closest leaf node to the concept *wolf* was the node that contained the concept *dog*, d_2 would be the shortest, which signifies that $d_{close} = d_2$.

7. Obtain $d_c - d_{close}$.

We calculated $d_c - d_{close}$, which is the difference between the average distances from the concept in first place. In other words, we calculated the difference between the average distance between *wolf* and concepts such as *hyena*, which is the node that *wolf* belongs to in WLSP, and the average distance between *wolf* and concepts such as *dog*, which was in first place. If all rankings of n were first place, the difference would be zero.

In this manner, we evaluated the concept embeddings that were generated using ranking and $d_c - d_{close}$.

4.2 Experimental Settings

For the parameters used in the calculation of word2vec, we used 200 dimensions, 5 window sizes, 1,000 batch sizes, and 5 iterations. We used CBOW as the algorithm. The training parameters used for fine-tuning were identical to the ones used when the original concept embeddings were generated in advance. Cosine similarity was used to compare the distances between the generated concept embeddings.

4.3 Results

Table 4 presents the average ranking of the correct nodes, which are the nodes that each concept whose embeddings were generated by this method belonged to. Table 4 also displays the average difference between the closest leaf node and the correct nodes, and the average number of leaf nodes. This table indicates that the poorest ranking of each concept embeddings was 6.868 for the manual vector. Because the average number of leaf nodes was 42, the average ranking of a random selected node was approximately 21. This suggests that, even when concept embeddings were generated using the worst method, the ranking of the nodes produced better results than when the random baseline was used.

5 Discussion

Table 4 indicates that the average ranking and difference of the all-words WSD-fine vector were smaller than those of the all-words WSD vector. In addition, the average ranking and difference of manual-fine vector were smaller than those of the manual vector. Smaller values of the average ranking and difference indicate better performance; therefore, these results demonstrate that fine-tuning improved the vectors. Table 4 also indicates that the results of the all-words WSD vector were superior to those of the manual vector, while the all-words WSD-fine vector was superior to the manual-fine vector. The poorest results were associated with the manual vector. These results suggest that the all-words WSD corpus is effective for generating concept embeddings without fine-tuning or for initial training of fine-tuning. We believe that a large corpus is necessary for generating improved word embeddings. We used the same parameters of word2vec for all vectors, which were tuned so that the results of the manual vector, the method with the poorest performance, could achieve the best performance. The other three vectors (i.e., all-words WSD vector, all-words WSD-fine vector, manual-fine vector) could be improved if the parameters were tuned for each method. This is because the results of vectors often improve when the parameters are tuned depending on the size and characteristics of the corpora. Table 5 presents the evaluation results of the all-words WSD vector and manual vector generated with 10 iterations. Other parameters are identical to those used in the experiments presented in Table 4. The results in Table 5 are inferior to those in Table 4; therefore, extensive experiments are necessary to tune the parameters suitable for each corpus.

The number of words in the all-words WSD corpus was approximately 69 times larger than the number of words in the manual corpus (see Table 3). In addition, according to Shinnou et al. (2018), the accuracy of the WSD system was approximately 80% for all words and approximately 70% for all ambiguous words in the test corpus (the annotation data of WLSP). In our experiments, the test corpus would be identical to the manual corpus and sub-corpus of the all-words WSD corpus if concept tags were removed and manually tagged. Therefore, we assume that the accuracy of the all-words WSD corpus would be approximately 70% or 80%. The results of the concept embeddings trained with the all-words WSD corpus were superior to the results of the concept embeddings trained with the manual corpus regardless of whether fine-tuning was used. This demonstrates that the all-words WSD corpus was superior

Concept Embeddings	Avg. Ranking	Avg. Difference from First Place	Number of Leaf Nodes
All-words WSD vector	2.945	0.059	42
All-words WSD-fine vector	2.644	0.046	42
Manual vector	6.868	0.102	42
Manual-fine vector	3.143	0.049	42

Table 4: Evaluation by ranking measured by distance

Concept Embeddings	Avg. Ranking	Avg. Difference from First Place	Number of Leaf Nodes
All-words WSD vector	3.217	0.043	42
Manual vector	7.52	0.105	42

Table 5: Evaluation by ranking using distance with 10 iterations

to the manual corpus in generating concept embeddings. In other words, our experiments revealed that the corpus that was concept-tagged with 70% or 80% accuracy and whose size was approximately 69 times larger, was more suitable for generating concept. However, it cannot be claimed that when generating concept embeddings, the corpus size is more important than the accuracy of the concept tags of the corpus. Therefore, we conducted additional experiments to investigate the effect of the size of the all-words WSD corpus. Table 6 presents the average ranking of correct nodes, average difference from the concept in first place, and the number of leaf nodes according to the size of the all-words WSD corpus. We tested 10% to 100% of the size of the entire corpus in increments of 10%. This figure indicates that the average ranking monotonically improved from 10% to 60%, worsened at 70%, 80% and 90%, and achieved the best value when the entire corpus was used.

Finally, according to Table 4, we can observe the effect of order of the data used for training and retraining of word-embeddings. All-words WSD-fine vector and manual-fine vector use both the manual corpus and the all-words WSD corpus. The difference of two method is order of the data. It indicates that not only the size of the data but also the order of the data used for training and fine-tuning is important to improve the quality of word embeddings.

However, additional experiments are necessary to investigate the relationship between accuracy and corpus size.

For future work, other algorithm for word2vec, skip-gram can be tried instead of CBOW algorithm. Also, other word embeddings such as GloVe or fastText could be other options.

6 Conclusion

In this study, we generated concept embeddings using a concept-tagged corpus that was tagged by an all-words WSD system, and using fine-tuning. In addition, we evaluated the concept embeddings using rankings measured by the distances between the concept embeddings based on the tree structure of WLSP. We compared four concept embeddings: 1) concept embeddings that were trained with a concept-tagged corpus tagged by an all-words WSD system, 2) concept embeddings that were trained with a small and manually tagged corpus, 3) concept embeddings of 1) that were fine-tuned with a small and manually tagged corpus, and 4) concept embeddings of 2) that were fine-tuned with a cocept-tagged corpus tagged by an all-words WSD system. Experiments revealed that fine-tuning was effective in generating better concept embeddings when we utilized a small, manually tagged corpus and a corpus that was concept-tagged by an all-words WSD system. The all-words WSD-fine vector, which represented the concept embeddings initially trained with a large corpus automatically tagged by an all-words WSD system and fine-tuned with a small, manually tagged corpus, was superior when the concept embeddings were evaluated using the tree structure of WLSP.

Percentage of corpus used	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
All-words WSD vector	5.458	4.531	4.156	4.055	3.843	3.707	3.848	3.705	3.770	3.455
All-words WSD-fine vector	4.689	4.004	3.750	3.663	3.449	3.474	3.470	3.447	3.613	3.087
manual-fine vector	5.184	4.694	4.331	4.205	3.896	3.888	4.054	3.917	4.017	3.619

Table 6: Evaluation by ranking using distance according to the size of the all-words word sense disambiguation (WSD) corpus

Acknowledgments

This work was supported by JSPS KAKENHI Grants Number 18K11421, 17H00917, and a project of the Center for Corpus Development, NINJAL.

References

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of *SEMEVAL 2001*, pages 1—5.

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.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. 2018. Annotation of 'word list by semantic principles' labels for the balanced corpus of contemporary written Japanese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December. Association for Computational Linguistics.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop 2013*, pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 1–9.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*, pages 746–751.

National Institute for Japanese Language and Linguistics. 1964. *Word List by Semantic Principles*. Shuuei Shuppan, In Japanese.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of *SEMEVAL 2007*, pages 30–35.

Katsuyuki Ouchi, Hiroyuki Shinnou, Kanako Komiya, and Minoru Sasaki. 2016. Construction of word sense embeddings from word embeddings using synonyms. *Proceedings of NLP 2016 (in Japanese)*, pages 99–102.

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, Rui Suzuki, and Kanako Komiya. 2018. All-words wsd with wlsp number as s sense label using a bidirectional lstm. *Proceedings of the Language Resources Workshop 2018 (in Japanese)*, pages 2–4.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of *SEMEVAL 2004*, pages 41–43.

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.

Shoma Yamaki, Hiroyuki Shinnou, Kanako Komiya, and Minoru Sasaki. 2017. Construction of word sense embeddings using training data. *Proceedings of NLP 2017 (in Japanese)*, pages 78–81.

Towards a Linguistically Motivated Segmentation for a Simultaneous Interpretation System

Youngeun Koo[1], Jiyoun Kim[1], Jungpyo Hong[1], Munpyo Hong[1*] and Sung-Kwon Choi[2]
[1]Dept. of German Linguistics & Literature, Sungkyunkwan University,
25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Korea
[2]Language Intelligence Research Section, Electronics and Telecommunications Research Institute(ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
{sarah8835,kite92,jphong2800,skkhmp}@skku.edu;choisk@etri.re.kr

Abstract

For simultaneous interpretation, it is very important to identify appropriate segmentation boundaries so that the source text can be translated accurately and promptly. This paper proposes four different segmentation methods for a simultaneous interpretation system. These methods are designed considering the balance between translation accuracy and translation latency. They employ various linguistic features such as prosodic, part-of-speech (POS), dependency, discourse, and cognitive features. This paper conducts experiments on segmentation in English to Korean and Korean to English simultaneous interpretation. Our finding shows that different segmentation method should be applied depending on the source language.

1 Introduction

Simultaneous interpretation aims to accurately translate what is being said in a source language into a target language quickly. To this aim, a strategy that segments the source text at appropriate points is often used by both human interpreters and simultaneous interpretation systems. Ideally, simultaneous interpretation systems should provide interpretation results as soon as possible while minimizing translation latency. However, there is a trade-off between translation accuracy and latency. The longer the segmentation unit is, the higher the translation accuracy will be, while the latency gets worse. In contrast, if the segmented unit is short, the latency will be better; however, the accuracy tends to be worsened.

In this paper, we investigate various segmentation features to determine the optimal segmentation points. The features were designed through a linguistic investigation into the prosodic, POS, dependency, and discourse-level characteristics in simultaneous interpretation. Also, we tried to find out what the appropriate segmentation length should be. In this paper, we propose four methods to derive optimal segmentation points employing these features. The segmentation features and methods considering both translation accuracy and latency may help to improve the performance of a simultaneous interpretation system.

In section 2, related works are introduced. In section 3 we suggest linguistic features for the segmentation. Section 4 shows our experimental setup with proposed features and methods and analyses the results. Finally, Section 5 concludes this paper and discusses future researches.

2 Related Works

Previous researches showed various approaches for investigating segmentation boundaries. The simplest way is to find a possible sentence unit (Cettolo and Federico, 2006; Sridhar et al., 2013b). Sridhar et al. (2013b) found segmentation boundaries based on predicting possible sentence end. Also, Sridhar et al. (2013b) utilized commas in sentences for segmentation.

* Corresponding author

Another approach for segmentation boundary detection is to use POS of the source text (Stolcke and Shriberg, 1996; Sridhar et al., 2013b; Nakabayashi et al., 2019). Stolcke and Shriberg (1996) tested two models for segmentation based on the POS of an input. The first model used POS tags labeled on every token and the second model used both POS and 'segmentation related' information, such as filled pause and discourse markers like 'okay', 'well'. Nakabayashi et al. (2019) found segmentation boundaries by aligning source text with target text made by human interpreters. Based on the analysis of segmentation boundaries, except for punctuation marks, coordinate conjunctions showed the highest rank followed by wh-words, adverbs, prepositions, and subordinate conjunctions.

Some researches focused on pause for segmentation (Kashioka, 2002; Bangalore et al., 2012). Bangalore et al. (2012) tried various lengths as a threshold of a meaningful pause and found that pauses over 100ms are meaningful for segmentation.

Such features mentioned above are derived from the aspect of a translation quality. Meanwhile, some studies focus on the translation latency (Cettolo and Federico, 2006; Rao et al., 2007; Sridhar et al., 2013b; Ma et al., 2019). Cettolo and Federico (2006) established segmentation boundaries every 10, 20, 30, 40, 50, 60, or 70 words and compared the translation quality of each approach. Ma et al. (2019) proposed to train a neural network model based on prefix-to-prefix and start translating source text from *k*-words behind ('*k*-wait'). This allowed the model to predict words at the sentence final position and translate with less latency. Ma et al. (2019) stated that '5-wait', approximately 3 seconds, results in the highest performance.

3 Segmentation for Simultaneous Interpretation

In this section, we propose linguistically motivated segmentation features and methods for a simultaneous interpretation system. In section 3.1, we introduce the linguistic features for segmentation. These features are taken into account to find out how suitable the point is to determine the segmentation boundary. In section 3.2, we suggest four segmentation methods. They differ in what they put stress on, when deciding segmentation boundaries.

3.1 Segmentation Features

We propose various linguistic features of segmentation for simultaneous interpretation: prosodic, POS, dependency, discourse, and cognitive information. In our method, the 'segmentation score' is calculated based on these features to decide the segmentation boundaries.

3.1.1 Prosodic Information

Prosodic information such as height and loudness of a sound can give clues to appropriate segmentation boundaries. However, to the best of our knowledge, there is not enough research on the impact of prosodic information on segmentation. Instead, many researches have been made on the impact of prosodic information on Transition Relevance Places (TRPs). TRP is a concept in Conversational Analysis. It denotes an end of Turn Construction Units (TCUs), unit of an utterance (Sacks et al., 1974). In human conversation, we can easily guess when the partner's utterance will end and when we can begin our turn. Ishimoto et al. (2011) investigated relation between prosodic information and TRPs in Japanese conversation.

Based on some similarities between simultaneous interpretation and conversation, Koo et al. (2019) applied the relation between prosodic information and TRPs to the relation between prosodic information and segmentation boundaries. Koo et al. (2019) analyzed pitch and power contours near segmentation boundaries. As a result, Koo et al. (2019) assumed that the fall of both pitch and power leads to segmentation.

In this sense, this paper sets the fall of pitch and power as one of the linguistic features for segmentation. Not only that, pauses in source text hint segmentation boundaries. This paper deals with two types of pause. Pauses marked as 'SENT_STR' by an automatic speech recognition system are relatively short and recognized as a start of a sentence by the system. Whereas pauses marked as 'SENT_END' are relatively long and recognized as an end of a sentence. We included these pauses as a linguistic feature of segmentation.

3.1.2 Part-of-speech (POS) Information

Syntactic structures take different forms depending on the grammatical features of each language. Some

POS information, of both English and Korean, characterizes the phrasal or clausal boundaries.

In English, a conjunction is used to connect two linguistic units (e.g. sentences, clauses). Therefore, we can effectively split two units by segmenting before conjunctions.

Korean is a SOV language and the end of each sentence is marked by the sentence-final ending. It is therefore the most obvious feature that can be used to make a segment between two sentences. While sentence-final ending marks the end position of a sentence, conjunctive ending connects two clauses. Since a clause is a syntactically complete unit, we can segment after conjunctive endings.

So, for most of the languages, POS information is a useful feature to find segmentation boundary.

3.1.3 Dependency Information

In simultaneous interpretation, due to speech situations such as pauses and lapses, a source text may be segmented at inappropriate segmentation points. To solve this problem, Koo et al. (2019) mentioned the need for semantic features that can prevent a semantically cohesive unit from being segmented into two separate units. As these units lose their original meaning when segmented, they should be maintained unsegmented.

In this study, we elaborate on this idea and suggest dependency features. The dependency features that we propose are summarized in Table 1.

Language	Feature name	Value
English	Adjective + Noun	JJ+N*
	Determiner + Noun	DT+N*
	Modal Auxiliary Verb + Verb	MD+VB*
	Auxiliary Verb + Verb	VB*+VB*
	Phrasal Verb(Verb + particle)	VB*+RP
	POS(not Noun) + Preposition	not N+IN
Korean	Adjectivalization Ending + Noun	ETJ+[N* or XPN]
	Adjective + Noun	D+[N* or XPN]
	Case Particle for Adjective + Noun	FM+[N* or XPN]
	Bound Noun	ND
	Auxiliary Verb	VX

Table 1. Dependency Features

3.1.4 Discourse Information

People try to be coherent when they are talking. This coherence is usually achieved by structuralizing the talk. Rhetorical Structure Theory (RST) is a theory that explains the structure of a text using a hierarchy between the sentences inside the text (Mann and Thompson, 1987). Texts in RST are hierarchic, built on partial texts which make a certain relation to each other. If two sentences have distinctive rhetorical characteristics, a linguistic marker appears between these two sentences to show such a transition. We call that Rhetorical Structure Markers (RSMs).

Therefore, RSM is an effective segmentation feature that can capture the general rhetorical relation of the text. In this study, we collected RSMs for each language: 160 for English, 140 for Korean. Table 2 is the example of RSMs.

Type	English	Korean
Addition	additionally, also, likewise	게다가(gedaga), 또한(ttohan), 유사하게(yusahage)
Contrast	although, conversely, in contrast	~에도 불구하고 (edo bulguhago), 반대로(bandaero), 대조적으로(daejojeogeuro)
Emphasis	in particular, specifically, without a doubt	특히(teukhi), 구체적으로(guchejeogeuro), 의심의 여지없이 (uisimui yeojieopsi)

Table 2. Examples of Rhetorical Structure Markers

3.1.5 Cognitive Information

While the features mentioned above are the features that guarantee the translation quality, the length of the segmentation unit is a feature for maintaining an appropriate translation latency. In order to do that, it is necessary to set an appropriate length of segmentation units. Based on the results of the previous studies and the analysis of the simultaneous interpretation data, the optimal length of the simultaneous interpretation unit was set to 4.5 seconds in this study.

First, the optimal length of the segmentation unit is based on the previous study in the field of simultaneous interpretation. Ear-Voice Span (EVS) refers to the time it takes for an interpreter to hear

the words spoken in the source language and then interpret them to the target language. In other words, EVS refers to the time it takes for a simultaneous interpreter to hear the source utterance and then obtain all the information needed to understand and interpret it.

Lederer (1978) showed that the average EVS occurring in the simultaneous interpretation of the English to French was measured between 3 and 6 seconds. Ono et al. (2008) analyzed the EVS in simultaneous interpretation between Japanese to English and English to Japanese. They found that the average EVS times were 4.532 seconds and 2.446 seconds, respectively. Also, according to Lee (2002), the English to Korean EVS averaged 3 seconds. Through the studies mentioned, it was found out that the segmentation unit for accurate translation was 3 seconds or more.

The length of the segmentation unit proposed in this study was set considering also the psychological state of the audience who listened to the interpretation in the target language. Sridhar et al. (2013a) found that the listeners feel psychologically tired when the lapse of more than 4 to 5 seconds occurs during the simultaneous interpretation. That is, in order to be a good simultaneous interpreter, the lapse does not occur for more than 5 seconds when interpreting the source text to target text.

The segmentation unit length of 4.5 seconds was also derived from the dataset, constructed by 'Electronics and Telecommunications Research Institute (ETRI)'. As a result of examining the 'SENT_END' tag in 19 English files and 10 Korean files, the source text length between 'SENT_END' and the next 'SENT_END' averages 4.55 seconds in English and 3.57 seconds in Korean. Refer to section 4.1 for more detailed information about the data.

3.2 Segmentation Methods

In this section, we propose four different segmentation methods using segmentation features, mentioned in section 3.1, to detect segmentation boundaries. We conducted experiments using these segmentation methods and compared the results in section 4.

3.2.1 Method 1 (Koo et al., 2019)

Segmentation method 1 was proposed in Koo et al. (2019). Method 1 segments a sentence when the 'priority feature of segmentation' appears. When it does not appear until 'optimal length of segmentation unit', then segments at the point, within 3.5~5.5 seconds, that has the highest segmentation score. Here, segmentation score is calculated by the sum of the values of segmentation feature. In detail, while the 'priority feature of segmentation' in English to Korean (En→Ko) simultaneous interpretation is RSMs, the 'priority feature of segmentation' in Korean to English (Ko→En) is RSMs and final endings.

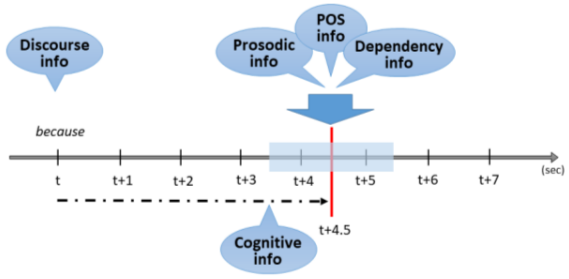


Figure 1. Flow of Segmentation Method 1

3.2.2 Method 2

Method 1 tends to segment only near 4.5 seconds, the range of 3.5~5.5 seconds, even if a better segmentation boundary is positioned right after that. Method 2 is designed to solve this limitation.

Like method 1, method 2 segments when 'priority feature of segmentation' appears. When it does not appear until 'optimal length of segmentation unit', then segmentation occurs at the point, after 4.5 seconds, where the segmentation score exceeds the threshold.

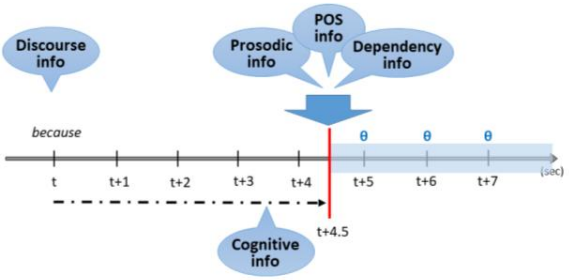


Figure 2. Flow of Segmentation Method 2

3.2.3 Method 3

Method 3 is similar to method 2, in that it segments when the segmentation score exceeds the threshold. However, method 3 gradually drops the threshold as time passes. This is inspired by human simultaneous interpreters. As time passes and the latency increases, they tend to accept less suitable points as segmentation boundaries, due to the pressure to give the audience a quick translation. Through this diminishing threshold, we expect less translation latency and guarantee translation quality.

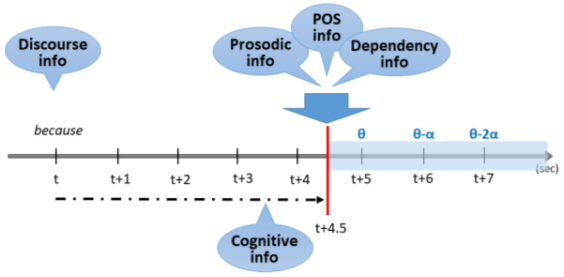


Figure 3. Flow of Segmentation Method 3

3.2.4 Method 4

Method 4 is similar to method 3, in that it considers both factors of simultaneous interpretation, translation quality and translation latency, when searching segmentation boundaries. However, method 4 directly utilizes latency as a variable for calculating the segmentation score.

To be specific, at the points before 4.5 seconds, method 4 focuses on guaranteeing only the translation quality and therefore uses linguistic features, mentioned in 3.1, for segmentation. On the other hand, at the points after 4.5 seconds, method 4 takes both the translation quality and translation latency into account for segmentation. Thus, methods 4 quantifies the latency and includes it as a feature for segmentation.

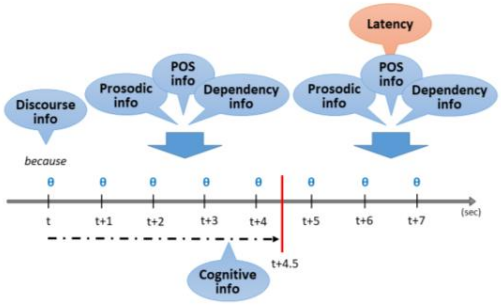


Figure 4. Flow of Segmentation Method 4

4 Experiments

In this section, we verify the linguistic features that we suggest and evaluate the best segmentation method for simultaneous interpretation systems. Section 4.1 explains about data used for the experiments. The evaluation results are shown in section 4.2. In section 4.3, we discuss and compare the results of the experiment.

4.1 Data

The experiment was conducted using the ETRI data, which are transcriptions of lectures. English data consist of 4 complete TED talks whose topics are artificial intelligence. The average length of videos is about 12 minutes long and the data include 6,711 tokens in total. Korean data are also composed of 4 lectures which are 'Sebasi' or 'K-MOOC' lectures with 5,193 tokens. Each video is about 12 minutes long on average.

As mentioned above, the transcription data include not only pause information but also POS tags. Additionally, we attached feature values for each token and determined whether the point should be segmented or not depending on each segmentation method.

4.2 Result

To evaluate the accuracy of the proposed methods, the acceptability of each segmentation point was calculated. As there is no absolutely correct answer for the segmentation points, only the acceptability of the points was taken into account. Table 3 shows the criteria that we set.

Grade	Criteria
Correct	The segmentation result contains all the syntactic and semantic pieces of the sentence, which are necessary for interpretation.
Acceptable	Some parts of information are missing, but still enough for interpretation.
Incorrect	Too much information is absent for interpretation.

Table 3. Criteria for Evaluation

Based on the criteria, three annotators who are native Korean speakers and possess a good command of English judged the appropriateness of segmentation points. Each annotator evaluated the

accuracy of segmentation points, thus three results of accuracy evaluation were derived. The agreement rate among three annotators is 76.6%. Then we took the average of these three as the final accuracy.

We analyzed two measurements: strict and loose accuracy. Strict accuracy only considers 'correct' segmentation points, while loose accuracy includes 'correct' and 'acceptable' points.

		Methods			
		1	**2**	**3**	**4**
Accuracy (%)	**strict**	68.6	70.5	78.3	**80**
	loose	76.3	80.7	88.7	**88.4**
Duration(sec)		4.6	7.5	7	**3.6**

Table 4. Evaluation of Segmented Units (English)

As Table 4 shows the results of evaluating segmented units, which are split depending on each method. We designated 0.33 as a threshold for English data and 0.01 as time weight. The threshold was calculated from the average of the feature values of each segmentation point in other data. These segmentation points are marked by professional human translators.

Though method 3 in Table 4 shows the highest accuracy, its average duration takes about 7 seconds per each segmented unit. However, method 4 represents slightly lower but relatively similar accuracy to method 3. Also, the average duration of segmented units of method 4 is about 3.6 seconds, which is the lowest latency. Considering the trade-off between accuracy and latency, it implies that method 4 is the most proper method for English simultaneous interpretation.

Table 5 compares an original text with texts segmented by using the method 4. Compared to the original text, the segmented text shows that the text is properly segmented without hurting the original meaning and showing lower latency. As for the first segment, pause information played the crucial role in segmentation. The second segmentation was geared by both pause and POS information. The third and last segmentation were caused by RSM, prosodic and pause features.

	Segmented Unit	**Time**
Original Text	In such a brutal environment entrepreneurs learned to grow very rapidly they learned to make their products better at lightning speed and they learned to hone their business models until they're impregnable.	14.52
Segmented Text (Method 4)	in such a brutal environment entrepreneurs learned to grow very rapidly	6.14
	they learned to make their products better at lightning speed	3.94
	and they learned to hone their business models	2.41
	until they're impregnable	2.01

Table 5. Examples of Segmented Units – Original Text vs. Method 4

Table 6 shows the comparison of the results of each segmentation method for Korean. We specified 0.26 as a threshold for Korean data and 0.01 as time weight. The threshold was assigned in the same way as for English data.

		Methods			
		1	**2**	**3**	**4**
Accuracy (%)	**strict**	76	85.6	**88.2**	80.1
	loose	78.9	88.4	**92.6**	87.9
Duration(sec)		3.7	5.5	**5.1**	3.4

Table 6. Evaluation of Segmented Units (Korean)

The evaluation results indicate that the method 3 seems to be the most powerful method to segment Korean data with the topmost loose/strict accuracy with lower latency. In contrast to English data, method 4 results in noticeably lower accuracy compared to the method 3.

Compared to Koo et al. (2019) we added and elaborated dependency features. Table 7 shows the effects of them that induce better segmentation points. With this example, we can confirm that segmentation between adjective and noun is prevented by dependency features.

	Segmented Unit	Time
Without dependency features	파이썬 이라는 단어는요 원래 저기 뱀 이 큰 (The word Python is actually a huge) (paisseon iraneun daneoneun-yo wonrae jeogi baem i keun)	4.78
	보아뱀이라고 하나요 (snake so-called Boa) (boabaemirago hanayo)	0.90
With dependency features	파이썬 이라는 단어는요 원래 저기 뱀 이 큰 보아뱀이라고 하나요 (The word 'Python' is actually a huge snake so-called Boa) (paisseon iraneun daneoneunyo wonrae jeogi baem i keun boabaemirago hanayo)	5.68

Table 7. Examples of Segmented Units – Effect of Dependency features

4.3 Discussion

As mentioned in Koo et al. (2019), method 1 tends to segment only near 4.5 seconds even if a better segmentation boundary is positioned right after that. Koo et al. (2019) expected that segmentation accuracy will increase if the system waits a little longer for a better segmentation boundary. As expected, method 2 showed better segmentation accuracy. Along with that, however, the average length of the segmented unit increased. This implies that method 2 caused more translation latency. Table 8 compares the segmentation result of method 1 and 2.

	Segmented Unit	Time
Method 1	now imagine an AI is helping a hiring manager find the next tech leader	5.15
	in the company	1.24
Method 2	now imagine an AI is helping a hiring manager find the next tech leader in the company	6.39

Table 8. Examples of Segmented Units – Method 1 vs. Method 2

We intended method 3 to alleviate translation latency by gradually dropping the segmentation threshold. As mentioned earlier, we expected less translation latency and guaranteed translation quality, when using segmentation method 3. As a result, it did keep high translation accuracy, but could not fully solve translation latency occurred in method 2. Refer to Table 9 for segmentation accuracy and an average length of segmented unit.

	Segmented Unit	Time
Method 2	Think about a pregnant woman in the Democratic Republic of Congo who has to walk seventeen hours to her nearest rural prenatal clinic to get a checkup what if she could get diagnosis on her phone instead	14.30
Method 3	Think about a pregnant woman in the Democratic Republic of Congo who has to walk seventeen hours to her nearest rural prenatal clinic to get a checkup	9.90

Table 9. Examples of Segmented Units – Method 2 vs. Method 3

The most critical problem of method 3 is that the segmented units are relatively long, which raises translation latency. To overcome this problem, method 4 brings in translation latency as a feature for segmentation. In addition, unlike method 3, method 4 checks every point whether it is an appropriate segmentation boundary. Consequently, method 4 resulted in a shorter average length of segmented unit, while maintaining high translation accuracy. Table 10 and 11 illustrates detailed information about the length of the segmented unit per methods for En→Ko and Ko→En.

Duration	Methods			
	1	2	3	4
Average	4.58	7.50	7.04	3.60
Minimum	0.15	0.10	0.10	0.10
Maximum	20.44	33.82	20.44	12.07

Table 10. Length of Segmented Units (English)

Duration	Methods			
	1	2	3	4
Average	3.73	5.35	5.10	3.36
Minimum	0.25	0.18	0.18	0.14
Maximum	9.03	21.21	14.55	8.95

Table 11. Length of Segmented Units (Korean)

Up to now, we looked through the segmentation results of each of four segmentation methods. We saw that each of them has different strengths and weaknesses. But not only that, they showed a different segmentation performance, depending on the source language.

According to the Table 10 and 11, regardless of the source language, segmentation accuracy is higher and average length of segmented unit is shorter, in the order of method 1, 2, and 3. Nevertheless, when comparing method 3 and 4, the result differs with regard to the source language. When segmenting English source text, method 3 and 4 led to similar segmentation accuracy, while method 4 produced considerably shorter segmented units. This indicates that method 4 can perform better for English when considering the trade-off between translation quality and translation latency.

Segmentation for Korean source text shows different aspects from that of English source text. When segmenting Korean source text, method 4 produced shorter segmented units, which implies less translation latency. However, method 4 caused relatively great decrease in segmentation accuracy when it was applied to Korean. This means that segmentation method 3 seems to work well for Korean source text.

This can be attributed to the typological difference between English and Korean. English is a head-initial language, so that a verb is located mostly in the front of a sentence. On the other hand, Korean is a head-final language and its verb appears in the back of a sentence. Since the latency is used as a feature for segmentation, method 4 results in more frequent segmentations after 4.5 seconds, the optimal length of segmentation unit. In this regard, when method 4 is applied to the Korean source text, it is likely that segmentation boundary occurs in between the verb phrase and leads to inappropriate segmentation. Therefore, different segmentation methods should be applied depending on the source language.

5 Conclusion and Future Works

In this paper, we proposed linguistically motivated segmentation features and methods to investigate segmentation units for simultaneous interpretation. Various features such as prosodic, POS, dependency, discourse and cognitive information were set for proper segmentation. Also, to prevent the length of the segment unit from being excessively long, we considered latency as a feature. Based on these features, four segmentation methods were proposed. The highest accuracy was achieved in method 4 (80%) for En→Ko and method 3 (88.2%) for Ko→En.

In the future study, the method of evaluating the segmented units should be further revised. In this study, when evaluating the segmented units, we judged only whether information in the segmented unit is sufficient to interpret. However, if the segmented unit contains other segmentation points inside itself, which should have been segmented, this unit should be penalized in the future.

Furthermore, we will check whether the interpretation result which is assisted by segmentation shows a significant performance difference compared to the interpretation result without segmentation. To this end, we plan to develop a suitable evaluation method for simultaneous interpretation that takes into account the differences between machine translation and simultaneous interpretation.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement)

References

Akiko Nakabayashi and Tsuneaki Kato. 2019. Simulating Segmentation by Simultaneous Interpreters for Simultaneous Machine Translation. In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC), 165-173.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), 2: 1005-1008.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-taking for Conversation, Language, 50: 696–735.

Hideki Kashioka. 2002. Translation Unit Concerning Timing of Simultaneous Translation. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), 142-146.

Marianne Lederer. 1978. Simultaneous Interpretation – Units of meaning and Other Features. Language interpretation and communication. Springer, 323-332.

Mauro Cettolo and Marcello Federico. 2006. Text segmentation criteria for statistical machine translation. In Proceedings of the International Conference on Natural Language Processing, 664-673.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 3025-3036.

Sharath Rao, Ian Lane, and Tanja Schultz. 2007. Optimizing Sentence Segmentation for Speech Translation. In Proceedings of Interspeech2007, 2845–2848.

Srinivas Bangalore, Vivek K. R. Sridhar, Prakash Kolan, LadanGolipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 437-445.

Tae-Hyung Lee. 2002. Ear Voice Span in English into Korean Simultaneous Interpretation. Meta, 47(4):596-606.

Takahiro Ono, HitomiTohyama, and Shigeki Matsubara. 2008. Construction and Analysis of Word-level Time-aligned Simultaneous Interpretation Corpus. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 3383-3387.

Vivek K. R. Sridhar, John Chen, and Srinivas Bangalore. 2013a. Corpus analysis of simultaneous interpretation data for improving real time speech translation. INTERSPEECH, 3468-3472.

Vivek K. R. Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013b. Segmentation strategies for streaming speech translation. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 230-238.

William C. Mann, and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243-281.

Youngeun Koo, Jiyoun Kim, Jungpyo Hong, Munpyo Hong and Sung-Kwon Choi. 2019. A Study on Segmentation Unit for the Real-time Simultaneous Interpretation System. In Proceedings of the 31st Annual Conference on Human & Cognitive Language Technology (HCLT), 229-235.

Yuichi Ishimoto, Mika Enomoto, and Hitoshi Iida. 2011. Projectability of Transition-Relevance Places Using Prosodic Features in Japanese Spontaneous Conversation. In Proceedings of Interspeech2011, 2061–2064.

Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation

Fajri Koto
School of Computing and Information System
The University of Melbourne
ffajri@student.unimelb.edu.au

Ikhwan Koto
Faculty of IT
Andalas University
ikhwan220397@gmail.com

Abstract

Although some linguists (Rusmali et al., 1985; Crouch, 2009) have fairly attempted to define the morphology and syntax of Minangkabau, information processing in this language is still absent due to the scarcity of the annotated resource. In this work, we release two Minangkabau corpora: sentiment analysis and machine translation that are harvested and constructed from Twitter and Wikipedia.[1] We conduct the first computational linguistics in Minangkabau language employing classic machine learning and sequence-to-sequence models such as LSTM and Transformer. Our first experiments show that the classification performance over Minangkabau text significantly drops when tested with the model trained in Indonesian. Whereas, in the machine translation experiment, a simple word-to-word translation using a bilingual dictionary outperforms LSTM and Transformer model in terms of BLEU score.

1 Introduction

Minangkabau (Baso Minang) is an Austronesian language with roughly 7m speakers in the world (Gordon, 2005). The language is spread under the umbrella of Minangkabau tribe – a matrilineal culture in the province of West Sumatra, Indonesia. The first-language speakers of Minangkabau are scattered across Indonesian archipelago and Negeri Sembilan, Malaysia due to "*merantau*" (migration) culture of Minangkabau tribe (Drakard, 1999).

Despite there being over 7m first-language speakers of Minangkabau,[2] this language is rarely used in the formal sectors such as government and education. This is because the notion to use Bahasa Indonesia as the unity language since the Independent day of Indonesia in 1945 has been a double-edged sword. Today, Bahasa Indonesia successfully connects all ethnicities across provinces in Indonesia (Cohn et al., 2014), yet threatens the existent of some indigenous languages as the native speakers have been gradually decreasing (Novitasari et al., 2020). Cohn et al. (2014) predicted that Indonesia may shift into a monolingual society in the future.

In this paper, we initiate the preservation and the first information processing of Minangkabau language by constructing a Minangkabau–Indonesian parallel corpus, sourced from Twitter and Wikipedia. Unlike other indigenous languages such as Javanese and Sundanese that have been discussed in machine speech chain (Novitasari et al., 2020; Wibawa et al., 2018), part-of-speech (Pratama et al., 2020), and translation system (Suryani et al., 2016), information processing in Minangkabau language is less studied. To the best of our knowledge, this is the first research on NLP in Minangkabau language, which we conduct in two different representative NLP tasks: sentiment analysis (classification) and machine translation (generation).

There are two underlying reasons why we limit our work in Minangkabau–Indonesian language pair. First, Minangkabau and Indonesian language

[1] Our data can be accessed at https://github.com/fajri91/minangNLP

[2] In Indonesia, Minangkabau language is the fifth most spoken indigenous language after Javanese (75m), Sundanese (27m), Malay (20m), and Madurese (14m) (Riza, 2008).

are generally intelligible with some overlaps of lexicons and syntax. The Indonesian language has been extensively studied and is arguably a convenient proxy to learn the Minangkabau language. Second, authors of this work are the first-language speakers of Minangkabau and Indonesian language. This arguably eases and solidifies the research validation in both tasks.

To summarize, our contributions are: (1) we create a bilingual dictionary from Minangkabau Wikipedia by manually translating top 20,000 words into Indonesian; 2) we release Minangkabau corpus for sentiment analysis by manually translating 5,000 sentences of Indonesian sentiment analysis corpora; 3) We develop benchmark models with classic machine learning and pre-trained language model for Minangkabau sentiment analysis; 4) We automatically create a high-quality machine translation corpus consisting 16K Minangkabau–Indonesian parallel sentences; and 5) We showcase the first Minangkabau–Indonesian translation system through LSTM and Transformer model.

2 Minangkabau–Indonesian Bilingual Dictionary

In the province of West Sumatra, Minangkabau language is mostly used in spoken communication, while almost all reading materials such as local newspaper and books are written in Indonesian. Interestingly, Minangkabau language is frequently used in social media such as Twitter, Facebook and WhatsApp, although the writing can be varied and depends on the speaker dialect. Rusmali et al. (1985) define 6 Minangkabau dialects based on cities/regencies in the West Sumatra province. This includes Agam, Lima Puluh Kota, Pariaman, Tanah Datar, Pesisir Selatan, and Solok. The variation among these dialects is mostly phonetic and rarely syntactic.

Crouch (2009) classifies Minangkabau language into two types: 1) Standard Minangkabau and 2) Colloquial Minangkabau. The first type is the standard form for intergroup communication in the province of West Sumatra, while the second is the dialectal variation and used in informal and familiar contexts. Moussay (1998) and Crouch (2009) argue that Padang dialect is the standard form of Minangkabau. However, as the first-language speaker, we contend that these statements are inaccurate because of two reasons. First, many locals do not aware of Padang dialect. We randomly survey 28 local people and only half of them know the existence of Padang dialect. Second, in 2015 there has been an attempt to standardize Minangkabau language by local linguists, and Agam-Tanah Datar is proposed as the standard form due to its largest population.[3]

Our first attempt in this work is to create a publicly available Minangkabau–Indonesian dictionary by utilising Wikipedia. Minangkabau Wikipedia[4] has 224,180 articles (rank 43rd) and contains 121,923 unique words, written in different dialects. We select top-20,000 words and manually translate it into Indonesian. We found that this collection contains many noises (e.g. scientific terms, such as *onthophagus*, *molophilus*) that are not Minangkabau nor Indonesian language. After manually translating the Minangkabau words, we use *Kamus Besar Bahasa Indonesia* (KBBI)[5] – the official dictionary of Indonesian language to discard the word pairs with the unregistered Indonesian translation. We finally obtain 11,905-size Minangkabau–Indonesian bilingual dictionary, that is 25 times larger than word collection in Glosbe (476 words).[6]

We found that 6,541 (54.9%) Minangkabau words in the bilingual dictionary are the same with the translation. As both Minangkabau and Indonesian languages are Austronesian (*Malayic*) language, the high ratio of lexicon overlap is very likely. Further, we observe that 1,762 Indonesian words have some Minangkabau translations. These are primarily synonyms and dialectal variation that we show in Table 1. Next, in this study, we use this dictionary in sentiment analysis and machine translation.

3 Sentiment Analysis

Sentiment analysis has been extensively studied in English and Indonesia in different domains such as movie review (Yessenov and Misailovic, 2009; Nurdiansyah et al., 2018), Twitter (Agarwal et al., 2011;

[3] https://id.wikimedia.org/wiki/Sarasehan_Bahasa_Minangkabau

[4] Downloaded in June 2020

[5] https://github.com/geovedi/indonesian-wordlist

[6] https://glosbe.com/min/id

Indonesian	English	Minangkabau
	Synonyms	
ibunya	her mother	*ibunyo, mandehnyo, amaknyo*
memperlihatkan	to show	*mampacaliak, mampaliekan*
kelapa	coconut	*karambia, kalapo*
	Dialectal variations	
berupa	such as	*barupo, berupo, berupa, barupa*
bersifat	is, act, to have the quality	*basipaik, basifaik, basifek, basifat, basipek*
Belanda	Netherlands	*Balando, Belanda, Bulando, Belando*

Table 1: Example of synonyms and dialectal variations in the Minangkabau–Indonesian dictionary

Koto and Adriani, 2015), and presidential election (Wang et al., 2012; Ibrahim et al., 2015). It covers a wide range of approaches, from classic machine learning such as naive Bayes (Nurdiansyah et al., 2018), SVM (Koto and Adriani, 2015) to pre-trained language models (Sun et al., 2019; Xu et al., 2019). The task is not only limited to binary classification of positive and negative polarity, but also multi classification (Liu and Chen, 2015), subjectivity classification (Liu, 2010), and aspect-based sentiment (Ma et al., 2017).

In this work, we conduct a binary sentiment classification on positive and negative sentences by first manually translating Indonesian sentiment analysis corpus to Minangkabau language (Agam-Tanah Datar dialect). To provide a comprehensive preliminary study, we experimented with a wide range of techniques, starting from classic machine learning algorithms, recurrent models, to the state of the art technique, Transformer (Vaswani et al., 2017).

3.1 Dataset

The data we use in this work is sourced from 1) Koto and Rahmaningtyas (2017); and 2) an aspect-based sentiment corpus.[7] Koto and Rahmaningtyas (2017) dataset is originally from Indonesian tweets and has been labelled with positive and negative class. The second dataset is a hotel review collection where each review can encompass multi-polarity on different aspects. We determine the sentiment class based on the majority count of the sentiment label, and simply discard it if there is a tie between positive and negative. In total, we obtain 5,000 Indonesian texts from these two sources. We then ask two native speakers of Minangkabau and Indonesian language to manually translate all texts in the corpus. Finally, we create a parallel sentiment analysis corpus with 1,481 positive and 3,519 negative labels.

3.2 Experimental Setup

We conducted two types of the zero-shot experiment by using Indonesian train and development sets. In the first experiment, the model is tested against Minangkabau data, while in the second experiment we test the same model against the Indonesian translation, obtained by word-to-word translation using the bilingual dictionary (Section 2). There are two underlying reasons to perform the zero-shot learning: 1) Minangkabau is intelligible with Indonesian language and most available corpus in the West Sumatra is Indonesian; 2) Minangkabau language is often mixed in Indonesian data collection especially in social media (e.g. Twitter, if the collection is based on geographical filter). Through zero-shot learning, we aim to measure the performance drop of Indonesian model when tested against the indigenous language like Minangkabau.

Our experiments in this section are based on 5-folds cross-validation. We conduct stratified sampling with ratio 70/10/20 for train, development, and test respectively, and utilize five different algorithms as shown in Table 2. For naive Bayes, SVM and logistic regression, we use byte-pair encoding (unigram and bigram) during the training and tune the model based on the development set. Due to data imbalance, we report the averaged F-1 score of five test sets.

For Bi-LSTM (200-d hidden size) we use two

[7] https://github.com/annisanurulazhar/absa-playground/

Method	Train $_{ID}$		Train $_{MIN}$
	Test $_{MIN}$	Test $_{ID'}$	Test $_{MIN}$
Naive Bayes	68.49	68.86	73.03
SVM	59.75	68.35	74.05
Logistic Regression	57.95	66.90	72.35
Bi-LSTM	58.75	65.62	72.37
Bi-LSTM + `fastText`	62.06	71.51	70.47
MBERT	62.71	67.60	**75.91**

Table 2: Results for Sentiment Analysis on Minangkabau test set. The numbers are the averaged F-1 of 5-folds cross validation sets. MIN = Minangkabau, ID = Indonesian, ID' = Indonesian translation through bilingual dictionary.

variants of 300-d word embedding: 1) random initialization; and 2) `fastText` pre-trained Indonesian embeddings (Bojanowski et al., 2016). First, we lowercase all characters and truncate them by 150 maximum words. We use batch size 100, and concatenate the last hidden states of Bi-LSTM for classification layer. For each fold, we train and tune the model for 100 steps with Adam optimizer and early stopping (patience = 20).

Lastly, we incorporate the Transformer-based language model BERT (Devlin et al., 2019) in our experiment. Multilingual BERT (MBERT) is a masked language model trained by concatenating 104 languages in Wikipedia, including Minangkabau. MBERT has been shown to be effective for zero-shot cross-lingual tasks including classification (Wu and Dredze, 2019). In this work, we show the first utility of MBERT for classifying text in the indigenous language, such as Minangkabau. In fine-tuning, we truncate all data by 200 maximum tokens, and use batch size 30 and maximum epoch 20 (2,500 steps). The initial learning rate is 5e-5 with warm-up of 10% of the total steps. We evaluate F-1 score of the development set for every epoch, and terminate the training if the performance does not increase within 5 epochs. Similar to Bi-LSTM models, we use Adam optimizer for gradient descent steps.

3.3 Result

In Table 2, we show three different experimental results. The first column is the zero-shot setting where the model is trained and tuned using Indonesian text and tested against Minangkabau data. Surprisingly, naive Bayes outperforms other models including MBERT with a wide margin. Naive Bayes achieves 68.49 F1-score, +6 points over the pre-trained language model and Bi-LSTM + `fastText`. This might indicate that naive Bayes can effectively exploit the vocabulary overlap between Minangkabau and Indonesian language.

In the second experiment, we hypothesize that a simple word-to-word translation using a bilingual dictionary can improve zero-shot learning. Similar to the first experiment, we train the model with Indonesian text, but we test the model against the Indonesian translation. As expected, the F-1 scores improve dramatically for all methods except naive Bayes with +0.37 gains. SVM, logistic regression and Bi-LSTMs are improved by 6–9 points while MBERT gains by +5 points by predicting the Indonesian translation.

In the third experiment, we again show a dramatic improvement when the model is fully trained in the Minangkabau language. Compared to the second experiment, all models are improved by 4–8 points with Bi-LSTM + `fastText` in exception. This is because the model uses `fastText` pre-trained Indonesian embeddings, and its best utility is when the model is trained and tested in the Indonesian language (second experiment). The best model is achieved by MBERT with 75.91 F1-score, outperforming other models with a comfortable margin.

Based on these experiments, we can conclude the necessity of specific indigenous language resource for text classification in Indonesia. These languages are mixed in Indonesian social media, and testing the Indonesian model directly on this Indonesian-type language can drop the sentiment classification performance by 11.41 on average.

3.4 Error Analysis

In this section, we manually analyze the false positive (FN) and false negative (FP) of MBERT model. We examine all misclassified instances in the test set by considering three factors:

- *Bias towards a certain topic*. In Indonesia, we argue that public sentiment towards government, politics and some celebrities are often negative. This could lead to bias in the train-

Category	Value
#FN	83
Bias towards certain topic (%)	34.84
Single polarity with negative words (%)	20.48
Mixed polarity (%)	12.05
#FP	56
Bias towards certain topic (%)	26.79
Single polarity with positive words (%)	26.79
Mixed polarity (%)	28.57

Table 3: Error analysis for False Negative (FN) and False Positive (FP) set.

ing and result in a wrong prediction in the test set. We count the number of texts in FP and FN set that contain these two topics: politics and celebrity.

- *Single polarity but containing words in opposite polarity*. The model might fail to correctly predict a sentiment label when contains words with the opposite polarity.
- *Mixed polarity*. A text can consist of both positive and negative polarity with one of them is more dominant.

In Table 3 we found that there are 83 FN (28% of positive data) and 56 FP (8% of negative data) instances. We further observe that 34.84% of FN instances contain politics or celebrity topic, while there is only 26.79% of FP instances with these criteria. In Figure 1, we show an FN example for the first factor: "*Iduik Golkar! Idrus jo Yorrys Bapaluak*" where "*Golkar*" is one of the political parties in Indonesia.

Secondly, 20.48% of FN instances contain negative words. As shown in Figure 1 the example uses words "give up" and "hurt" to convey positive advice. We notice that the second factor is more frequent in FP instances with 26.79% proportion. Lastly, we find that 28.57% of FP instances have mixed polarity, 2 times larger than FN. We observe that most samples with mixed polarity are sourced from the hotel review. It highlights that mixed polarity is arguably a harder task, and requires special attention to aspects in text fragments.

4 Machine Translation

Machine translation has been long run research, started by Rule-based Machine Translation (RBMT) (Carbonell et al., 1978; Nagao, 1984), Statistical Machine Translation (SMT) (Brown et al., 1990), to Neural Machine Translation (NMT) (Bahdanau et al., 2015). NMT with its continuous vector representation has been a breakthrough in machine translation, minimizing the complexity of SMT yet boosts the BLEU score (Papineni et al., 2002) into a new level. Recently, Hassan et al. (2018) announce that their Chinese–English NMT system has achieved a comparable result with human performance.

Although there are 2,300 languages across Asia, only some Asian languages such as Chinese and Japanese have been extensively studied for machine translation. We argue there are two root causes: a lack of parallel corpus, and a lack of resource standardization. Apart from the Chinese language, there have been some attempts to create a parallel corpus across Asian languages. Nomoto et al. (2019) construct 1,3k parallel sentences for Japanese, Burmese, Malay, Indonesian, Thai, and Vietnamese, while Kunchukuttan et al. (2017) release a large-scale Hindi-English corpus. Unlike these national languages, machine translation on indigenous languages is still very rare due to data unavailability. In Indonesia, Sundanese (Suryani et al., 2015) and Javanese (Wibawa et al., 2013) have been explored through statistical machine translation. In this work, our focus is Minangkabau–Indonesian language pair, and we first construct the translation corpus from Wikipedia.

4.1 Dataset

Constructing parallel corpus through sentence alignment from bilingual sources such as news (Zhao and Vogel, 2002; Rauf and Schwenk, 2011), patent (Utiyama and Isahara, 2003; Lu et al., 2010), and Wikipedia (Yasuda and Sumita, 2008; Smith et al., 2010; Chu et al., 2014) have been done in various language pairs. For Indonesian indigenous language, Trisedya and Inastra (2014) has attempted to create parallel Javanese–Indonesian corpus by utilizing inter-Wiki links and aligning sentences via Gale and Church (1993) approach.

In this work, we create Minangkabau–Indonesian

Minangkabau	English
Bias towards certain topic: Iduik Golkar! Idrus jo Yorrys Bapaluak	**Bias towards certain topic:** Glory for Golkar! Idrus and Yorrys are hugging
Positive text that uses negative words: Ado masonyo dima awak harus marelaan nan awak sayangi untuak pai.. Yo mamang sakik! Tapi tu demi kebaikan awak surang. :)	**Positive text that uses negative words:** there are times when we have to give up on someone we care about. Yes indeed hurt! but this is for our good :)
Mixed polarity: kamarnyo rancak, patamu kali pakai airy kironyo dapek toolkit dan snack lo, cuma wifi hotelnyo indak bisa connect.. overall rancak, apolai diskon 80%	**Mixed polarity:** the room was good, the first time I used Airy, it turns out I got snacks and toolkits, it's just that hotel wifi was not functioning. Overall is good, especially 80% discounts.

Figure 1: Example of False Negative.

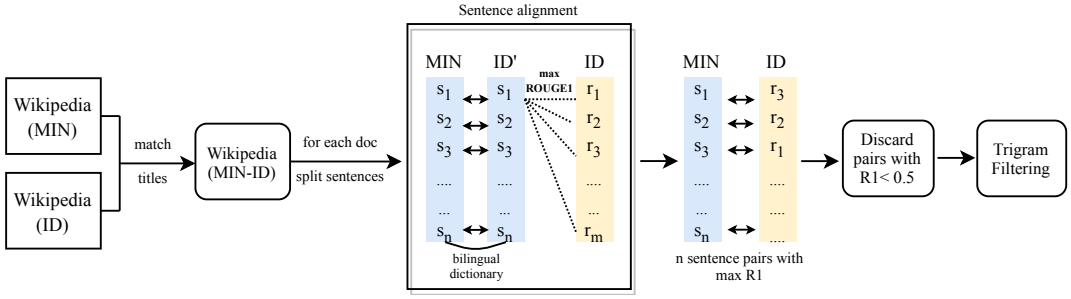


Figure 2: Flow chart of MIN-ID parallel corpus construction.

(MIN-ID) parallel corpus by using Wikipedia[8] (Figure 2). We obtain 224,180 Minangkabau and 510,258 Indonesian articles, and align documents through title matching, resulting in 111,430 MIN-ID document pairs. After that, we do sentence segmentation based on simple punctuation heuristics and obtain 4,323,315 Minangkabau sentences. We then use the bilingual dictionary (Section 2) to translate Minangkabau article (MIN) into Indonesian language (ID'). Sentence alignment is conducted using ROUGE-1 (F1) score (unigram overlap) (Lin, 2004) between ID' and ID, and we pair each MIN sentence with an ID sentence based on the highest ROUGE-1. We then discard sentence pairs with a score of less than 0.5 to result in 345,146 MIN-ID parallel sentences.

We observe that the sentence pattern in the collection is highly repetitive (e.g. 100k sentences are about biological term definition). Therefore, we conduct final filtering based on top-1000 trigram by iteratively discarding sentences until the frequency of each trigram equals to 100. Finally, we obtain 16,371 MIN-ID parallel sentences and conducted manual evaluation by asking two native Minangkabau speakers to assess the adequacy and fluency (Koehn and Monz, 2006). The human judgement is based on scale 1–5 (1 means poor quality and 5 otherwise) and conducted against 100 random samples. We average the weights of two annotators before computing the overall score, and we achieve 4.98 and 4.87 for adequacy and fluency respectively.[9] This indicates that the resulting corpus is high-quality for machine translation training.

[8]Downloaded in June 2020

Category	Wiki		SentC	
	MIN	ID	MIN	ID
mean(#word)	19.6	19.6	22.3	22.2
std(#word)	11.6	11.5	12.8	12.7
mean(#char)	105.2	107.7	98.9	99.2
std(#char)	59.1	60.4	57.9	58.1
#vocab	32,420	27,318	13,940	13,698
Overlapping #vocab	21,563		9,508	

Table 4: Statistics of machine translation corpora.

4.2 Experimental Setup

First, we split Wikipedia data with ratio 70/10/20, resulting in 11,571/1,600/3,200 data for train, de-

[9]The Pearson correlation of two annotators for adequacy and fluency are 0.9433 and 0.5812 respectively

velopment, and test respectively. In addition, we use parallel sentiment analysis corpus (Section 3) as the second test set (size 5,000) for evaluating texts from different domain. In Table 4, we provide the overall statistics of both corpora: Wikipedia (Wiki) and Sentiment Corpus (SentC). We observe that Minangkabau (MIN) and Indonesian (ID) language generally have similar word and char lengths. The difference is in the vocabulary size where Minangkabau is 5k larger than Indonesian in Wiki corpus. As we discuss in Section 2, this difference is due to various Minangkabau dialects in Wikipedia.

We conducted two experiments: 1) Minangkabau to Indonesian (MIN → ID); and 2) Indonesian to Minangkabau (ID → MIN) with three models: 1) word-to-word translation (W2W) using bilingual dictionary (Section 2); 2) LSTMs; and 3) Transformer. We use Moses Tokeniser[10] for tokenization, and sacreBLEU script (Post, 2018) to evaluate BLEU score on the test sets. All source and target sentences are truncated by 75 maximum lengths.

Our encoder-decoder (LSTM and Transformer) models are based on Open-NMT implementation (Klein et al., 2017). For LSTM models, we use two layers of 200-d Bi-LSTM encoder and 200-d LSTM decoder with a global attention mechanism. Source and target embeddings are 500-d and shared between encoder and decoder. For training, we set the learning rate of 0.001 with Adam optimizer, and warm-up of 10% of the total steps. We train the model with batch size 64 for 50,000 steps and evaluate the development set for every 5,000 steps.

The Transformer encoder-decoder (each) has 6 hidden layers, 512 dimensionality, 8 attention heads, and 2,028 feed-forward dimensionalities. Similar to LSTM model, the word embeddings are shared between source and target text. We use cosine positional embedding and train the model with batch size 5,000 for 50,000 steps with Adam optimizer (warm-up = 5,000 and Noam decay scheme). We evaluate the development set for every 10,000 steps.

4.3 Result

In Table 5, we present the experiment results for machine translation. Because Indonesian and Minangkabau language is mutually intelligible, and Table 4 shows that roughly 75% words in two vocabularies overlap, we set the BLEU scores of raw source and target text as the baseline. We found for both MIN→ID and ID→MIN, the BLEU scores are relatively high, more than 30 points.

Method	MIN → ID		ID → MIN	
	Wiki	SentC	Wiki	SentC
Raw (baseline)	30.08	43.73	30.08	43.73
W2W	**64.54**	**60.99**	**55.08**	**55.22**
LSTM	63.77	22.82	48.50	15.52
Transformer	56.25	10.23	43.50	8.86

Table 5: BLEU score on the test set. SentC is parallel sentiment analysis corpus in Section 3.

We observe that a simple word-to-word (W2W) translation using a bilingual MIN-ID dictionary achieves the best performance over LSTM and Transformer model in all cases. For MIN→ID, the BLEU scores are 64.54 and 60.99 for Wiki and SentC respectively, improving the baseline roughly 20–30 points. The similar result is also found in ID→MIN with disparity 12–25 points in the baseline.

Both LSTM and Transformer models significantly improve the baseline for Wiki corpus, but poorly perform in translating SentC dataset. For the Wiki corpus, the LSTM model achieves a competitive score in MIN→ID and ID→MIN, improving the baseline for 33 and 18 points respectively. The Transformer also outperforms the baselines, but substantially lower than the LSTM. In out-of-domain test set (SentC), the performance of both models significantly drops, 20–30 points lower than the baselines. We further observe that this is primarily due to out of vocabulary issue, where around 65% words in SentC are not in the vocabulary model.

4.4 Analysis

In Figure 3, we show translation examples in Wiki corpus. In MIN→ID, the LSTM translation is slightly more eloquent than word-to-word (W2W) translation. Word "*tamasuak*" (including) in W2W translation is Minangkabau language and not properly translated. This is because the word "*tama-*

[10] https://pypi.org/project/mosestokenizer/

Wiki corpus Example (MIN-ID)	Wiki corpus Example (ID-MIN)
Source (MIN): saketek nan dikatahui tantang kahidupan awal ching shih, tamasuak namo lahia jo tanggal lahirnya (*There is little information about Ching Shih early life, including her birth name and birth date*) **Target (ID):** **Reference:** sedikit yang diketahui tentang kehidupan awal ching shih, termasuk nama lahir dan tanggal lahirnya **W2W with Bilingual Dictionary:** sedikit yang diketahui tantang kehidupan awal ching shih, tamasuak nama lahir dengan tanggal lahirnya **LSTM:** sedikit yang diketahui tentang kehidupan awal ching shih, termasuk nama lahir dan tanggal lahirnya **Transformer:** sedikit yang diketahui tentang kehidupan awal rambut, termasuk nama lahir dan tanggal kelahiran	**Source (ID):** laba-laba ini biasanya banyak ditemui di amerika serikat, guatemala, antigua (*This spider is mostly found in the United States, Guatemala, Antigua*) **Target (MIN):** **Reference:** lawah iko biasonyo banyak ditamui di amerika sarikat, guatemala, antigua **W2W with Bilingual Dictionary:** laba-laba ko biasonyo banyak ditemui di amerika serikat, guatemala, antigua **LSTM:** lawah iko biasonyo banyak ditamui di amerika serikat, guatemala, moldavia **Transformer:** lawah iko biasonyo banyak ditamui di amerika sarikat, guatemala, alaska

Figure 3: Examples of model translation in Wikipedia corpus.

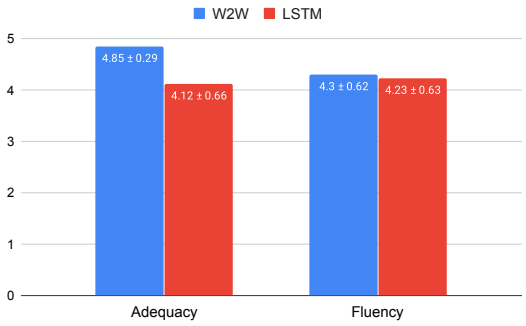


Figure 4: Human judgment on MIN→ID (Wiki)

suak" is not registered in the bilingual vocabulary. In this sample, Transformer model hallucinates as mentioning "*kehidupan awal rambut*" (early life of hair), resulting in a poor fluency and adequacy. Next in ID→MIN, W2W translation is better than LSTM and Transformer. W2W translation is relatively good, despite word "*ditemui*" (found) that is not translated into Minangkabau. In this example, LSTM and Transformer hallucinate mentioning incorrect location such as "*moldavia*", and "*alaska*".

For further analysis, we conduct a manual evaluation on two best models: W2W and LSTM in MIN→ID (Wiki) experiment. Like manual evaluation in Section 4.1, we ask two native Indonesian and Minangkabau speakers to examine the adequacy and fluency of 100 random samples with scale 1–5. Figure 1 shows that W2W translation significantly better than LSTM in terms of adequacy, but similar in terms of fluency. This is in line with our observation, that the LSTM model frequently generates incorrect keywords in a fluent and coherent translation. This is possibly due to the out-of-vocabulary (OOV) case in the test set, triggered by the small-size of our train set. A proper training scheme for the low-resource setting can be leveraged in future work, so it can reduce hallucination issue in the LSTM model.

5 Conclusion

In this work, we have shown the first NLP tasks in Minangkabau language. In sentiment analysis task, we found the necessity of indigenous language corpus for classifying Indonesian texts. Although Indonesian and Minangkabau languages are from *Malayic* family, the Indonesian model can not optimally classify Minangkabau text. Next, in the machine translation experiment, although the word-to-word translation is superior to LSTM and Transformer, there is still a room of improvement for fluency. This can be addressed by training seq-to-seq model with a larger corpus.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful feedback and suggestions. In this research, Fajri Koto is supported by the Australia Awards Scholarship (AAS), funded by Department of Foreign Affairs and Trade (DFAT) Australia.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. pages 30–38.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *A statistical approach to machine translation*, 16(2): 79–85.

Jaime G Carbonell, Richard E Cullinford, and Anatole V Gershman. 1978. *Knowledge-based machine translation.* Technical report, Yale University, Department of Computer Science, Connecticut, US.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese–Japanese Parallel Corpus from Wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. pages 642–647.

Abigail C Cohn, and Maya Ravindranath. 2014. Local languages in Indonesia: Language maintenance or language shift. *Linguistik Indonesia*, 32(2): 131–148.

Sophie Elizabeth Crouch. 2009. *Voice and verb morphology in Minangkabau, a language of West Sumatra, Indonesia.* Master Thesis, The University of Western Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 4171–4186.

Jane Drakard. 1999. *A Kingdom of Words: Language and Power in Sumatra.*

William A. Gale, and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistic*, 19(1): 75–102.

Raymond G. Gordon. 2005. *Ethnologue: languages of the world, Fifteenth Edition.* SIL International, Dallas, Texas.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono, and Mirna Adriani. 2015. Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. pages 1348–1353.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*.

Philipp Koehn 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.

Fajri Koto and Mirna Adriani. 2015. A Comparative Study on Twitter Sentiment Analysis: Which Features are Good? In *20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015*. pages 453–457.

Fajri Koto and Mirna Adriani. 2015. The Use of POS Sequence for Analyzing Sentence Pattern in Twitter Sentiment Analysis. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*. pages 547–551.

Fajri Koto and Gemala Y. Rahmaningtyas. 2017. InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*. pages 391–394.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The IIT Bombay English-Hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. pages 74–81.

Shuhua Monica Liu and Jiun-Hung Chen. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems With Applications*, 42(3): 1083–1093

Bing Liu. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*. pages 627–666.

Bin Lu, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *IJCAI'17 Proceedings of the 26th International Joint Conference on Artificial Intelligence*. pages 4068–4074.

Gérard Moussay. 1998. Tata Bahasa Minangkabau. Kepustakaan Populer Gramedia, Jakarta.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*. pages 173–180.

Hiroki Nomoto, Kenji Okano, Sunisa Wittayapanyanon, and Junta Nomura. 2019. Interpersonal meaning annotation for Asian language corpora: The case of TUFS Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fifth Annual Meeting of the Association for Natural Language Processing*. pages 846–849.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis. In *SLTU/CCURL@LREC*. pages 131–138.

Yanuar Nurdiansyah, Saiful Bukhori, and Rahmad Hidayat. 2018. Sentiment Analysis System for Movie Review in Bahasa Indonesia using Naive Bayes Classifier Method. In *Journal of Physics: Conference Series*. pages 12011.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. pages 311–318.

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.

Ryan Armiditya Pratama, Arie Ardiyanti Suryani, and Warih Maharani. 2020. Part of Speech Tagging for Javanese Language with Hidden Markov Model. In *Journal of Computer Science and Informatics Engineering (J-Cosine)*, 4(1): 84–91.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. In *Machine Translation*, 25(4): 341–375.

Hammam Riza. 2008. Resources Report on Languages of Indonesia. In *ALR@IJCNLP*. pages 93–94.

Marah Rusmali, Amir Hakim Usman, Syahwin Nikelas, Nurzuir Husin, Busri Busri, Agusli Lana, M. Yamin, Isna Sulastri, and Irfani Basri. 1985. Kamus Minangkabau – Indonesia. Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan, Jakarta.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 403–411.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *NAACL-HLT*. pages 380–385.

Arie Ardiyanti Suryani, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. 2015. Experiment on a phrase-based statistical machine translation using POS Tag information for Sundanese into Indonesian. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. pages 1–6.

Arie Ardiyanti Suryani, Isye Arieshanti, Banu W. Yohanes, M. Subair, Sari D. Budiwati, and Bagus S. Rintyarna. 2016. Enriching English into Sundanese and Javanese translation list using pivot language. In *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. pages 167–171.

Bayu Distiawan Trisedya and Dyah Inastra. 2014. Creating Indonesian-Javanese parallel corpora using wikipedia articles. In *2014 International Conference on Advanced Computer Science and Information System*. pages 239–245.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. pages 72–79.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. pages 5998–6008.

Hao Wang, Dogan Can, Abe Kazemzadeh, Franois Bar, and Shrikanth Narayanan. 2012. A System for Realtime Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations*. pages 115–120.

Aji P. Wibawa, Andrew Nafalski, A. Effendi Kadarisman, and Wayan F. Mahmudy. 2013. Indonesian-to-Javanese Machine Translation. In *International journal of innovation, management and technology*.

Jaka Aris Eko Wibawa, Supheakmungkol Sarin, Chen Fang Li, Knot Pipatsrisawat, Keshan Sodimana, Oddur

Kjartansson, Alexander Gutkin, Martin Jansche, and Linne Ha. 2018. Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *2019 Conference on Empirical Methods in Natural Language Processing*. pages 833–844.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. *arXiv preprint arXiv:1904.02232*.

Keiji Yasuda and Eiichiro Sumita. 2008. Method for Building Sentence-Aligned Corpus from Wikipedia. In *2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08)*. pages 263–268.

Kuat Yessenov and Saša Misailovic. 2009. Sentiment analysis of movie review comments. In *Methodology*, 17: 1–7.

Bing Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of 2002 IEEE International Conference on Data Mining, 2002*. pages 745–748.

Vowel Effects on L2 Perception of English Consonants by Advanced Learners of English

Yizhou Lan
Shenzhen University
3688 Nanhai Avenue, Nanshan District,
Shenzhen
yzlan@szu.edu.cn

Abstract

A perceptual identification experiment has been designed to study the effects of vowel context in Mandarin speakers' perceptual identification of L2 English fricatives and affricates. The identification task elicited the preferred Mandarin equivalent and a fitness rate of each of about 70 stimuli words with these English consonants from 67 Mandarin-speaking advanced learners of English (advanced EFL learners). The degree of mapping between Mandarin and English consonants, ranging from poor to fair, and good, were compared against predictions by the Perceptual Assimilation Model, a theoretic model that predicts learning outcomes by phonetic distances. Overall, the perceived phonetic distances between Mandarin and English consonants predicted the learners' correct identification of the L2 consonants except for a few number of individual sounds, which showed enormous vowel variation. The variation of mapping patterns across vowel context data and individual sound results suggests that factors other than phonetic similarity, such as articulatory gestures should be accounted for in the prediction of L2 learning outcomes. The findings are discussed along the lines of gestural economy, present L2 speech learning models and pedagogical applications in EFL classrooms.

1 Introduction

Pronunciation was once regarded as a "peripheral" and "unteachable" component in ESL/EFL teaching and learning, especially in the trending of "meaning-focused", "communicative" context, which is an area of SLA research that lacks the marriage between theory and practice (Derwing, 2008). However, when communicating with native speakers, learners who fail to discern phonemes of meaningful significance may result in misinterpretation of the message from the interlocutor and thus end up with increased learner anxiety or lack of confidence. Therefore, L1 influences were often identified as sources of these cases. As effective aids to predict L2 sound acquisition issues under L1 influence, speech learning models such as Best (1995) and Flege (2003) respectively used the perceived L1-L2 distance as an indicator of learner performance.

In the Perceptual Assimilation Model (PAM), Best (1995) and Best & Tyler (2007) denote that three native to non-native assimilation types (TC, CG and SC) are discernible when a learner mentally compare two related incoming L2 categories with our native categorie(s) in storage. The L2 sounds may be assimilated to two different L1 sounds, which is the Two Category (TC) type, or to a single L1 category, the Single Category type (SC), or alternatively to a single native category with one being a better candidate than the other, the Category Goodness type (CG). PAM's postulations also include predictions of levels of learners' difficulties in comprehending L2 sounds. The easiest is the TC, then CG, and the hardest one being SC. On the other hand, SLM posits that speakers' L1 and L2 sound systems interact and exist in a common phonological space (Flege, 2003). Whether new L2 phonetic categories are established or not depends on the perceived dissimilarities of an L2 sound from the closest L1 or L2 sounds. Learners' ability to establish such new phonetic categories increases with increased L2 experience. Too close similarity actually blocks the formation of new L2 categories (Flege, 2003, Flege et al., 2003)

The Mandarin Chinese language has a rich inventory of fricative and affricate consonants ("z/ts/" "c/ts^h/" "s/s/" "j/tɕ/", "q/tɕ^h/", "x/ɕ/", "zh/tʂ/", "ch/tʂ^h/", "sh/ʂ/,", "r/ʐ/") that is more densely categorized than that of English (/f, v, s, z, ʃ, ʒ, ʧ, ʤ/). Specifically, all English fricatives are laminal sounds, i.e., articulation with the tongue tip pointing downwards, but Chinese has a distinction of laminal and apical (upward pointing of tongue tip) sounds.

Wang and Chen (2019) did an experimental study on non-native speakers' perception of Chinese L2 sounds by elementary and intermediate learners, in which they addressed and affirmed the robustness of PAM. The study has explored what English substitutes (among /s, ʃ, ʧ, ʤ, ʒ, t, z, r/) English learners of Chinese resort to when they have to choose one to label Chinese fricatives and affricates audios being played. Their findings suggest that non-native speakers' identification of sounds like "q/tɕ^h/" and "c/ts^h/" are often two- or three-folds, suggesting complicated assimilation patterns. However, their study did not further inquire into whether phonetic contexts play a role in L2 perception.

As pointed out in Lan (2013), certain vowel context can trigger co-articulation in L1 and L2 production, and such gestural economy may also in turn covertly result in varied L2 perception accuracies. Therefore, intending to add original contribution to current literature, this study is especially interested in the impact of vowel context may have on consonant perception.

On a slightly different note, the study is original in two more aspects. In the previous literature, the assimilation types of English fricatives and affricates were experimented mostly on naive listeners. Few studies up to date has given a detailed account of the actual assimilation types as well as learning outcomes of advanced Chinese learners of English who have already received considerably abundant numbers of input. Moreover, previous literature has reported that English learners have problems in distinguishing these phonetic affricates /tr, dr, ts/ from other real phonemic affricates /ʧ, ʤ/ (Lee, 2003; Cruttenden, 2014). In this study, such phonetic affricates (/tr, dr/) were included to find if such confusion exist for Chinese learners.

All being considered, this study aims to tackle the following three research questions based on the previous findings. The study is believed to be significant in finding the following:

Q1. How do Mandarin Chinese advanced learners of English categorize English fricatives and affricates (including phonetic affricates) based on Mandarin Chinese candidates?

Q2. Do vowel contexts of stimuli interact with the consonant perception?

Q3. Do findings align well with the PAM predictions of category goodness? If not, how can it be explained?

2 Methods

2.1 Participants

Participants are 67 students in 2 parallel English classes the author was teaching in the fall semester of 2019/20. They were all majoring in biomedical sciences, attending a top-ranking provincial university in Guangdong Province, China. Their first language vary in Mandarin (56 cases) and Cantonese (11 cases), but they all use Mandarin daily in their boarding school-life. The average period of formal EFL instruction was around 10-12 years when the experiment took place, and none of the participants and their parents had lived abroad. Their English scores on the National College Entrance Examination ranged between 114 and 138 (out of 150), with an average of 120. Considering their test results and duration of formal English learning, the participants could be collectively described as upper intermediate to advanced learners of English. Prior to the experiment, they have all gone through a 30-minute English IPA training so that they can identify the proper IPAs for the English affricates.

2.2 Stimuli

The stimuli material of the identification task are English fricatives and affricates produced by a professionally-trained, 30 years old, male native American English phonetician, as AmE is the most received variety of pronunciation in that specific area in China, and most participants followed AmE as the language model in their pre-tertiary English-education materials. The target stimuli used in the experiment are monosyllabic English words in CVC structure with /s, ʃ ɹ, ʒ, ʧ, ʤ, tr, dr/ as initials except for /ʒ/. Since its lack of onset instances, two

CVʒVC structured words were used instead. Each target phoneme contains variations of three stimuli words, respectively in three extreme vowel contexts /i, æ(a), u/. When a frequent word does not exist under a certain vowel condition, we have increased the number in other vowel contexts to make up the sum of 3. Apart from target stimuli, control words also accompany the target words. We have used plosives irrelevant to the current inquiry (/t, d/) as control fillers, each with three instances, too. The complete list of stimuli can be found in the following table:

Con-sonant	/i/	/æ/	/u/	2nd syllable
s	seed	sack	sued	
ʃ	sheep	shack		
ʧ	cheap	chat	chewed	
ʤ	jeep	Jack	Jude	
tr		track	truth troop	
dr	dream	drag	drew	
ʒ	genre Jean			visual

Table 1: Stimuli words grouped by consonants and vowels.

2.3 Procedure

Recording of perception experimental material was prepared 1 month prior to the study. During the recording phase, the AmE speaker was asked to read aloud the stimuli in front of a MD recorder in a sound booth. The recording sampling rate was set at 44100 Hz in mono channel. The recorded stimuli were put in a carrier sentence "Now I say _____", which was also produced by the AmE speaker. The recordings were edited and composed on a PC computer using Praat (Boersma and Weenink, 2020). The target words were separated from the sentences using waveform editing, normalized for peak volume, and saved as wave form for presentations. The stimuli were arranged in random order. The students were given an ISI of 8 seconds after the stimuli to identify the onset of the given syllable of that English word by selecting one of Chinese pinyin "z/ts/" "c/ts^h/" "s/s/" "j/tɕ/", "q/tɕ^h/", "x/ɕ/", "zh/tʂ/", "ch/tʂ^h/", "sh/ʂ/,", "y/j/", "r/ʐ/".

In the identification experiment, the students are required to listen to the audio presentations designed as above, and press on a button on their cellphone representing those pinyin choices through an online instant-respond survey system (www.wjx.cn). All participants were then asked to rate the goodness of the English sounds with regard to Mandarin. The goodness was represented in a Likert scale of 0 (very poor) - 5 (exactly the same) on the same system.

One methodological specificity on the presentation of stimuli in tasks have taken controversy: shall we use IPA or Romanization in the presentation of token choices? The commonly used method of phoneme inventory comparisons is not sufficient as the IPA symbols do not provide the detailed phonetic properties of sounds across languages. Especially, Mainland Chinese students often learn *pinyin*, a Latinate Romanization for the language, before the Chinese writing system, hence opening to the possibility that the coincidences of orthography of Chinese and English may play a part in their confusion of L2 sounds in actual acquisition. Therefore, we have utilized pinyin in all L1 Mandarin identification force-choice tasks.

3 Results

3.1 General findings

Overall, the Mandarin speakers had shown organic and yet varied results of cross-linguistic identification. Figure 1 below lays out a straightforward quantitative pattern of participants' choice of sounds, showing a general tendency of identifications patterns over sibilants, affricates and /ʒ/.

The fricative sounds /s, ʃ, ʒ/ behaved differently in the learner's perceptual categorization. /ʃ/ only map on /ʃ/ with a good 4.3 rating out of 5 and 75% categorization. /s/ on the other hand can be mapped both on "s" and "x" , with "s" being a better candidate. The situation of /ʒ/ is more complicated with three divided identifications, with the best candidate /r/ only taking up 34.2% of all identifications with the rating fitness at 3.5. The rest two candidates were "y", with 14.7% identification and "zh" at 11.8%, both accompanying low fitness rates of less than 3. More cross-affricate confusion were exposed in the identification of affricates /ʤ, dr, ʧ, tr/. Both /ʤ/ and /dr/ can be mapped on to j and zh with varying percentages from 22.9 to 55.4, with moderate levels of fitness ranging from 3.3 to 4.3. What worth noticing is that the assimilation of /ʤ/ favors "j" and that of /dr/ favors "zh". As for /ʧ/ and /tr/,

both are mostly mapped onto “ch” at over 70%, and the fitness ratings were at a high level of 4.4 and 4.1. /ʧ/ can be identified with an alternative mapping candidate “q”, but with much less percentage, only 25%, but participants who have chosen it showed a high fitness rate at 4.2.

Finally, participants’ perception of control sounds has witnessed over 96% correct categorizations of "t" and "d". The English plosives has voice onset times different from those of Mandarin but they have been categorized as TC assimilation. Therefore, Mandarin listeners of English naturally map aspirated affricates onto voiceless ones; and unaspirated onto voiced ones.

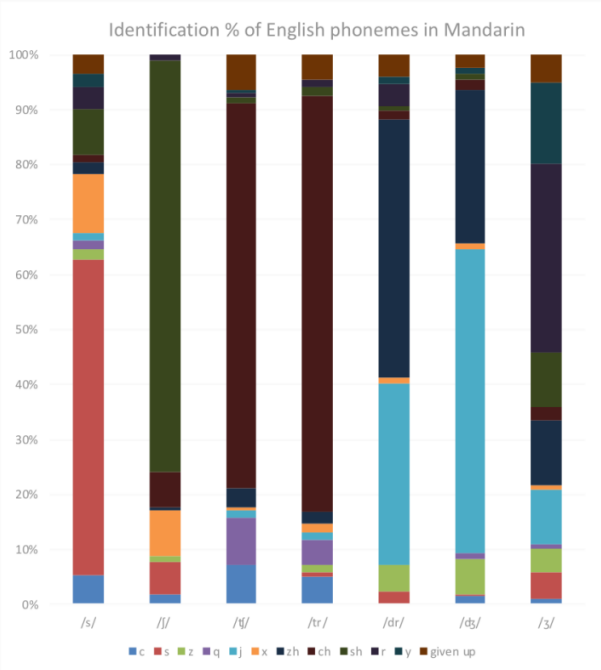


Figure 1: The % identifications of English consonants as each Chinese consonant, stacked to 100%.

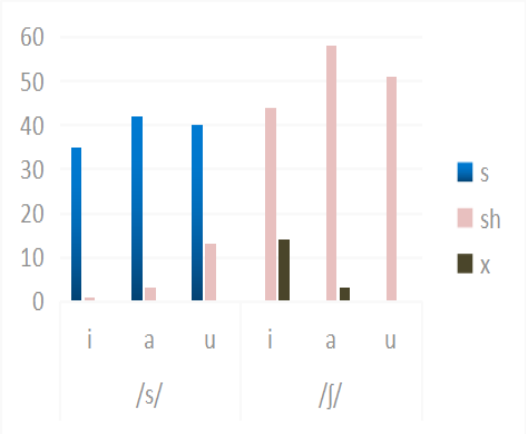
3.2 Variation across vowel contexts

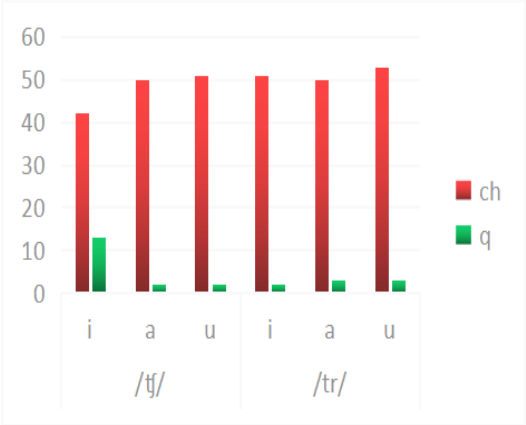
Surprisingly, distinctive perceptual assimilation patterns were seen across vowels within each consonant condition. The following Figure 3 demonstrates the frequency of choice of Mandarin candidates in different consonant (pairs) under vowel contexts of /i, a, u/. The order of the presentation is /s, ʃ/, /ʧ, tr/, /ʤ, dr/ and /ʒ/ from the upper to the lower panels. In the /s, ʃ/ pair, although the consonants favored “s” and “sh” respectively in general (both at more than 80%), the choice of other stand-alone candidates across vowels showed variation. As for /s/, 13 out of 134 outlier cases, a significant higher rate of choosing “sh” in the /u/ context than 1-4 cases in other two contexts can be identified, whereas for /ʃ/, only consonants in words with vowel /i/ are prone to be categorized as “x”, at 14 out of 170. A reverse pattern can be identified.

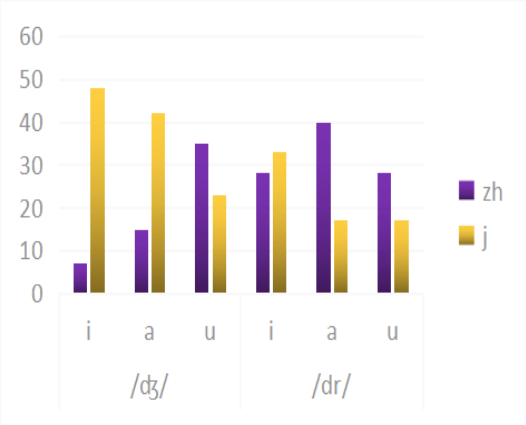
The vowel-context variation for /ʧ/ showed that the sounds with /i/ behaves slightly different from those with the other vowels: /æ(a)/ and /u/ both had fewer than 3 cases categorized as “q” but /i/ had a considerable 13 cases as “q”. As for “tr” however, almost all sounds were mapped onto “ch” regardless of the vowel context, all with exceptions of as few as 3 instances of categorizations as “q”.

The /ʤ, dr/ pair showed a distinct vowel context variation that worth noticing. In terms of general tendency, participants associate /ʤ/ with “j” and /dr/ with “zh”. Though both sounds can be categorized as “zh” and “j”, /ʤ/ in the /u/ context showed a reversed pattern against the other two contexts, favoring “zh” at 35 cases of “zh” to 23 cases of “j”, whereas /dr/ in the /i/ context behaved opposite from /tr/, favoring “j” at 28 cases of “zh” to 33 cases of “j”.

The behavior of the consonant /ʒ/ is the most complicated across vowels within these four panels of graphs. Participants’ categorization of /ʒ/ as “r” exists in all three conditions, and topped in /i/ at 49 cases, but significantly fewer in /u/ conditions at only 19 cases. The “zh” choice exist only in /i/ and /æ(a)/ contexts, but not /u/. The third candidate “y” is the favored choice of participants only in the /u/ context, which was labeled 26 instances among the total 45 cases in this vowel context, well over half; but not popular at all in the other two contexts, constituting a distinct pattern. There is a clear reverse pattern between /r/ and /y/ in the /i/ and /u/ vowel contexts.







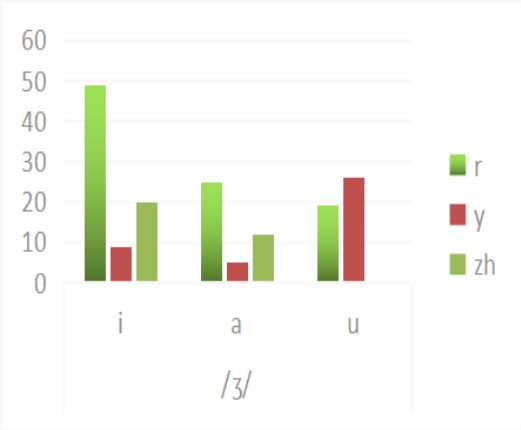


Figure 2: The % correct Mandarin categorizations of English consonants (From top to bottom: /s, ʃ, /ʧ, tr/, /ʤ, dr/ and /ʒ/) clustered by vowels.

4 Discussions

As predicted, Mandarin candidate consonants were mapped onto English sounds within a range of fitness ratings from 1.0 (poorest) to 6.4 (best) by native Mandarin listeners. Some assimilation patterns elicited from the results do conform to predictions by PAM. For example, English sibilants /s, ʃ/ showed largely similar assimilation tendencies as reported in the previous literature.

However, the variation within categories may not be predicted by PAM. For example, /ʧ/ and /tr/, which supposedly constituted a classic SC (single category) assimilation pattern, showed CG-like patterns. For example, PAM could explain that some participants preferred the "ch" in Chinese as a candidate for /tr/. Since the /r/ sound in English does not exist in Chinese, so this "cluster" is seldom considered as separated sounds by Chinese listeners. So it is reasonable to predict that they would regard "ch" as the equivalent to /tr/. However, some other listeners preferred the candidate "q" but only in limited vowel contexts of /i/ disregarding its similarity to "ch". This is probably due to the fact that "q" and /tr/ share similar places of articulation in the vowel context of /i/, where the palatal "q" and post-alveolar /tr/ approaches physiologically. This leads to the main argument of this research that vowel context may greatly shift the L1-L2 perceptual space, which questions the idea that such a space may remain intact during a specific learning stage (Flege, 2003).

The variation of vowel context shown in many consonant perceptions could be traced of

physiological roots. For example, the /s/ and /ʃ/ context showing /i/ and /u/ as behaving differently from the other two contexts is clearly affected from the gestural proximity of articulatory gestures. The /s/+/u/ combination will result in palatalization because the upward movement of the tongue dorsom by /u/ can pull backwards the articulation of fricative /s/, creating post-alveolar frication (cf. Lan, 2013). Similarly, /ʃ/+/i/ in the front vowel context will add the degree of frontness of the tongue tip to the post-alveolar sound, pulling the tongue body further front, palatalizing the sound to make it similar to the Chinese "x" which only exists before front vowels /i, y/ phonologically.

Vowel influence can also be seen in /ʧ, ʤ, dr/. The /ʧ/ sound being categorized as "q" and /ʤ/ as "j" in the /ʧ/+/i/ or /ʤ/+/i/ combination has a phonological constraint of only allowing front vowels /i, y/ after "j" and "q", which is in line with gestural economy: the palatalization may come from the co-articulation due to increased degree of frontness of the tongue tip (see Browman and Goldstein, 1992; Gick et. al., 2006 for similar examples of /r/ in different vowel contexts). Such a tendency can also be seen grammaticalized in many loanwords such as Jeep (ji pu), Cheetah (qi ta) and so on. As for the /u/ context, /ʤ/+/u/ is often regarded as a retroflex affricate in Chinese, which was also noticed as an unexpected finding in Wang and Chen (2019). What they did not specify is that these two sounds share the common place of articulation, and only differs in apicality. Therefore, we can see that the difference between apical and laminal sounds was considered not sensitive by Mandarin speakers of English, who are more sensitive to places of articulation.

The English voiced fricative /ʒ/ is another sound that has caused mass confusion in categorization. It is generally categorized as "r", or the semi-vowel "y"(/j/). The distribution of the assimilation shows that "r" is the preferred variant especially in the context of /i/ and "y" mainly preferred in the context of /u/. The reason behind it may lie in that "r" as a fricative sound /ʐ/ shares similar place of articulation (post-alveolar) and only differs from /ʒ/ in terms of apicality. Phonologically, "r" can only exist in the context of a front vowel /ɹ /, which is an allophonic variation and graphic equivalent of /i/ in Mandarin. However, as for /u/, the reason of learners favoring "y" is unclear.

Although /dr/ and /ʤ/ are both similar to the "zh" sound in Mandarin, they are not assimilated into a single category SC because an additional CG candidate, "j" also exists in the Mandarin phonemic inventory, with "zh" as the more preferred target (70%) for "dr". Although Wang and Chen (2019) found something similar, our study showed ,interestingly, that /ʤ/ has the more similar representation with "zh" when considered phonetically: "zh" is more phonetically similar to /ʤ/, which is proven in the very stable categorization of /ʃ/ as "sh", but not the other candidate "x", which only differs in voicing (see Table 2). Therefore, we could conclude that the above findings cannot be solely attributed to phonetic similarity claimed by PAM. A possible explanation to this anomaly is the orthographic coincidence of Chinese "j" and English /ʤ/ in Romanization. Such orthographic influence will be studied and argued for in another study.

Mandarin	phonetic difference	% of identification	predicted match
/ʃ/			
x	place *	9.5%	poor
sh	manner (apicality)	75%	very good
/ʤ/			
zh	manner (apicality)	22.9%	poor
j	both manner and place *	55.4%	good to very good

Table 2: Assimilation mismatches of English /ʃ/ and /ʤ/ in terms of phonetic-similarity-predicted match and actual identification proportion.

Finally, the findings reflects on Derwing and Munro (2004) and Major (2014)'s concern that in actual L2 perception, learners may utilize other phonetic details such as duration, or phonetic context, which may win them high confidence in discerning those sounds. But, some of them may not be correct and helpful to learning, which can be seen in various modes of assimilation. As for the SC-type pairs in the current experiment, i.e., English /ʧ/ and /tr/, learners show very high identification as Chinese "ch" (70.1% and 75% respectively), with high goodness ratings. For TC pairs, For example, the /ʃ/ sound has experienced a relatively low accuracy compared to the apparent categorization could be predicted in the L1 Chinese candidate "sh". However, both

identifications may lead to low accuracy in actual perception, though not attested in the current inquiry. That being said, a good category for L1-L2 mapping may be still perceptually confusing. At this point, we could affirm that the perception of L2 sounds is more than a low-level acoustic process that is merely generated by automatic distance-comparison, but must include higher-level cognitive processes that alter with par linguistic and discourse contexts.

5 Conclusion

The study presents a novel design in an attempt to understand the vowel influence on L2 consonant perception. Both theoretical and pedagogical implications can be drawn from the findings. Much of the assimilation patterns surfaced as predicted by those from TC, CG and SC types in the PAM model. However, on the other hand, the perceived phonetic distance between L1 and L2 is not the only factor in play. We have witness that the assimilation patterns were under the influence from physiological attributes such as gestural economy levels in various vowel contexts.

Theoretically, we could see from the results that vowel contexts clearly interact with L2 phonology. In other words, the universal human speech apparatus enables the phonetic sounds to express in phonology without boundaries, which has affected and complicated the "should-be" holistic interlanguage in phonological terms. The L2 evidence may be seen as a side proof of the view of gestural nature of phonological perception (Liberman and Mattingly, 1985; Browman and Goldstein, 1992).

Pedagogically, the study invites instructors and learners to admire and tackle the complexity of speech learning and the L1 influence, especially in families of sounds like fricatives and affricates of L1 Chinese and L2 English where multiple entangling mappings can be found. Future studies should include perceptual accuracy tests and production tests to further solidify the implications of current findings, especially whether the actual learner accuracy performance was as predicted by theoretic models so that more theoretical and suggestions can be made. The non-linguistic effect, for example orthographic influences, should also be carefully explored so as to discover more myths in the complex system of L2 consonant perception.

Acknowledgements

This study is financially supported by the Shenzhen University Young Scholar Start-up Fund (No. 2019082).

References

A. Cruttenden. 2014. Gimson's Pronunciation of English. 8th edition, Routledge Tylor & Francis, England.

A. M. Liberman and I. G. Mattingly. 1985. The motor theory of speech perception revised. Cognition 2 (1): 1–36.

B. Gick, F. Campbell, S. Oh, & L. Tamburri-Watt. 2006. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. Journal of Phonetics, 34, 49-72.

C. P. Browman and L. Goldstein. 1992. Articulatory phonology: an overview. Phonetica, 49, 155-180.

C. T. Best. 1995. A direct-realist view of cross-language speech perception," Speech perception and linguistic experience: Issues in cross-language research, 171-204.

C. T. Best and M. Tyler. 2007. Nonnative and second-language speech perception: commonalities and complementarities. In O.S. Bohn & M. Munro (Eds.). Second-language Speech Learning: the Role of Language Experience in Speech Perception and Production. A Festschrift in Honour of James. E. Flege. Amsterdam: John Benjamins, 13-34.

J. E. Flege. 2003. Assessing constraints on second-language segmental production and perception," Phonetics and phonology in language comprehension and production: Differences and similarities, 319–355.

J. E. Flege, C., Schirru., and I. R. A. MacKay. 2003. Interaction between the native and second language phonetic subsystems, Speech Communication, 40, 467-491.

P. Boersma and D. Weenink. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.09, retrieved 26 January 2020 from http://www.praat.org/

R. C. Major. 2014. Foreign Accent—the ontogeny and phylogeny of second language phonology. New York: Routledge.

S. Lee. 2003. A comparison of cluster realizations in first and second language. The Journal of Studies in Language, 19(2): 341-357.

T. Derwing and M. J. Munro. 2005. Second Language Accent and Pronunciation Teaching: A Research‐Based Approach. TESOL Quarterly 39 (3): 379-397.

T. Derwing. 2008. Curriculum issues in teaching pronunciation to second language learners. In J. Hansen Edwards & M. Zampini (Eds.). Phonology and Second Language Acquisition. Amsterdam, Netherlands: John Benjamins, 347-369.

X. Wang and J. Chen. 2019. English Speakers' Perception of Mandarin Consonants: The Effect of Phonetic Distances and L2 Experience. ICPhS 2019 Proceedings, 1-5.

Y. Lan. 2013. Towards a Revised Motor Theory of L2 Speech Perception, Proceedings of PACLIC 27, 136-142.

Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features

Andreas Liesenfeld, Gábor Parti, Yu-Yin Hsu, Chu-Ren Huang
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong
amliese@polyu.edu.hk, gabor.parti@connect.polyu.hk,
yu-yin.hsu@polyu.edu.hk, churen.huang@polyu.edu.hk

Abstract

This paper examines gender and age salience and (stereo)typicality in British English talk with the aim to predict gender and age categories based on lexical, phrasal and turn-taking features. We examine the SpokenBNC, a corpus of around 11.4 million words of British English conversations and identify behavioural differences between speakers that are labelled for gender and age categories. We explore differences in language use and turn-taking dynamics and identify a range of characteristics that set the categories apart. We find that female speakers tend to produce more and slightly longer turns, while turns by male speakers feature a higher type-token ratio and a distinct range of minimal particles such as "eh", "uh" and "em". Across age groups, we observe, for instance, that swear words and laughter characterize young speakers' talk, while old speakers tend to produce more truncated words. We then use the observed characteristics to predict gender and age labels of speakers per conversation and per turn as a classification task, showing that non-lexical utterances such as minimal particles that are usually left out of dialog data can contribute to setting the categories apart.

Author's note (Oct 2020): statement on the use of social categories in this study *This work involves the labelling of participants for social categories related to gender and age. We caution against the use of this heuristic due to the risk of promoting a biased view on the topics. We would like to encourage those interested in the computational modelling of social categories to join the discussion on these concerns and consider participating in efforts to build more inclusive resources for the study of the topics.*

1 Introduction

One of the most interesting topics in language studies has been on how speakers' gender and age differences influence their communicative behaviour. Transcriptions of real-world, naturally-occurring conversations provide us a window to examine such differences in talk-in-interaction.

Gendered and age-salient elements of talk have long been studied from a range of perspectives, including linguistics (e.g. Lakoff 1973, Tannen 1990), psychology (for an overview see, e.g. James and Drakich 1993, Tannen 1993), and conversation analysis (e.g. Jefferson 1988). This topic has also been extensively studied from a computational perspective, focusing on how the differences can be formally described and modelled. In recent years, the research interest has been extended to various applications using different types of data, such as using movie subtitles to identify gender distinguishing features (Schofield and Mehr, 2016); email interactions to study gender and power dynamics (Prabhakaran and Rambow, 2017); video recordings of human-robot interactions to study gendered and age-related differences in turn-taking dynamics (Skantze, 2017); literary and weblog data to study differences between male and female writing (Herring and Paolillo 2006; Argamon et al. 2003); and multimodal audiovisual and thermal sensor data for gender detection (Abouelenien et al., 2017).

Recent studies on gendered and age-salient behaviour in conversations also focus on the use of

specific constructions or classes of constructions, such as swear words (McEnery and Xiao, 2004), amplifiers (Xiao and Tao, 2007), *do* constructions (Oger, 2019), and minimal particles (Acton, 2011). In the current study, we use transcriptions of recordings of naturally-occurring talk in British English to explore distributional differences in language use across gender and age groups, testing well-known tropes such as tendencies that women speak more politely, or that men use more swear words (Baker 2014, Lakoff 1973, Tannen 1990), and also shedding light on other under-explored aspects of gendered and age-salient elements in talk such as the use of non-lexical vocalizations, laughter and other turn-taking dynamics. Our interest in this topic derives from work in computational modelling of dialog and conversation, especially studies aiming to automatically identify speaker properties for the use in voice technology and user modeling (Joshi et al. 2017, Wolters et al. 2009, Liesenfeld and Huang 2020).

This pilot study explores whether non-lexical vocalizations and turn-taking properties can contribute to the prediction of age and gender categories. We investigate this question using a dataset of naturalistic talk that includes a range of elements other than words, such as laughter, pauses, overlaps and minimal particles. Can authentic and often "disfluent" and "messy" transcriptions of natural talk be used for a classification task? How will different behavioural cues contribute to a statistical investigation and prediction of gender and age salience and typicality?

2 Data description

Our dataset comes from the Spoken BNC2014 (Love et al., 2017), a corpus of contemporary British English conversations recorded between 2012-2016. It consists of transcriptions of talk on a range of topics covering everyday life in casual settings between around 2 to 4 speakers with a wide variety of social relationships such as between family members or good friends, and among colleagues or acquaintances. For classification, we extract two subcorpora from the SpokenBNC using the speaker labels "female" and "male" as well as age labels for speakers above 70 years and under 18 years. Table 1 provides an overview of the two subcorpora. For age, we chose to only include the youngest and oldest speakers to tease out more significant differences by removing the bulk of middle-aged speakers. The downside of this approach is that this subcorpus is relatively small.

Feature	Category	Count
Speakers	Female	365
	Male	305
	Old	56
	Young	49
Words	Female	6,671,774
	Male	4,080,524
	Old	737,398
	Young	792,039
Turns	Female	742,973
	Male	478,851
	Old	96,994
	Young	102,433
Average turn length (in words)	Female	9.42
	Male	8.950
	Old	8.05
	Young	8.18
Type-token ratio	Female	0.0073
	Male	0.011
	Old	0.0231
	Young	0.0235

Table 1: Properties of the dataset obtained from the SpokenBNC2014 corpus. "Old" refers to speakers above 69 years of age, "young" includes speakers up to 18 years old.

3 Methods

Comparing the behaviour of speakers across categories, we first look at lexical and phrasal differences. Then we examine non-lexical vocalizations such as laughter, minimal particles, and turn-taking dynamics such as overlaps and pauses. For both parts, we tokenize the corpora and remove stopwords using the NLTK and SpaCy libraries (Bird et al. 2009, Honnibal and Montani 2017).

3.1 Lexical features

We select a number k_i of speaker's of each label from conversations i and build a language model

with the n-gram frequencies for all turns per category. We then examine the characteristic differences in the use of lexical items using Scaled F-score. Scaled F-score is a modified metric based on the F-score, the harmonic mean of precision and recall. It addresses issues related to harmonic means dominated by precision, as well as a better representation of low-frequency terms.[1]

We plot gender and age categories using the Scattertext library (Kessler 2017) to visualize the cross-categorical differences at the n-gram level. Figure 1 and 2 show words and phrases that are more characteristic of each category, while also reporting their frequencies and a list of top characteristic items.

We observe that a range of terms reflect (stereo)typicality of gender and age categories in our corpus. For instance, top characteristic terms of male speakers feature the nouns "mate", "game", "cards", "quid" and "football", while female speaker's talk more prominently features "baby", "weekend", "hair", "birthday" and "cake" (see Figure 1). More interesting for us, the characteristic terms per gender category also feature a number of verb constructions, exclamations and minimal particles such as "ain't", "innit", "eh" and "uh" for male speakers and "my God", "mhm", "blah". "huh" and "hm" for female.

For age categories, notably a much smaller corpus, we also observe that a range of items features more prominently in talk of speakers labelled as young or old (see Figure 2). Likewise, we observe that some non-lexical utterances, exclamations and particles exhibit category salience, such as "em", "innit" and "oh dear". Based on these observations we decide to further explore the role of non-lexical vocalizations, exclamations and minimal particles in the corpus.

1 Positive responses and continuers:
S1: you're so good at hair
S2: really?
S1: **mm**
S2: hair is my weakness I feel like I'm really bad

S1: no I'd rather sit inside
S2: **uhu**
S1: if it was just a little bit sunnier

2 Turn stalling:
S1: you don't like riding them?
S2: I do but [short pause] **hmm** [short pause] you don't really
S3: [overlapping] I don't have a bike

3 Turn management:
S1: oh lemon balm yeah you can do that as well
S2: **erm** what what is very good for colds i [truncated] is **er erm** purple sage
S1: yeah pur [truncated] yeah [short pause] I know that one yeah

4 Repair initiators:
S1: are all all the actors are redubbed for the songs aren't they
S2: **hm?**
S1: are the all the actors redubbed for the songs? I can't remember

5 Change-of-state tokens:
S1: she was just awake screaming for hours
S2: **oh**
S1: so that took its toll

Table 2: Overview of minimal particle types

3.2 Non-lexical and turn-taking features

Next, we move beyond lexis and examine non-lexical vocalizations and a range of other aspects of turn-taking such as laughter, pauses, overlaps, and truncation. Our dataset contains non-lexical vocalizations of different functions, such as the minimal particles (also known as interjections) "hm", "mhm", "hmm", "er", "erm", "um", "aha", "oh". In fact, this type of utterance ranks among the most frequent in the corpus. These utterances can format a wide range of functions that may be relevant to gender and age category prediction. Unfortunately, our corpus does not annotate functional information of these utterance which makes it difficult to

[1] https://github.com/JasonKessler/scattertext

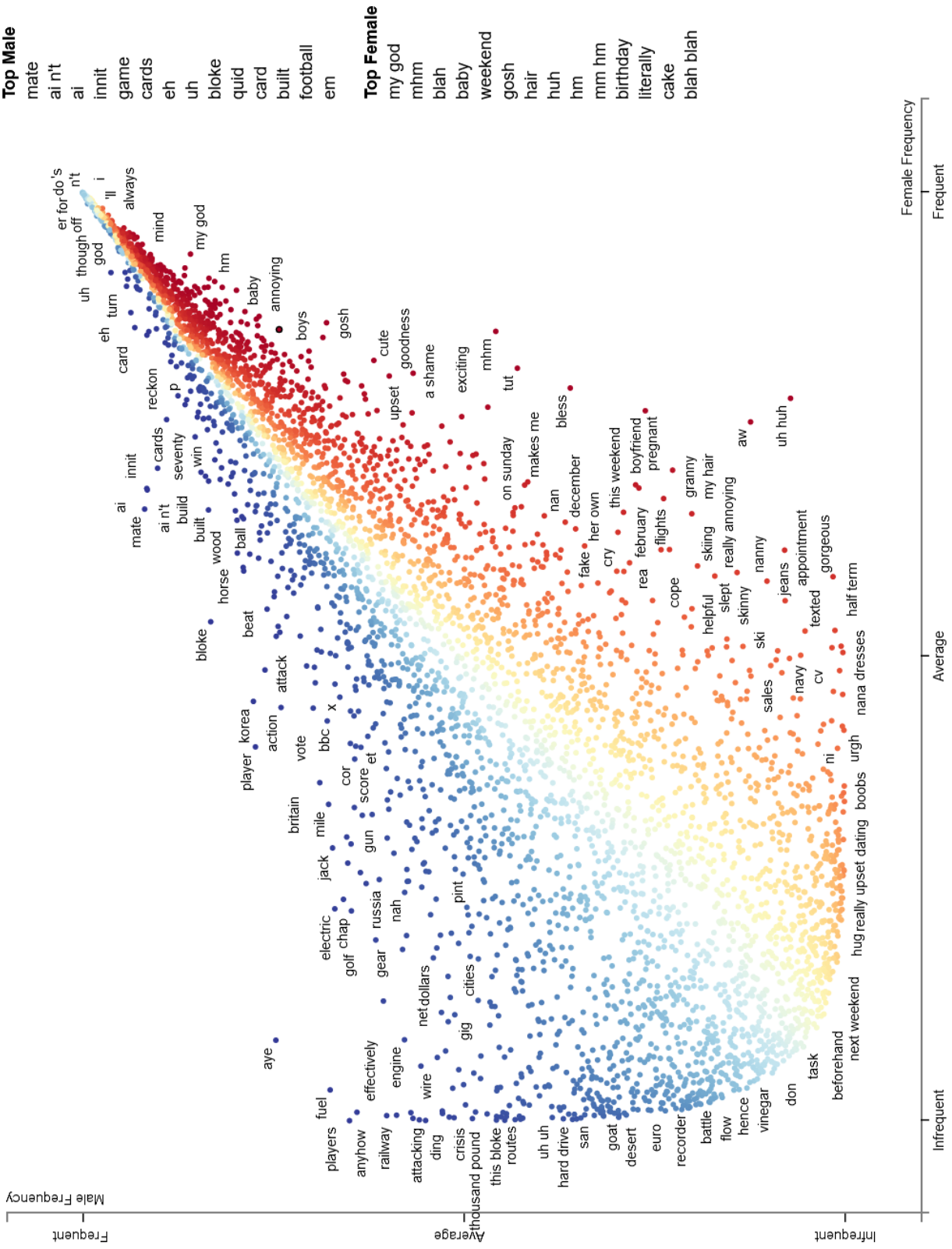


Figure 1: Overview of characteristic terms by gender category, blue=Male, Red=Female, plotted by frequency

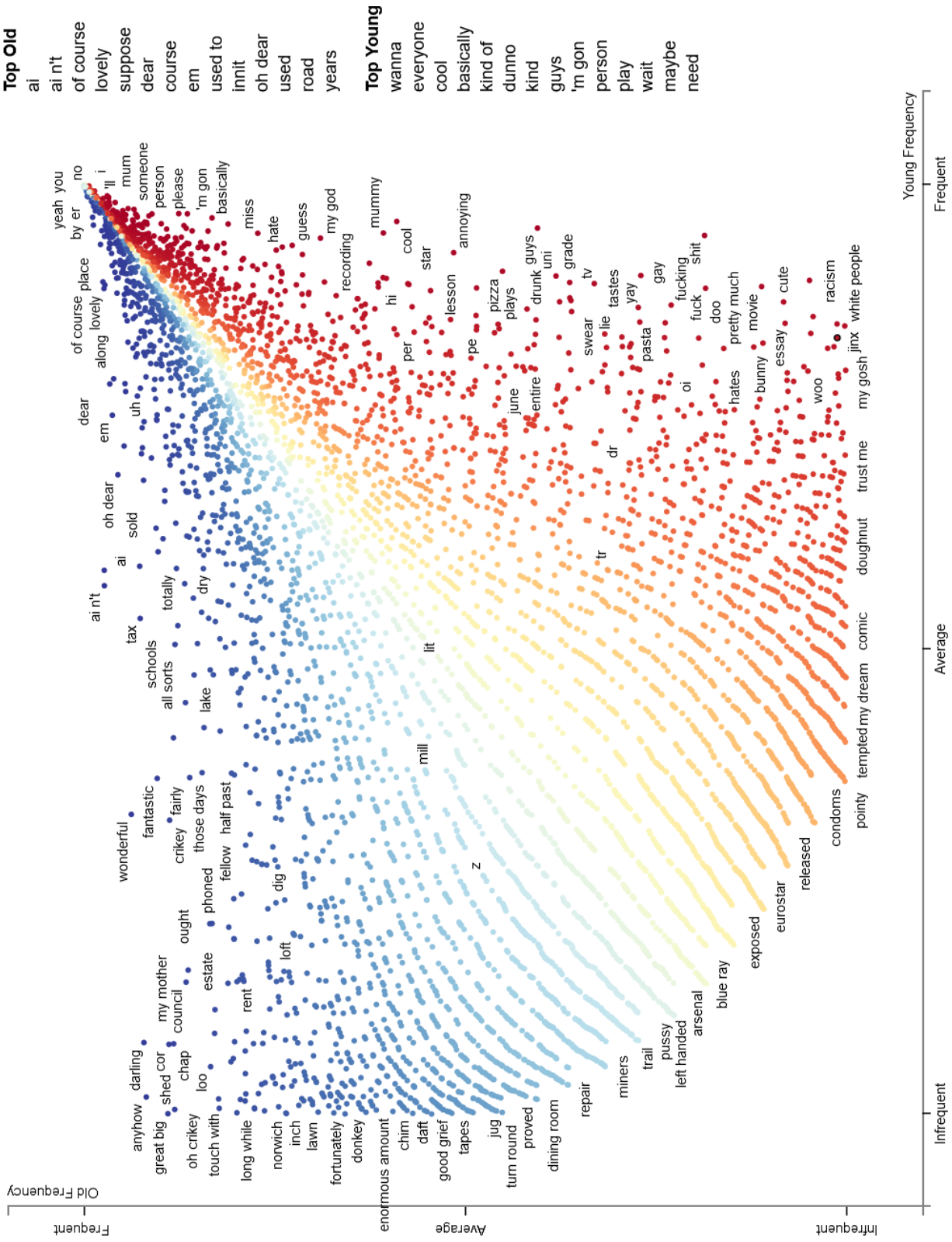


Figure 2: Overview of characteristic terms by age category, blue=Old, Red=Young, plotted by frequency

consistently group this type of utterance into functional categories in retrospect (Liesenfeld 2019b). Inspired by Couper-Kuhlen and Selting (2017), we therefor decided to only group these particles into five broad form-function mappings based on their typical forms. This way we aim to capture at least some functional variety, even though this unfortunately does not accommodate inter-speaker variation. Table 2 shows the functions we differentiate.

In addition, laughter, truncation , pauses and overlap is also annotated in our corpus as single labels that indicate the occurrence of laughter-related sounds, abandoned words, as well as the occurrence of overlap between two turns by speakers. Table 3 provides an overview of these non-lexical vocalizations and turn-taking properties.

The cells in dark blue show the highest occurrence of a feature per category as relative frequencies. For example, laughter is most prominent among young speakers, while turn management tokens (”er”, ”erm”, ”um”) are typical for old speakers. The lighter blue cells compare the prominence of the same item across categories, displayed as the percentage of the highest ranking category.

First we look at minimal particles that typically format positive responses (as for acknowledgments) and continuers. This includes nasal utterances such as “mm”, “mhm”, “mm_hm”, as well as vocalic utterances such as “aha”, “uhu”, “uhuh”, “uh_huh”. Turns by female and old speakers feature these utterances more often as those of other speakers. Second, we examine utterances typically related to turn stalling or management. We distinguish two types of forms that typically format this, nasal “hmm” and “hmmm” sounds as well as vocalic sounds “um”, “er”, and “erm”. Turn stalling tokens appear most frequently in turns by female speakers while turn management tokens appear predominantly in turns by old speakers. Next, we examine nasal utterances featuring a rising pitch which are annotated as “hm?”. This type of utterance can format doubt, disbelieve, or serve as a repair initiator. It appears predominantly in turns by female speakers (notably raw counts for this token are very low). Lastly, the utterance “oh”, that (as a response and with rising pitch) commonly formats a change-of-state token that expresses an insight or understanding, features most prominently in talk by old speakers.

4 Prediction

Can we predict the speaker’s gender and age category based on lexical, non-lexical and turn-taking features alone? And how does including non-lexical vocalizations impact the binary classification task?

4.1 Controlling the data

One challenge of working with transcriptions of unscripted conversations is that various subcorpora that one would like to compare are rarely of the same size. For binary classification it is therefore essential to select equal numbers of speakers of each category. We also checked the amount of utterances of each speaker and removed those which only produced a minimal amount of talk.

Furthermore, we considered controlling for gender pairs, making sure our subcorpora feature male-male, female-male, and female-female talk in equal measure, but ultimately we decided that, in this case, the resulting dataset would not be big enough for the task. Similarly, we decided against using a leave-one-label-out split to control for the language of a particular conversation.

4.2 Results

First, we predicted the label of a single speaker based on all their utterances. We obtained 305 speakers each for classifying gender and around 50 each for age. Especially the size of the age corpus is therefore almost unsuitable for a classification task, so we caution the reader to treat the resulting classification accuracy with a grain of salt. Using the features discussed in Section 3, we began with only considering lexical features, and then considering both lexical and non-lexical. We then train/tested (50/50) a Logistic Regression classifier to predict gender and age of each speaker with 10-fold cross-validation. We obtained a classification accuracy of 71% for gender labels and 90% for age labels using only lexical features, after added non-lexical features the accuracy increased to 81% and 92% respectively.

Second, we also tried to predict the label of a speaker per individual turn. Similarly, we split the dataset into equal amounts of turns per category (around a million turns for gender, and around 200,000 for age), and trained a classifier using a

Feature		Category			
		Female	Male	Old	Young
minimal particles	**Positive responses and continuers** (mm, mhm, mm_hm, aha, uhu, uhuh, uh_huh) gender n=86,098; age n=10,506	**highest**	67.8%	97.5%	54.6%
	Turn stalling (hmm, hmmm) gender n=2,722; age n=132	**highest**	64.2%	25%	33.4%
	Turn management (um, er, erm) gender n=98,442; age n=16,777	62.8%	77.3%	**highest**	64%
	Repair initiators (hm?) gender n=195; age n=15	**highest**	61.9%	19%	71.4%
	Change-of-state token (oh) gender n=96,566; age n=15,852	80.6%	62.4%	**highest**	71.2%
turn-taking properties and other vocalizations	**laughter** gender n=92,417; age n=15,603	72.5%	59.1%	58.6%	**highest**
	pause (short, 1-5 sec) gender n=236,885; age n=29,703	**highest**	91.2%	94.4%	76.6%
	truncated words gender n=68,122; age n=11,065	80%	89%	**highest**	91.%
	overlaps (by total turns ratio) gender n=250,628; age n=46,285	87%	80%	90%	**highest**

Table 3: Overview of non-lexical features in the dataset: minimal particles, turn-taking properties and other vocalizations (in relative frequencies, first rank is displayed as "highest" and rank 2-4 in percentage of first rank, blue and teal intensity indicates rank, n = counts of each feature by subcorpus)

90/10 train/test ratio to predict the gender and age label of a single turn. Here, we obtain 62% accuracy for gender and 67% for age using only lexical features, and 63% and 67% after adding non-lexical features.

In both cases non-lexical features increased the accuracy of the classifier which indicates that non-lexical vocalizations and other turn-taking dynamics are useful discriminators of the gender and age labels.

5 Conclusion

We examined gender and age salience and (stereo)typicality in British English conversation. The results of this pilot study show that a range of lexical, phrasal, non-lexical, and turn-taking-related features exhibit a tendency to appear more prominently across binary gender and old categories. We were especially interested in the use of non-lexical vocalizations, particles, exclamations and other turn-taking dynamics. Here, we found that female speakers produce significantly more and slightly longer turns. Talk by female speakers also tends to feature the minimal particles "huh", "hm" and "mm" more prominently. In contrast, male speakers' talk tends to be characterized by the minimal particles "eh", "uh" and "em". Overall, male speakers tend to produce shorter turns with fewer words and a higher type-token ratio.

Looking at generational differences, we found that young speakers laugh more and their turns overlap more and tends to feature more words typically related to swearing such as "shit", "fuck" or "fucking". Talk by old speakers tends to feature more truncated words and turn management tokens.

Based on such observations of characteristics across categories, we set up a classification task to predict gender and age labels of both single speakers and individual turns. We found that predicting speaker labels per conversation yields significantly higher classification accuracy in comparison to the prediction of labels for individual turns. This is likely due to the high number of very short turns that don't feature utterances with label bias. The classification results of around 80% for predicting speaker labels per conversation show that a simple logistic regression classifier does a reasonably good

Features	Gender Accuracy±Std. Error	Age Accuracy±Std. Error
per conversation		
baseline	50±0%	50±0%
lexical	70.95±5.38%	89.50±3.53%
lexical + non-lexical	80.71±5.62%	92.00±3.27%
per turn		
baseline	50±0%	50±0%
lexical	62.39±0.18%	67.33±0.34%
lexical + non-lexical	63.57±0.13%	67.74 ±0.46%

Table 4: Results of category prediction as a binary classification task of "male" - "female", "young" - "old" labels, per single-speaker and per turn

job even when confronted with "unstructured" and "messy" transcribed speech. Notably, we show that non-lexical utterances and minimal particles, which are often filtered out in dialog and speech corpus datasets, contribute to more accurate prediction.

Lastly, we would like to highlight the potential conceptual pitfalls of thinking of speaker's gender and age category prediction as a binary classification task. The use of labels for participants can lead to the dissemination of biased conceptions of gender and age salient performances in conversation. We would like to stress the need to be very cautious when making inferences based on data labelled for social categories such as used in this study. No study on related topics can be a study on computational modelling eo ipso. This underscores the need for more inclusive language resources in the area.

Nonetheless, we hope that our preliminary results yielded some interesting insights to gender and age (stereo)typicality in contemporary British English talk and will draw more attention to much-needed computational work based on authentic, real-world recordings instead of sterile, polished datasets.

6 Limitations and further studies

In the real world, gender and age performances in talk-in-interaction are not classification tasks. Challenges for big data prediction approaches are plenty. For instance, a more comprehensive model needs to take into account that speakers perform gender and age differently across various conversational settings. When more datasets become available, a natural extension to the existing prediction study would be explorations of differences across various conversational compositions. Would we observe similar patterns in conversations with speakers of the same or different gender and age?

Another important extension to the current type of study are more detailed explorations of turn-taking dynamics that look into more fine-grained aspects of different types of actions in conversation. Classifiers such as those used in this study work well as soon as they are fed items with strong class bias for prediction, but humans are often able to make informed guesses on speaker traits based on style and format of a very short sequence of talk. Modelling this requires a more detailed typology of turn types and conversational moves, which makes it necessary to dig deeper into the fine-grained systematics of talk-in-interaction. However, quantitative methods often brush away the details of how speakers format various action types, which leads to challenges of how to model the sequential unfolding of action sequences computationally (Liesenfeld, 2019a).

A critical challenge for the data-driven prediction of gender and age salience in talk is therefore how to take variation in formats of specific actions and activities into account, especially those that have been described as gendered or age-salient such as hedging or "troubles talk" (Lakoff 1973; Jefferson 1988). Focusing on specific actions would enable a more fine-grained analysis of how speakers negotiate their concepts of gender and age in interaction as part of specific sequences in conversation and how navigating these concepts in interaction relates to (stereo)typical gender and age salience.

References

Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., and Burzo, M. (2017). Multimodal gender detection. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 302–311.

Acton, E. K. (2011). On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics*, 17(2):2.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

Baker, P. (2014). *Using corpora to analyze gender*. A&C Black.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Couper-Kuhlen, E. and Selting, M. (2017). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.

Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Honnibal, M. and Montani, I. (2017). Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

James, D. and Drakich, J. (1993). Understanding gender differences in amount of talk: A critical review of research. In Tannen, D., editor, *Oxford studies in sociolinguistics. Gender and conversational interaction*, page 281–312. Oxford University Press.

Jefferson, G. (1988). On the sequential organization of troubles-talk in ordinary conversation. *Social problems*, 35(4):418–441.

Joshi, C. K., Mi, F., and Faltings, B. (2017). Personalization in goal-oriented dialog. In *NIPS 2017 Conversational AI Workshop, 4-9 Dec 2017*.

Kessler, J. S. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. In *Proceedings of ACL-2017 System Demonstrations, 30 July - 4 August 2017*, Vancouver, Canada. Association for Computational Linguistics.

Lakoff, R. (1973). Language and woman's place. *Language in society*, 2(1):45–79.

Liesenfeld, A. (2019a). *Action formation with janwai in Cantonese Chinese conversation*. PhD thesis, Nanyang Technological University.

Liesenfeld, A. (2019b). Cantonese turn-initial minimal particles: annotation of discourse-interactional functions in dialog corpora. *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)*.

Liesenfeld, A. and Huang, C. R. (2020). NameSpec Asks: What's Your Name in Chinese? A Voice Bot to Specify Chinese Personal Names through Dialog. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Love, R., Dembry, C., Hardie, A., Brezina, V., and McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

McEnery, A. and Xiao, Z. (2004). Swearing in modern British English: the case of fuck in the BNC. *Language and Literature*, 13(3):235–268.

Oger, K. (2019). A Study of Non-Finite Forms of Anaphoric do in the Spoken BNC. *Anglophonia. French Journal of English Linguistics*, (28).

Prabhakaran, V. and Rambow, O. (2017). Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse*, 8(2):21–55.

Schofield, A. and Mehr, L. (2016). Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature, 16 June, 2016*, pages 32–39.

Skantze, G. (2017). Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 9-11 March, 2017*, pages 196–204. IEEE.

Tannen, D. (1990). *You just don't understand: Women and men in conversation*. Morrow New York.

Tannen, D. (1993). *Gender and conversational interaction*. Oxford University Press.

Wolters, M., Vipperla, R., and Renals, S. (2009). Age recognition for spoken dialogue systems: Do we need it? In *Tenth Annual Conference of the International Speech Communication Association, 6-10 September 2009*.

Xiao, R. and Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic studies*, 1(2):241–273.

Simple is Better! Lightweight Data Augmentation for Low Resource Slot Filling and Intent Classification

Samuel Louvan
University of Trento
Fondazione Bruno Kessler
slouvan@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

Neural-based models have achieved outstanding performance on slot filling and intent classification, when fairly large in-domain training data are available. However, as new domains are frequently added, creating sizeable data is expensive. We show that *lightweight augmentation*, a set of augmentation methods involving word span and sentence level operations, alleviates data scarcity problems. Our experiments on limited data settings show that lightweight augmentation yields significant performance improvement on slot filling on the ATIS and SNIPS datasets, and achieves competitive performance with respect to more complex, state-of-the-art, augmentation approaches. Furthermore, lightweight augmentation is also beneficial when combined with pre-trained LM-based models, as it improves BERT-based joint intent and slot filling models.

1 Introduction

In task-oriented dialogue systems, a spoken language understanding component is responsible for parsing the user utterance into a semantic representation. This is often modeled as a semantic frame (Tur and De Mori, 2011), and typically involves *slot filling* and *intent classification*. For example, in the utterance "*book in Southern Shores for 8 at Ariston Cafe*", the intent is BOOKING A RESTAURANT and the corresponding slot values and slot names are "*Southern Shores*" (CITY_NAME), "*8*" (NUMBER_OF_PEOPLE), and "*Ariston Cafe*" (RESTAURANT_NAME).

Although neural-based models (Qin et al., 2019; Goo et al., 2018; Mesnil et al., 2015) have achieved stellar performance in slot filling (SF) and intent classification (IC), their performance depend on the availability of large labeled datasets. Consequently, they suffer in *data scarcity* situations, which regularly happen when new domains are added to the system to support new functionalities.

One of the methods proposed to alleviate data scarcity is *data augmentation* (DA), which aims to automatically increase the size of the training data by applying data transformations, ranging from simple word substitution to sentence generation. Recently, DA has shown promising potential for several NLP tasks, including text classification (Wei and Zou, 2019; Wang and Yang, 2015), parsing (Sahin and Steedman, 2018; Vania et al., 2019), and machine translation (Fadaee et al., 2017). As for SF and IC, DA approaches typically generate synthetic utterances by leveraging Seq2Seq (Hou et al., 2018; Zhao et al., 2019; Kurata et al., 2016), Conditional VAE (Yoo et al., 2019), or pre-trained Natural Language Generation (NLG) models (Peng et al., 2020). Such approaches make use of in-domain data, and are relatively *heavyweight*, as they require training neural models, which may involve several phases to generate, filter, and rank the produced augmented data, thus requiring more computation time. It is also relatively challenging for deep learning-based models to generate semantically preserving synthetic utterances in limited data settings.

In this paper, we show that *lightweight augmentation*, a set of simple DA methods that produce utterance variations, is very effective for SF and IC

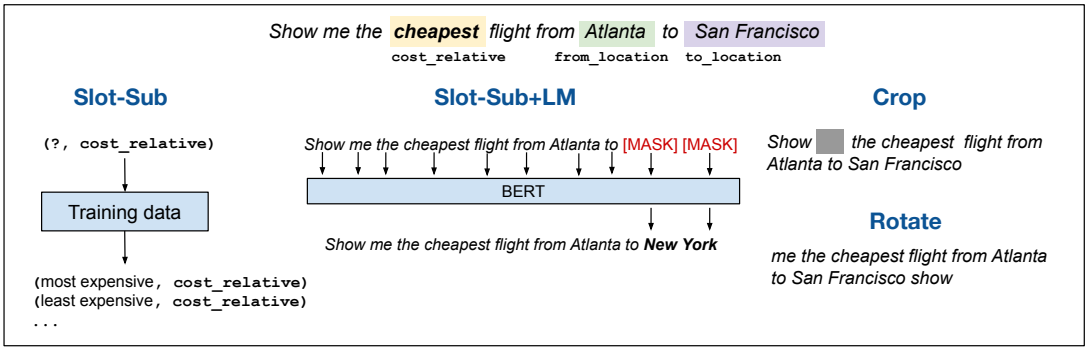


Figure 1: Examples of applying *lightweight augmentation* on an utterance in the ATIS dataset.

in a low-resource setting. Lightweight augmentation considers both *text span* and *sentence* variations. The span-level augmentation aims to diversify slot values in a particular text span through a *semantically preserving* substitution of slot values. The sentence-level augmentation seeks to produce alternative sentence structure through crop and rotate (Sahin and Steedman, 2018) operations based on a dependency parse structure.

We investigate the effect of lightweight augmentation both on typical biLSTM-based joint SF and IC models, and on large pre-trained LM transformers based models, in both cases with a limited data setting. Our contributions are as follows:

- We present a lightweight text span and sentence level augmentation for SF and IC. We show that, despite its simplicity, lightweight augmentation is competitive with more complex, deep learning-based, augmentation.
- We show that big self-supervised models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020) can perform well under a low data regime, and still benefit from lightweight augmentation.
- The combination of our span based augmentation and transfer learning (e.g. BERT fine-tuning) yields the best performance for most cases.

2 Lightweight Data Augmentation

Given the original training data $\mathcal{D}$, DA aims to generate additional training data $\mathcal{D}'$. For each sentence S in $\mathcal{D}$, an augmentation operation is applied N times, which can be empirically determined. Each augmented sentence S' is added to $\mathcal{D}'$, and the union of $\mathcal{D}$ and $\mathcal{D}'$ is then used to train the model for SF and IC. We describe the lightweight DA operations in the following subsections.

2.1 Slot Substitution (SLOT-SUB)

Our first lightweight method, slot substitution, is similar to Gulordava et al. (2018), which is based on substituting a token in a sentence with another token with a consistent syntactic annotation (i.e., part-of-speech or morphology tags). However, unlike Gulordava et al. (2018), our method is not limited to single tokens. As slot filling is a *semantic* task, rather than syntactic, we can naturally extend the method from single tokens (i.e., slot names composed by a single token) to multiple tokens (i.e., slot names composed by multiple tokens, or *spans*[1]), still preserving the semantics associated to a certain slot.

Practically, for slot substitution we take advantage of the fact that SF training data are typically annotated with the BIO format[2]. We exploit the fact that two text spans in different utterances in $\mathcal{D}$ are likely to be semantically similar if they share the same slot label. We randomly pick one span in the S and then perform the substitution (Figure 1 *Left*). For instance, we can substitute the span "*cheapest*", with other spans having the same slot label (i.e.,

[1]We define a span as a sequence of one or more tokens that convey a slot value.

[2]`B` indicates the beginning of the span, `I` indicates the inside of the span. `O` indicates that a token does not belong to any slot. For example, "San Francisco" will be annotated as `B-to_location I-to_location`.

COST_RELATIVE), such as "*least*" or "*most expensive*".

More formally, we denote a span sp in a sentence S as a slot-value pair $sp = (y, val)$, and we aim to produce an alternative pair $sp' = (y', val')$ such that the slot values are different ($val \neq val'$) and the slot labels are the same ($y = y'$) for both slot-value pairs. To obtain sp', we collect a set of candidates $\mathcal{SP}' = \{sp'_1, sp'_2, ..., sp'_n\}$, by looking for slot spans in other sentences in $\mathcal{D}$ that satisfy our criteria. After that, we randomly sample a span from $\mathcal{SP}'$ to obtain a $sp^{'}$. We replace sp in S with sp' to produce the new augmented sentence $S^{'}$. For example, in the utterance "*show me the cheapest flight from Atlanta to San Francisco*", one of the spans that can be substituted is $sp = (\text{COST\_RELATIVE}, \text{"}cheapest\text{"})$. Assuming that from $\mathcal{D}$ we can obtain $\mathcal{SP}^{aug} = \{(\text{COST\_RELATIVE}, \text{"}least\ expensive\text{"}), (\text{COST\_RELATIVE}, \text{"}most\ expensive\text{"}), \dots\}$, we then sample a sp' from $\mathcal{SP}'$ and replace sp in S with sp' to produce S'. Notice that the slot values in sp' are not necessary synonyms of the original slot value, although their slot label must be the same to preserve semantic compatibility.

2.2 Slot Substitution with Language Model (SLOT-SUB-LM)

Our second lightweight method, SLOT-SUB-LM, shares the goal with SLOT-SUB, i.e., to substitute sp with sp'. However, we do not use $\mathcal{D}$ to look for substitute candidates, instead we use a large pre-trained language model to generate the slot value candidates, using the *fill-in-the-blank* style (Donahue et al., 2020). The expectation is that large pre-trained LMs, being trained on massive amount of data, can produce a sensible text span given a particular sentence context, and possibly produce slot values that do not occur in $\mathcal{D}$. While we use BERT for our purpose, virtually any pre-trained LM can be used for SLOT-SUB-LM. Existing works on DA using LMs (Kobayashi, 2018; Kumar et al., 2020) are applied on text classification to replace random tokens in the text, which is not directly applicable to SF. Our approach focuses on *spans* conveying slot values, and include a filtering mechanism to reject retrieved slot spans that are not semantically compatible.

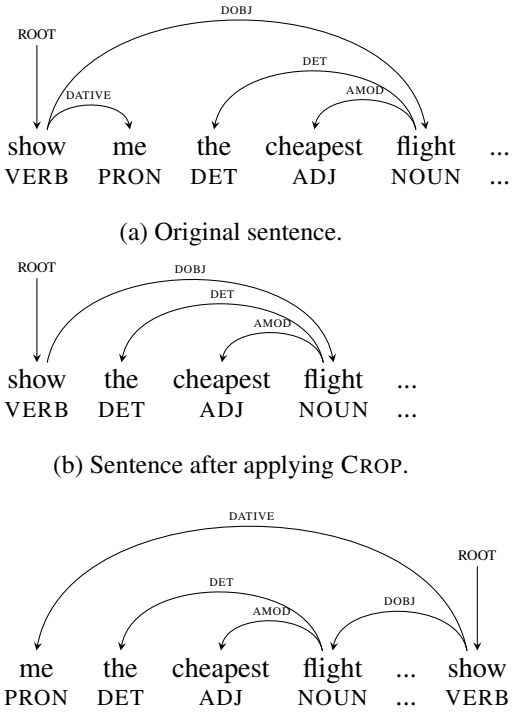
Generating New Slot Values. Given an utterance consisting of one or more slot value spans, we "blank" one of the span and then let the LM to predict the new tokens in the span. For instance, we give "*show me the ____ round trip flight from Atlanta to Denver*" to the LM for blank prediction. Practically, blank tokens are encoded as special `[MASK]` tokens[3] to let the pre-trained LM performing prediction. The decoding of the new tokens is carried out iteratively from left to right (Figure 1 *Middle*) and, to produce the surface form of a token, we apply nucleus sampling (Holtzman et al., 2020) using the top-*p* portion of the probability mass. Nucleus sampling has been empirically shown to be better than beam search, and top-*k* sampling (Fan et al., 2018) to produce fluent and diverse texts.

Filtering. While pre-trained LMs are expected to generate sensible replacements for a span in the utterance, a possible issue is that the new slot span is not semantically consistent with the original one. For example, for the original span "*cheapest*" in "*show me the cheapest round trip flight from Atlanta to Denver*", the LM could output "*earliest*" as a substitution, which does not fit the slot label COST_RELATIVE. To mitigate this issue, we use a binary sentence classifier as a *filter* (**SLOT-SUB-LM+Filter**) to decide whether S and S' are semantically compatible, based on the change made on the slot span. The training of the classifier is composed of a pair S and S', with its binary decision label (i.e., accept or reject S'). To construct the training data, for positive examples (*accept*) we take advantage of the sentence pair produced by SLOT-SUB, while for the negative examples (*reject*) we sample sp' in $\mathcal{D}$ where $y \neq y'$ and replace sp in S with sp' to produce S'. We use the BERT model as the sentence pair classifier and we encode the tokens, w, in both S and S' sentence pairs, as `[CLS]` $w_1^S w_2^S \dots w_n^S$ `[SEP]` $w_1^{S'} w_2^{S'} \dots w_m^{S'}$. On top of BERT, we add a feed-forward layer that uses the hidden state of the sentence representation, $h_{\texttt{[CLS]}}$, for prediction.

[3] We set the number of masked tokens to be the same as the tokens of the original slot value, e.g. `san francisco` is masked as `[MASK][MASK]`, although this number could actually be sampled as well.

2.3 CROP and ROTATE

The third lightweight method that we present augments an utterance by changing its syntactic structure. We adopt the augmentation approach from Sahin and Steedman (2018) (Figure 1 *Right*), which is based on two operations, CROPand *rotate*, applied to the dependency parse tree of a sentence. To our knowledge, this approach has not yet been applied to slot filling and intent classification, which is a contribution of our work. *Crop* focuses on particular fragments of a sentence (e.g., predicate and its subject, or predicate and its object), and removes the rest of the fragments, including its sub-tree, to create a smaller sentence. *Rotate* aims to rotate the target fragment of a sentence around the root of the dependency parse structure, producing a new utterance. For example, in the utterance "*show me the cheapest flight from Atlanta to San Francisco*", the word *"me"* can be cropped as it is one of the children of the *root* verb "*show*". While for rotation, the direct object ("*flight*") and its children ("*the cheapest*") are rotated around the root verb. Figure 2 illustrates the relevant dependency structure manipulation.



(a) Original sentence.

(b) Sentence after applying CROP.

(c) Sentence after applying ROTATE.

Figure 2: Examples of dependency tree operations on a sentence.

3 Experiments and Results

We experimented our lightweight augmentation approach on three well-known datasets for SF and IC, namely ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018) and FB (Schuster et al., 2018). All datasets are in English. ATIS contains utterances related to flight domain (e.g., searching flight, booking). SNIPS includes multi-domain utterances such as weather, movie, restaurant, etc. FB contains utterances from 3 domains, weather, alarm, and reminder. To simulate the *data scarcity* setting, we follow previous works (Hou et al., 2018; Yoo et al., 2019) and only use *medium*-size (i.e., 1/10) of training data for each dataset. Statistics on the three datasets are reported in Table 1.

As for evaluation, we use standard evaluation metrics, namely the F1-score for SF and accuracy for IC[4]. Performance are calculated as the average score of ten different runs. In order to compare our methods, we use two baselines for slot filling and intent detection: a simple BiLSTM-CRF model, and a state of the art BERT-based model, which is fine-tuned to SF and IC[5]. Each model is trained for 30 epochs, and we apply early stopping criteria.

For both slot substitution (SLOT-SUB) and slot substitution with language model (SLOT-SUB-LM) augmentation methods, we tune the number of augmented sentences per utterance, N, on the dev set of each dataset. For crop and rotate, we use the default parameters from Sahin and Steedman (2018). To produce the dependency parse structure for the utterances in our datasets, we use Spacy[6]. All hyperparameters are tuned on the dev set. More details on the settings is provided in Appendix A. For training the binary classifier for SLOT-SUB-LM+Filter, we generate the same number of positive (*accept*) and negative (*reject*) training instances[7].

In order to allow comparison with more complex data augmentation approaches, we also report

[4]Metric is computed using the standard evaluation script https://www.clips.uantwerpen.be/conll2000/

[5]We use the `bert-base-uncased` model

[6]https://spacy.io/

[7]Details in Appendix B

	Label		#Utterances ($\mathcal{D}$)			#Augmented Training Utterances ($\mathcal{D}'$)			
Dataset	**#slot**	**#intent**	**#train**	**#dev**	**#test**	**SLOT-SUB**	**SLOT-SUB-LM**	**CROP**	**ROTATE**
ATIS	79	18	0.4K	500	893	3.9K	0.8K	0.8K	1.1K
SNIPS	39	7	1.3K	700	700	6.3K	2.5K	2.6K	3.7K
FB	16	12	3K	4.1K	8.6K	5.4K	5.4K	5.9K	8.5K

Table 1: Statistics of both the original training data $\mathcal{D}$ and the augmented data $\mathcal{D}'$. #train denotes our medium-size training data setup (10% of full training data). $\mathcal{D}'$ is produced by each augmentation method, where the number N of augmentations per sentence is tuned on the dev set.

results obtained with state of the art approaches based on Seq2Seq (Hou et al., 2018) and Conditional Variational Auto Encoder (CVAE) (Yoo et al., 2019). Our implementation is based on the Huggingface library (Wolf et al., 2019), available at `https://github.com/slouvan/saug`.

3.1 Results

Table 2 reports the results on the test sets used in our experiments. As for comparison, we include best-reported scores from two state of the art augmentation methods, namely a sequence-to-sequence (Seq2Seq) based on Hou et al. (2018) and a VAE based methods from Yoo et al. (2019). Results in Table 2 (*test set*) show that lightweight augmentation is beneficial for both Bi-LSTM CRF and BERT, on both ATIS (single domain) and SNIPS (multi-domain) datasets. SLOT-SUB yields the best results for both the BiLSTM+CRF and BERT models, with SF performance up to 90.43 on ATIS and 90.66 on SNIPS, and IC performance to 95.49 on ATIS and 97.11 on SNIPS. As for the FB dataset, models only gain marginal improvement across lightweight augmentation. We hypothesize that FB is relatively easy to solve, compared with ATIS and SNIPS, as the slot filling performance of BiLSTM without augmentation already achieves a very high F1 score. The improvement using augmentation is more significant for SF rather than for IC.

Out of all lightweight augmentation methods, SLOT-SUB obtains the best performance, particularly on slot filling on ATIS and SNIPS. The overall best performing configuration is a combination of BERT fine-tuning with SLOT-SUB augmentation. Given limited training data, BERT fine-tuning without augmentation surpasses BiLSTM-CRF without augmentation by a large margin. Yet, performance can be boosted even further with lightweight augmentation, suggesting that even a big, self-supervised model, such as BERT can still benefit from augmentation on limited data settings. The improvements on BiLSTM-CRF indicate that lightweight augmentation improves the model's robustness when trained on small amounts of data. We find that SLOT-SUB-LM is suboptimal for SF. Our qualitative observation shows that SLOT-SUB-LM often generates slot values that are semantically incompatible with the original slot label. CROP and ROTATE can help IC in some cases, although their improvement is marginal.

Despite its simplicity, SLOT-SUB is also competitive with state-of-the-art heavyweight data augmentation approaches (Seq2Seq and CVAE), significantly boosting Bi-LSTM and BERT performance for SF on ATIS and SNIPS. We believe that the key advantage of SLOT-SUB is its capability to maintain semantic consistency over the slot spans, which has revealed to be stronger than that of heavyweight approaches. This also shows that slot consistency is crucial for obtaining good performance, particularly for SF. While the CVAE based approach from Yoo et al. (2019) has injected slot and intent labels in the model, it seems that generating semantically consistent utterances is still challenging for deep learning models, especially when data is limited.

4 Analysis and Discussion

In this Section we discuss several aspects of data augmentation applied to slot filling and intent detection.

Impact of number of augmented sentences. To better understand the effect of the number of augmented sentences per utterance (N), we now

Model	DA	ATIS		SNIPS		FB	
		Slot	Intent	Slot	Intent	Slot	Intent
BiLSTM+CRF	None	86.83	90.64	84.51	95.94	93.83	98.47
	Seq2Seq (Hou et al., 2018)	88.72	-	-	-	-	-
	VAE (Yoo et al., 2019)	89.27	90.95	-	-	-	-
	SLOT-SUB	89.89†	93.37†	86.45†	96.30†	93.70	98.45
	SLOT-SUB-LM	87.03	92.96†	82.82	96.14	91.52	98.20
	SLOT-SUB-LM+Filter	87.19	92.01†	82.77	96.08	92.18	98.37
	CROP	88.62†	92.32†	85.84†	96.07	93.91	98.64
	ROTATE	88.83†	92.33†	85.65	96.39†	94.04	98.56
BERT	None	89.39	94.98	89.17	96.70	94.22	98.61
	SLOT-SUB	**90.43**†	**95.49**†	**90.66**†	**97.11**†	94.01	98.59
	SLOT-SUB-LM	87.88	94.49	85.65	96.59	91.84	98.47
	SLOT-SUB-LM+Filter	88.37	94.57	86.23	96.60	92.60	98.59
	CROP	89.47	94.55	89.77	96.78	94.20	98.73
	ROTATE	89.57	94.48	89.37	96.81	**94.32**	**98.80**

Table 2: Overall results on the test set. Underlined numbers indicate best performing methods for a particular slot filling + intent model. **Bold** numbers indicate best overall methods. † indicates significant improvement over the baseline without augmentation (p-value < 0.05, Wilcoxon signed rank test).

observe the performance of our best performing method, SLOT-SUB, while changing N values (we use $\{2, 5, 10, 20, 25\}$) on the dev set. As for ATIS, increasing N yields a F1 improvement from 90.68 up to 91.62; SNIPS performance increased from 87 F1 and to 88 F1 when increasing N from 2 to 5 and it is stable around 88 F1 when using N larger than 5; finally, FB is stable around 93.4 to 93.7 F1. Overall, the biggest improvement is when N is increased from 2 to 5, while with higher values only minor improvements can still be obtained on ATIS.

Performance on different training data size ($\mathcal{D}$). Figure 3 displays the gain obtained by SLOT-SUB for various data size for slot filling. Using smaller data size (i.e., 5%) than our default setting, SLOT-SUB still obtains a F1 gain for all datasets. On the other hand, as we increase the number of training data, the SLOT-SUB benefit diminishes, without hurting performance on ATIS and SNIPS. As for FB we observe a performance drop of less than 1 F1, which is still relatively low.

Is lightweight augmentation beneficial to very large language models? Motivated by the encouraging results that lightweight augmentation has obtained on a strong pre-trained LM such as BERT on low-resource settings (see Table 2), we now further examine the advantage of lightweight augmentation for other very large pre-trained LM models, namely Albert (Lan et al., 2020) and Roberta (Liu et al., 2019). We use the largest trained models for each of the

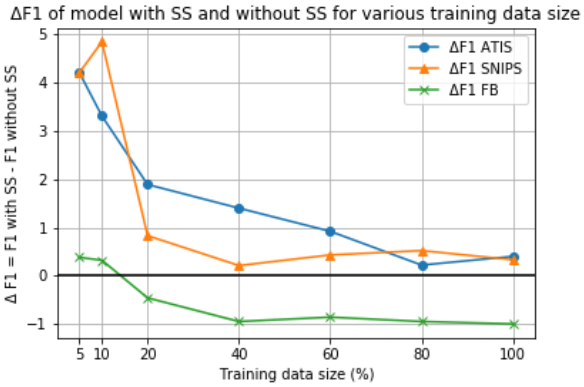


Figure 3: Gain ($\Delta F1$) obtained by SLOT-SUB (SS) on various training data size. Positive numbers mean that the model with SS is better than without SS.

Model	Aug.	ATIS		SNIPS	
		Slot	Intent	Slot	Intent
BERT (large)	None	91.6	95.0	89.8	95.0
	SS	92.8	95.4	92.8	95.4
Albert (xxl)	None	92.1	94.8	89.5	99.0
	SS	92.9	95.0	**93.6**	**99.2**
Roberta (large)	None	90.6	92.8	89.2	98.9
	SS	**93.2**	**95.9**	92.5	98.8

Table 3: Lightweight augmentation SLOT-SUB (SS) applied to very large pre-trained LMs.

pre-trained LM, namely `bert-large-uncased`, `roberta-large`, and `albert-xxl`. Results, reported in Table 3, show that on limited data settings, all the very large models still benefit from SLOT-SUB, notably on the performance for SF.

Qualitative Analysis of slot values from SLOT-SUB vs SLOT-SUB-LM. The performance of SLOT-SUB especially in SF is better than SLOT-SUB-LM, as SLOT-SUB maintains semantic consistency on the span level. We observe that SLOT-SUB-LM often generates slot values that fit the sentence context but that do not maintain the semantics of the slots, which hampers the performance in SF (Table 4). The fact that SLOT-SUB-LM often generates "wrong" slot values makes SLOT-SUB-LM+Filter also less effective. A possible future direction is to cast SLOT-SUB-LM as a *conditional* NLG problem, incorporating labels at the token-level, although this is still challenging when data is limited.

5 Related Work

Data augmentation methods have been widely applied in computer vision, ranging from geometric transformations (Krizhevsky et al., 2012; Zhong et al., 2020), data mixing (Summers and Dinneen, 2019), to the use of generative models (Goodfellow et al., 2014) for generating synthetic data. Recently, data augmentation has been applied to various NLP tasks, including text classification (Wei and Zou, 2019; Wang and Yang, 2015), parsing (Sahin and Steedman, 2018; Vania et al., 2019), and machine translation (Fadaee et al., 2017). Augmentation techniques for NLP tasks range from operations on tokens (e.g., substituting, deleting) (Wang and Yang, 2015; Kobayashi, 2018; Wei and Zou, 2019), to manipulation of the sentence structure (Sahin and Steedman, 2018), to paraphrase-based augmentation (Callison-Burch et al., 2006).

Data augmentation has been also experimented in the context of slot filling and intent classification. Particularly, recent methods have focused on the application of generative models to produce synthetic utterances. Hou et al. (2018) proposes a method that separates the utterance generation from the slot values realization. A sequence to sequence based model is used to generate utterances for a given intent with slot values placeholders (i.e., delexicalized), and then words in the training data that occur in similar contexts of the placeholder are inserted as the slot values. Zhao et al. (2019) also uses a sequence to sequence model by exploiting a small number of template exemplars. Yoo et al. (2019) proposes a solution based on Conditional Variational Auto Encoder (CVAE) to generate synthetic utterances. In this case the CVAE takes into account both the intent and the slot labels during training, and the model generates the surface form of the utterance, slot labels, and the intent label. Recent work from Peng et al. (2020) make use of GPT-2 (Radford et al., 2019), and fine-tuned it to intent and slot-value pairs to generate utterances.

In comparison to existing, state of the art, augmentation methods for slot filling and intent detection, the augmentation methods proposed in this paper can be considered as *lightweight* because they do not require any separate training based on deep learning models for generating additional data. Still, lightweight augmentation maintains consistent slot semantic substitutions, a feature that is crucial for effective data augmentation. In the spectrum of existing augmentation methods, i.e., from words manipulation to paraphrasing-based methods, our lightweight approaches lie in the middle, as we focus either on particular *text spans* that convey slot values, or on particular structures in the dependency parse tree of the utterance.

6 Conclusion

We showed that lightweight augmentation for slot filling and intent detection in low-resource settings is very competitive with respect to more com-

Dataset	Slot	Original Sentence	SLOT-SUB	SLOT-SUB-LM
ATIS	DEPART_TIME	List all flights leaving Denver on Continental on Sunday after **934 pm**	List all flights leaving Denver on Continental on Sunday after **7 pm**	List all flights leaving Denver on Continental on Sunday after **Christmas day**
	FROMLOC_CITYNAME	List all flights leaving **Denver** on Continental on Sunday after 934 pm	List all flights leaving **Atlanta** on Continental on Sunday after 934 pm	List all flights leaving **Boston** on Continental on Sunday after 934 pm
	AIRLINES_NAME	I need a flight on **Air Canada** from Toronto to San Diego with a layover in DC	I need a flight on **Northwest Airlines** from Toronto to San diego with a layover in DC	I need a flight on **a Thursday** from Toronto to San Diego with a layover in DC
SNIPS	CONDITION_DESCRIPTION	Will it be **sunny** in Eyota Hawaii on February seventh 2025	Will it be **humid** in Eyota Hawaii on February seventh 2025	Will it be **held** in Eyota Hawaii on February seventh 2025
	SPATIAL_RELATION	What is the **closest** cinema today playing animated movies	What is the **close-by** cinema today playing animated movies	What is the **underground** cinema today playing animated movies
	RESTAURANT_TYPE	I need to book a **pub** in Cammack village Wyoming for a party of seven	I need to book a **fast food restaurant** in Cammack village Wyoming for a party of seven	i need to book a **lodge** in Cammack village Wyoming for a party of seven
FB	DATE_TIME	Set alarm **for 4 am** tomorrow morning	Set alarm **at 6** tomorrow morning	Set alarm **for me** tomorrow morning
	LOCATION	How hot is it in **Hong Kong** ?	How hot is it in **Fairbanks** ?	How hot is it in **the mornings** ?

Table 4: Samples of sentences from SLOT-SUB and SLOT-SUB-LM. The bold text span denotes the span that is substituted. The text span in blue denotes semantically consistent replacements, while red indicates semantically inconsistent substitutes.

plex deep learning based data augmentation. A lightweight method based on slot values substitution, while preserving the semantic consistency of slot labels, has proven to be the more effective. We also show that large self-supervised models, like BERT, can benefit from lightweight augmentation, suggesting that a *combination* of data augmentation and transfer learning is very useful, and has the potential to be applied to other NLP tasks.

As for future work, it would be interesting to see the effect of using the augmented data generated by SLOT-SUB as additional training data for deep learning based augmentation models. Encouraged by the results of our lightweight augmentation, our work can also be experimented on semantic tasks with similar characteristics, such as Named Entity Recognition.

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. 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 2492–2501. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *HLT*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1234–1245. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2077–2083. The Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised

learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:530–539.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models. *CoRR*, abs/2004.13952.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. 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 2078–2087. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gözde Gül Sahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. 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 5004–5009. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL-HLT*.

Cecilia Summers and Michael J Dinneen. 2019. Improved mixed-example data augmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1262–1270. IEEE.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116. Association for Computational Linguistics.

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 Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2557–2563. The Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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 6381–6387. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7402–7409. AAAI Press.

Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. 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 3635–3641. Association for Computational Linguistics.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13001–13008. AAAI Press.

Appendix A. Hyperparameters

Hyperparameter	Value
Learning rate	10^{-5}
Dropout	0.1
Mini-batch size	16
Optimizer	BertAdam
Number of epoch	30 (`bert-base-uncased`) 10 (`bert-large`, `roberta-large`, `albert-xxl`)
Early stopping	10
nb_{aug}	Tuned on {2, 5, 10}
Nucleus sampling	top-p = 0.9
Max rotation	3
Max crop	3

Table 5: Hyperparameters used for the Transformer based models and data augmentation methods

Appendix B. Training Data for SLOT-SUB-LM+Filter

Dataset	#train
ATIS	7,846
SNIPS	24,472
FB	52,798

Table 6: Total training examples for SLOT-SUB-LM+Filter. The number of positive and negative examples are the same.

Dialogue policy optimization for low resource setting using Self-play and Reward based Sampling

Tharindu Madusanka, Durashi Langappuli, Thisara Welmilla,
Uthayasanker Thayasivam and Sanath Jayasena
Department of Computer Science and Engineering,University of Moratuwa ,Sri Lanka
{stharindu.16, durashi.16, welmilla.16, rtuthaya, sanath}@cse.mrt.ac.lk

Abstract

Reinforcement Learning is considered as the state of the art approach for dialogue policy optimization in task-oriented dialogue systems. However, these models demand a large corpus of dialogues to learn effectively. Training Reinforcement Learning agent with low data amount tends to overfit the agent. Although synthesizing dialogue agendas with dialogue Self-play using rule-based agents and crowdsourcing has demonstrated promising results with the low amount of samples, these methods hold limitations. For instance, rule-based agents acquire specific domain and language while crowdsourcing demands a high price and domain experts, especially in local languages. In this paper, we address these limitations by proposing a novel approach for synthetic agenda generation by acknowledging the underlying probability distribution of the user agendas and a reward-based sampling method that prioritizes failed dialogue acts. Evaluations conducted shows leveraged performance without overfitting, compared to the baseline method. Also, the reward-based sampling method improves the overall mean task success rate by an average of 11.307%.

1 Introduction

A dialogue system, or conversational agent, denotes a system that can conduct a conversation with another agent, usually a human (Perez-Marin and Pascual-Nieto, 2011). One type of conversational agent are Task-oriented conversational agents, that can help users accomplish tasks ranging from meeting scheduling to vacation planning. The structure of a task-oriented conversational agent is outlined in Figure 1. It consists of (i) a natural language understanding(NLU) module for identifying intents of user utterances (ii) a dialogue state tracker(DST) for tracking conversation state (iii) a dialogue policy learner(POL) which selects the next action based on the current state (iv) a natural language generator(NLG) for converting the agent action to a natural language response (Gao et al., 2018).

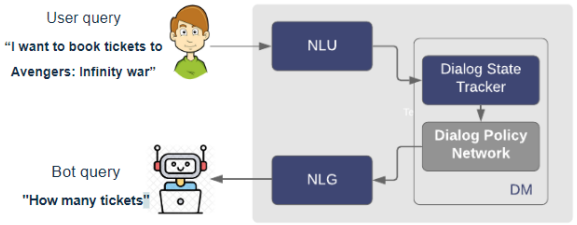


Figure 1: Components of a task-oriented conversational agent

Dialogue Policy Network(learner) plays a critical role in task-oriented conversational agents since they require logical reasoning and planning over several dialogue turns. Policy Network has been developed under rule-based methods(i.e. ontology-based, finite state machines) and model-based methods(i.e. Supervised Learning(SL) based (Bordes et al., 2016; Dinan et al., 2018), Reinforcement Learning(RL) based (Lu et al., 2019; Liu and Lane, 2018; Su et al., 2016a)). Due to enhanced data availability, dialogue policy optimization has leveraged with RL based methods. However, these state-of-the-art RL methods have barely experimented in the low resource

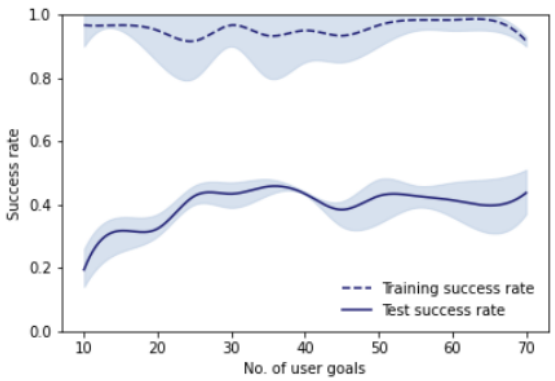


Figure 2: Training and test success rates of user goals with low amount of samples

setting because these models tend to overfit when the number of available training data is low(see Figure 2).

The accuracy of the dialogue policy network directly depends on the availability of quality dialogue samples to train. As depicts in Figure 2, the agent tends to overfit when the number of training samples is low. The dialogue Self-play approach has proposed to overcome this issue with (i) crowdsourcing(Shah et al., 2018a) or manually synthesize agendas and (ii) rule-based agents to synthesize agendas (Shah et al., 2018b). However, these methods have limitations since crowdsourcing is expensive especially considering low resource local languages, and rule-based agents are limited to a specific domain and language. Thus an alternative approach to address this issue is to synthesize agendas considering all the possibilities. But this method conduces to create unrealistic agendas besides awarding equal probability to every possible state that affects the speed of convergence of Reinforcement Learning agents.

To address these limitations, we propose a novel approach for synthetic agenda generation by acknowledging the underlying probability distribution of the user agendas. Since this methodology applies to a low amount of samples, this method can lead to an insufficient exploration of agendas by the agent. Therefore we further developed the methodology by introducing a selective sampling method based on the reward function that prioritizes the failed dialogue acts, where the agent actively decides what agendas to use. The intuition is that the agent can learn more from failed dialogues over successful ones.

The rest of the paper is organized as follows: Section 2 describes the Background on RL and data synthesis. Section 3 presents the related work for dialogue policy optimization using RL and low resource settings. Our methodology is fully described in section 4. We conduct our experiments and show the results in section 5. Finally, we conclude our work in section 6.

2 Background

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a learning paradigm where an intelligent agent learns to make optimal decisions by interacting with an initially unknown environment(Sutton and Barto, 2018). The agent interacts with the environment by observing the state s_t and taking an action a_t. Depend on the action the agent receives a reward r_{t+1} and observes the state change s_{t+1}. This continues until the episode ends. The agent's goal is to maximize the cumulative reward at each step and the cumulative reward at step t is denoted by G_t

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$$
$$= \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}$$

where the γ is the discount factor. The action an agent takes at state s defined by the policy the agent follows. The policy is a function that maps states to actions and denoted by π. The agent's goal is to find the optimal policy denoted by π^*.

There are mainly three types of methods for finding the optimal policy. They are the value-based methods, the policy-based methods, and the actor-critic methods. In value-based methods, a value function is used to express the value of the state or state-action pair. Note the value function define with respect to the policy and the optimal value function denoted by either by V^* or Q^*. So if the optimal value function is known the optimal policy can be found by,

$$\pi^*(s) = argmax_a(Q^*(s, a))$$

The Monte-Carlo method, Q-learning, and DQN (Mnih et al., 2013) which use deep neural networks are popular value-based methods.

The policy-based methods find the optimal policy directly without using any value function. The well-known policy-based methods are Reinforce (Williams, 1992) and Proximal policy optimization (PPO) (Schulman et al., 2017). The actor-critic methods(Konda and Tsitsiklis, 2000) combine the two methods. Here, there is an actor which acts according to certain policy and a critic which tries to estimate the value function for a given state.

The dialogue policy learning is a policy optimization problem, and many researchers have used Reinforcement Learning to find the optimal policy (Pineau and Thrun, 2004; Gasic and Young, 2014; Williams and Young, 2007). Here, the policy network is the agent, and the environment is the user or the user simulator. The agent keeps track of the state of the dialogue, and the state is defined using intents, entities, and slot values. The agent takes action depending on the state. This action can be generating a dialogue act or sending an API call etc. For the action taken by the agent, a reward is given by the user simulator or user at the end of dialogue or each step. Usually, in task-oriented dialogue systems, the reward is associated with the task-success rate. The task-success rate is the measure of how much of the user's task is achieved or not.

2.2 Data synthesization

Machine learning specifically in the context of Supervised Learning can be defined as an approach that tries to get $min_\theta \sum L(\theta, D)$, where D is the data and θ is the parameters of the model. Usually, when models are trained, in each episode of training a random sample(batch), D_s is drawn from the data D used for updating model parameters. However if enough data is not available, we may need to generate data or use a different type of learning method like Meta-learning. If data is generated then this generated data, S is used for training either with D or instead of D. S can be created using a Generator function, G or manually.

$$S = G(D) \text{ and } G \sim P(D)$$

The usual method of sampling from synthetic data, S is random sampling,

$$D_s \sim random(S)$$

3 Related work

The Reinforcement Learning approaches are more suitable for modeling the policy network of the conversational agent and it has already shown very promising results in dialogue policy optimization research(Larsson and Traum, 2000; Li et al., 2009; Lucie et al., 2019; Gasic and Young, 2014). It is capable of exploring a large action state space and approximating the optimal policy. However, the Reinforcement Learning methods are rarely used for the low resource setting as these approaches require more samples with respect to the rule-based and Supervised Learning approaches.

Most of the research works in the area of using Reinforcement Learning for dialogue policy learning are mostly focused on optimizing the reward function(Su et al., 2016a; Liu and Lane, 2018), improving task success rate(Li et al., 2009), achieving personalization(Mo et al., 2018; Hengst et al., 2019) and making system end-to-end (Dhingra et al., 2016; Wen et al., 2016) or interactive(Shah et al., 2016; Liu et al., 2018).

Some research work has been done on making the Reinforcement Learning approaches more sample efficient. One such approach is using warm-start such as the imitation learning where the agent mimics an expert provided policy(Li et al., 2015), Replay Buffer Spiking(RBS)(Lipton et al., 2018) which is used with DQN and use expert generated dialogues(Henderson et al., 2008; Chen et al., 2017). Another approach for improving the RL-based approach is using an efficient method for exploration(Lipton et al., 2018). Even though this research work has made the novel Reinforcement Learning approaches more sample efficient, they still cannot be used in low resource settings as they still require large amount of training dialogues to reach a sufficient level of accuracy.

Several dialogue Self-play approaches have proposed to synthesize agendas with (i) crowdsourcing(Shah et al., 2018a) or manually synthesizing (ii) rule-based agents for synthesizing (Shah et al., 2018b) (iii) synthesize agendas considering all possibilities. However, these methods have limitations. Crowdsourcing is an expensive process and may lead to additional practical problems especially in the local language setting and rule-based agents are limited to a specific language or domain. Although these is-

sues are addressed by the last method, it generates unrealistic agendas and can affect the rate of convergence.

Our goal is to solve the overfitting problem that occurs when the dialogue policy network is implemented with the Reinforcement Learning technique in a low resource setting. Thus, we propose a combination of two techniques. Addressing current issues in the Self-play approaches for agenda synthesis in the conversational domain, we proposed a Self-play mechanism that uses underlying probability distributions of dialogue acts to synthesize new agendas. However, this proposed method may cause insufficient exploration of agendas by the agent, due low amount of data. To address this issue, we introduce a selective sampling method that prioritizes failed dialogue acts based on the reward function.

4 Proposed Methodology

Proposed methodology has two main components. The first one is the Self-play mechanism that calculate underlying probability distributions from the training data and feed dialogue acts sample from these probability distributions to the user simulator. The second component is responsible for reward based sampling technique that prioritizes failed dialogues over successful ones.

4.1 Definitions

The objective of the methodology is to synthesize data that reflect the true distribution of the data and then sample and train the agent in an efficient manner. Let training data(Agendas) is denoted by D and $P(D)$ denotes the data distribution. Let there be N number of slots and s_q denotes the q^{th} slot ($1 \leq q \leq N$). To capture the $P(D)$ we consider 3 independent discrete probability distributions.

1. $P(s_q)$ - Probability of slot s_q exists in an Agenda

2. $P(req|s_q)$ - Probability that slot s_q being requestable slot given that it exists($P(req|s_q)$ = 1 - $P(inform|s_q)$ where $P(inform|s_q)$ denotes the probability that slot s_q being informable slot)

3. $P(kb|inform, s_q)$ - Probability that the informable slot value available in the knowledge base given that slot exists in the agenda and it is informable.

We use these probabilities to capture $P(D)$. The methodology we describe and the mathematical equations are applied to all 3 probability distributions mention above. So for generalization we'll denote any of the above probability distributions with the term P. Let there are m elements in the probability distribution P. Let x_k denotes the k^{th} element in P ($1 \leq k \leq m$). $p(x_k)$ denotes the probability of x_k

4.2 Self-play Mechanism

An overview of the proposed methodology is illustrated in Figure 3. Instead of directly using the training dataset for training which leads to overfitting the model, we use the training dataset for calculating the underlying probability distribution and use that probability distribution for sampling agendas to optimize the policy network.

We add a small noise value ϵ_k for each element x_k.

$$\epsilon_k \sim \mathcal{N}(0,\, \delta^2) \tag{1}$$

where δ is chosen as a very small number. We add the noise for two purposes. One is to promote exploration by making $\forall x_k, p(x_k) > 0$, and the other is to avoid overfitting. We add that noise in a way such that,

$$0 < p(x_k) < 1 \quad \text{and} \quad \sum_{k=1}^{m} p(x_k) = 1$$

Once the probability distributions are calculated and noise is added, we use the mechanism that we created to sample dialogue acts from the probability distribution and feed it to the user simulator. This way agents can be trained without overfitting to the training dataset.

4.3 Prioritizing Failed Dialogues (Reward Based Sampling Technique)

We use a selective sampling technique that prioritizes failed dialogues over successful ones. The intuition behind the concept is that we believe the agent can learn more from the failed dialogues than successful ones. We use the reward as the indicator of the dialogues being successful or failed. We assume that the reward is negative for failed dialogues while it is positive when dialogues are successful. So the method we proposed uses the reward function and the

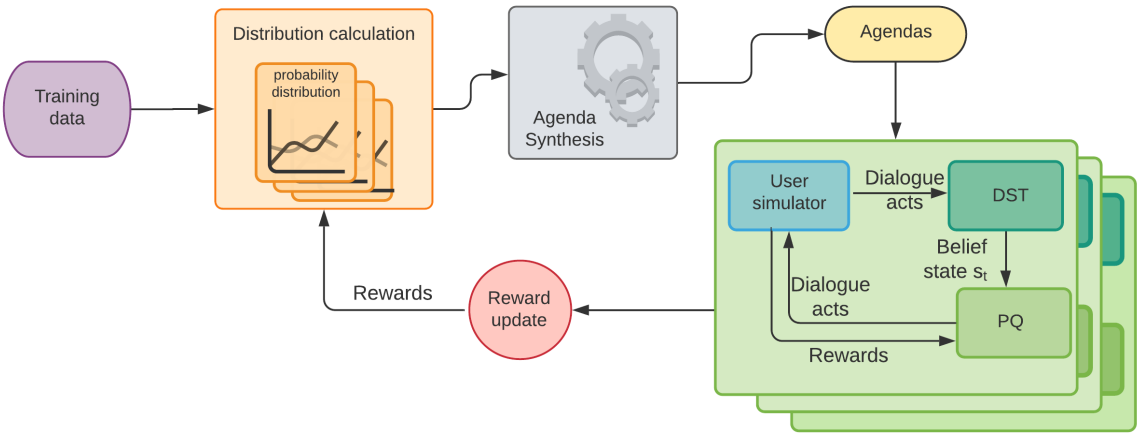


Figure 3: Our proposed Self-play framework, Distribution calculation and Agenda synthesis components responsible for synthesizing agendas, which then feed into the simulator. Reward update feed the reward signal for re-calculating distributions.

prior probabilities to recalculate the probabilities for sampling such that failed dialogues are prioritized.

So before any recalculation,

$$\sum_{k=1}^{m} p(x_k) = 1 \tag{2}$$

Let's consider element x_k. Let there be n_k agendas with element x_k in them and for each $i \leq n_k$, r_i^k denotes the reward. Since we want to prioritize the failed dialogues over successful ones we consider the negation. Also, we add the maximum possible reward to it to make sure that the resultant value is non-negative. Let $max(r)$ denotes the maximum possible reward while $min(r)$ denotes the minimum possible reward. Then we can calculate,

$$z_i^k = -r_i^k + max(r) \tag{3}$$

So that $\forall i, k, z_i^k \geq 0$. We can calculate the mean reward related to the element x_k by,

$$\begin{aligned} \frac{1}{n_k}\sum_{i=1}^{n_k} z_i^k &= \frac{-1}{n_k}\sum_{i=1}^{n_k} r_i^k + max(r) \\ &= w_k \end{aligned} \tag{4}$$

So w_k denotes the mean negative reward related to the element x_k. Then we can recalculate the probabilities from by considering the mean reward and previous probabilities.

$$f(x_k) = p(x_k)(\frac{w_k}{max(r)})^{\alpha} \tag{5}$$

The α is a hyperparameter which controls the influence of the reward when recalculating probability distributions. We divide by $max(r)$ for normalization. Then the resultant term $\frac{w_k}{max(r)}$ get normalized to a value between 0 and $(1 + \frac{|min(r)|}{max(r)})$. Then we can recalculate the probabilities by,

$$\begin{aligned} p(x_k) &= \frac{(f(x_k))^{\beta}}{\sum_{j=1}^{m}(f(x_j))^{\beta}} \\ &= \frac{(p(x_k)(\frac{\frac{-1}{n_k}\sum_{i=1}^{n_k} r_i^k}{max(r)} + 1)^{\alpha})^{\beta}}{\sum_{j=1}^{m}(p(x_j)(\frac{\frac{-1}{n_j}\sum_{i=1}^{n_j} r_i^j}{max(r)} + 1)^{\alpha})^{\beta}} \end{aligned} \tag{6}$$

$$\begin{aligned} p(x_k) &= \frac{exp\{(f(x_k)\beta\}}{\sum_{j=1}^{m} exp\{f(x_j)\beta\}} \\ &= \frac{exp\{(p(x_k)(\frac{\frac{-1}{n_k}\sum_{i=1}^{n_k} r_i^k}{max(r)} + 1)^{\alpha})^{\beta}\}}{\sum_{j=1}^{m} exp\{(p(x_j)(\frac{\frac{-1}{n_j}\sum_{i=1}^{n_j} r_i^j}{max(r)} + 1)^{\alpha})^{\beta}\}} \end{aligned} \tag{7}$$

Equation 6 and 7 normalize the $f(x_k)$ so that the $\sum_{j=1}^{m} p(x_k) = 1$. Equation 6 uses a division by the sum(naive normalization), while equation 7 uses the Softmax equation for normalization. In both cases, the β is a hyperparameter. So, we can use either equation 6 or 7 for recalculating probability distributions such that failed dialogues are prioritized (we can use equation 6 and 7 for reward-based sampling).

Algorithm 1 explain the flow of the process from getting the dataset to the continuous recalculation of probability distributions.

Algorithm 1: Self-play with reward based Sampling

Require: training, D
Require: hyperparametrs; α,β, and maximum reward; max(r)
calculate probability distributions from the D
add noise $\epsilon_k \sim \mathcal{N}(0, \delta^2)$ *for each* x_k *in each of the probability distributions* P
initialize reward buffer; R = {}
for *each training episode t* **do**
  sample agenda, A_t *from probability distribution*
  interact and train the agent
  store agenda A_t *and reward* r_t *in reward buffer* $R \cup \{A_t, r_t\}$
  if *recalculate probability* **then**
    recalculate probability distribution using either equation 6 or 7
  end if
end for

4.4 The Influence of α in Prioritizing Failed Dialogues

Consider the mean negative reward related to the element x_k, w_k and normalized mean negative reward $\frac{w_k}{max(r)}$, where max(r) is the maximum possible reward.

$$0 \leq \frac{w_k}{max(r)} \leq 1 + \frac{|min(r)|}{max(r)} \tag{8}$$

Note that if the more dialogues are successful then w_k is less than $max(r)$, while if more dialogues failed then w_k is greater than $max(r)$. Let w_i denotes a negative mean reward of a element where most dialogues succeed(more than half of dialogues succeed) while w_j denotes the negative mean reward of a element where most dialogues failed. Then,

$$0 \leq \frac{w_i}{max(r)} \leq 1 \tag{9}$$

$$1 \leq \frac{w_j}{max(r)} \leq 1 + \frac{|min(r)|}{max(r)} \tag{10}$$

So if we consider α_1,α_2, where $\alpha_1 < \alpha_2$ and α_3, where $\alpha_3 > 1$. Then,

$$(\frac{w_i}{max(r)})^{\alpha_1} > (\frac{w_i}{max(r)})^{\alpha_2} \tag{11}$$

$$(\frac{w_j}{max(r)})^{\alpha_1} < (\frac{w_j}{max(r)})^{\alpha_2} \tag{12}$$

and,

$$(\frac{w_i}{max(r)})^{\alpha_3} < \frac{w_i}{max(r)} \tag{13}$$

$$(\frac{w_j}{max(r)})^{\alpha_3} > \frac{w_j}{max(r)} \tag{14}$$

This means not only α controls the affect of the reward, but also it controls the separation between elements with majority failed dialogues and elements with majority successful dialogues.

5 Experiments and Results

We performed a set of well-designed experiments to evaluate our proposed methods. All the experiments are done in the movie booking domain. As the user simulator, we use the open-source simulator described by Li et al. (2016). The agent train by the user simulator for 200 episodes then evaluated against a test set. Each episode consists of 20 simulated dialogues followed by one epoch of training. Totally 29 slots are used and all of them are available from the beginning of the training process. To represent the dialogues we constructed 268-dimensional a feature vector.

From the dataset, we randomly sampled 30 user goals at each experiment as the test dataset. The rest of the dataset is used for training the agent. Purpose of the experiment is to measure the performance of the proposed methodology and evaluated against a baseline when number of available training samples are low. So we vary the number of training samples available for the agent in each experiment to train(or in case of self play, available for calculating probabilities). Experiments are done for different sizes of training datasets ranging from 10 to 70(we use a interval of size 5 when varying size of training samples). Each experiment-setting is done (i) without Self-play (ii) Self-play with naive normalization (iii) Self-play with softmax normalization (iv) Self-play without reward-based sampling. For each experiment setting, we plot the results as the mean of the five

Training dataset size	Without selfplay	Selfplay with naive	Selfplay with softmax	Selfplay without reward based sampling
10	0.19	0.63	**0.84**	0.58
15	0.32	0.71	**0.81**	0.60
20	0.32	0.79	**0.81**	0.66
25	0.43	0.78	**0.86**	0.76
30	0.43	0.68	**0.80**	0.73
35	0.46	0.73	**0.84**	0.68
40	0.43	0.73	**0.76**	0.68
45	0.38	0.75	**0.79**	0.76
50	0.43	**0.81**	0.72	0.68
55	0.43	**0.84**	0.79	0.76
60	0.41	**0.78**	0.76	0.73
65	0.40	**0.77**	0.76	0.68
70	0.44	**0.80**	**0.80**	0.74

Table 1: Mean test success rate for (i) without Self-play (ii) Self-play with naive normalization (iii) Self-play with softmax normalization (iv) Self-play without reward-based sampling. The mean is calculated for 5 different random seeds.

experiments. In all cases we have used a Deep-Q-Network(Mnih et al., 2013) as the agent. The results of the experiment are shown in the Figure 4.

Training details: First, the rule-based agent interacts with the simulator with 120 dialogues for the replay buffer spike. Next, the RL agent interacts with a simulator for 200 episodes each consist of 20 simulated dialogues. The resultant state-action reward pairs store in the replay buffer. Each episode followed by one epoch of training with a batch size of 16. During the training for the epoch, the agent freezes the target network parameters and then update the local Q -function.

Self-play mechanism also contains hyperparameters. We use the ϵ_i with $\delta = 0.0005$ for $P(s_q)$, $\delta = 0.01$ for $P(req|s_q)$ and $\delta = 0.01$ for $P(kb|inform, s_q)$. Also we use $\alpha = 0.5$ and $\beta = 1$ for reward based sampling with naive normalization and $\alpha = 0.5$ and $\beta = 10$ for reward based sampling with softmax normalization. α and β values are determined by grid-search. We start reward based sampling from 60^{th} episode onward.

Architectural details: All models are implemented as multilayer perceptrons(MLP's) with ReLU activations. Each model consists of 2 hidden layers each having 64 hidden neurons. The Adam optimizer (Kingma and Ba, 2014) used as the optimizer with the learning rate of 0.0005. The training batch size is 16. The target network update by soft update instead of hard update.

Results: As shown in Figure 4, our proposed methodology did not overfit and perform better than the baseline in every case. Directly training using the training data tends to overfit the model, hence a higher training accuracy is achieved, but the respective test accuracy is low. However, the test accuracy is increasing for the baseline model as the number of training samples increase. This is because as the amount of training samples increase, the model can generalize better. In most cases, reward-based sampling(prioritizing failed dialogues) method outperform the random sampling method(without prioritizing failed dialogues), especially when the number of training samples is extremely low. This is because the reward base sampling provides a better exploration of agenda space. Also, when the number of training examples available is low, reward base sampling with softmax function yields the best result. Since Softmax function takes the exponent of the probability and the mean negative reward's multiplication, the final probability calculation for elements has a higher standard deviation. This increase the degree of exploration, which is helpful when the number of training samples is low.

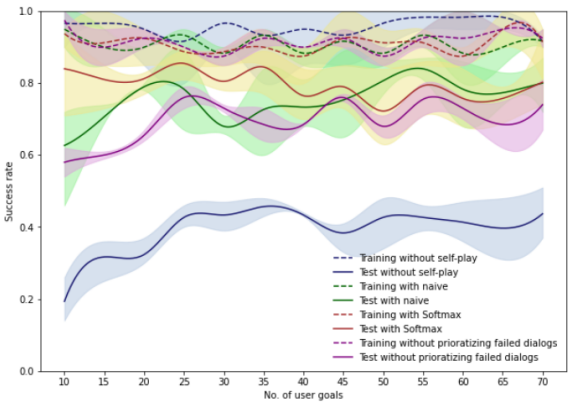


Figure 4: The mean train and test success rate for (i) without Self-play (ii) Self-play with naive normalization (iii) Self-play with softmax normalization (iv) Self-play without reward-based sampling. The mean is calculated for 5 different random seeds and curve is smoothed using interpolate spline technique.

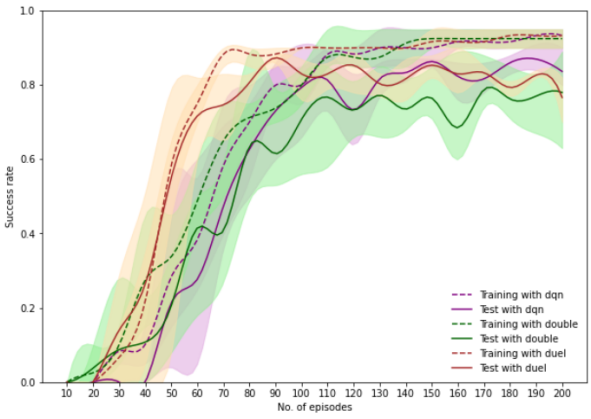


Figure 5: The mean train and test success rate for different RL methods. Kept number of training samples constant at 40 and vary the number of training episodes. The mean is calculated for 5 different random seeds and curve is smoothed using interpolate spline technique

Rate of convergence To verify that our proposed method train the model in a way such that models improve the testing success rate with their training success rate, we conduct an experiment keeping the training sample size constant and testing at different stages when training. This way, we can track how the test success rate improves with the number of episodes. In all cases, we use the reward-based sampling with naive normalization with an $\alpha = 0.5$ and $\beta = 1$.

We use the same simulator with the same movie booking dataset for this experiment as well. We randomly separate the data into train and test set, and in each run of the experiment, we use 40 samples for training and 40 samples for testing. We use the 40 training samples for calculating probability distributions, then use these probability distributions for Self-play mechanism. The agents interact with the user simulator for 200 episodes where each episode consists of 20 simulation dialogues followed by an epoch of training. After every 10 episodes, an evaluation is conducted using the test set. We start the reward-based sampling at 60^{th} episode. We experiment 5 times with each time with different train and test set that is sampled from the data. The mean train and test success rate of each of the RL method is shown in Figure 5.

The results which are shown in Figure 5 shows that our method improves the test success rate along with its train success rate. In all algorithms, our proposed method performs well without overfitting. All RL algorithms improve their performance as the number of episodes increase, and the test success rate is improving along with train success rate without an apparent lag.

6 Conclusion

We proposed a method to train RL agents for dialogue policy learning without overfitting. The methodology includes a Self-play technique that uses underlying probability distribution for agenda generation and reward-based sampling technique that prioritizes failed dialogues. We have shown that our method performs well compared to baseline as well as that the method can train a policy agent without overfitting. We also have shown that reward-based sampling method performs well in the exploration of agenda space. It also performs better than the random sampling method. We see several possible paths for future work. One of the main is using a better reward function. Also, we like to expand the work so that a full task-oriented conversational agent can be made by modeling the problem as an active learning problem with a better reward function. So an end-to-end task-oriented conversational agent can be made for low resource setting.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. *Learning end-to-end goal-oriented dialog*. arXiv preprint arXiv:1605.07683.

Lu Chen, Xiang Zhou, Cheng Chang, Runzhe Yang, and Kai Yu. 2017. *Agent-aware dropout dqn for safe and efficient on-line dialogue policy learning*. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2454–2464.

Lucie Daubigney, Matthieu Geist, Senthilkumar Chandramohan, and Olivier Pietquin. 2012. *A comprehensive reinforcement learning framework for dialogue management optimisation*. IEEE Journal of Selected Topcis in Signal Processing, 6.

Floris Den Hengst, Mark Hoogendoorn, Frank Van Harmelen, and Joost Bosman. 2019. *Wizard of wikipedia: Knowledge-powered conversational agents*. arXiv preprint arXiv:1811.01241.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. *Towards end-to-end reinforcement learning of dialogue agents for information access*. arXiv preprint arXiv:1609.00777.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. *Wizard of wikipedia: Knowledge-powered conversational agents*. arXiv preprint arXiv:1811.01241.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. *Neural approaches to conversational AI*. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 1371–1374.

Milica Gasic and Steve Young. 2014. *Gaussian processes for pomdp-based dialogue manager optimization*. Audio, Speech, and Language Processing,IEEE/ACM Transactions on, 22:28–40.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. *Hybrid Reinforcement/Supervised Learning of Dialogue Policies from Fixed Data Sets*. Computational Linguistics, 34:487–511.

Diederik P. Kingma and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Vijay R Konda and John N Tsitsiklis. 2000. *Actorcritic algorithms*. In Advances in neural information processing systems, pages 1008–1014.

Staffan Larsson and David Traum. 2000. *Information state and dialogue management in the trindi dialogue move engine toolkit*. Natural Language Engineering, 6:323–340.

Lihong Li, He He, and Jason Williams. 2015. *Temporal supervised learning for inferring a dialog policy from example conversations*. 2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings, pages 312–317.

Lihong Li, Jason D. Williams, and Suhrid Balakrishnan. 2009. *Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection*. Tenth Annual Conference of the International Speech Communication Association.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. *A user simulator for task-completion dialogues*. arXiv preprint arXiv:1612.05688.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. *Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems*. In ThirtySecond AAAI Conference on Artificial Intelligence.

Bing Liu and Ian Lane. 2018. *Adversarial learning of task-oriented neural dialog models*. arXiv preprint arXiv:1805.11762.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. *Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems*. arXiv preprint arXiv:1804.06512.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. *Goal-oriented dialogue policy learning from failures*. Proceedings of the AAAI Conference on Artificial Intelligence.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. *Playing atari with deep reinforcement learning*. arXiv preprint arXiv:1312.5602.

Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, andQiang Yang. 2018. *Personalizing a dialogue system with transfer reinforcement learning*. Thirty-Second AAAI Conference on Artificial Intelligence.

Diana Perez-Marin and Ismael Pascual-Nieto. 2011. *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global.

Joelle Pineau and Sebastian Thrun. 2004. *Spoken Dialogue Management Using Probabilistic Reasoning*. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. *Proximal policy optimization algorithms*. arXiv preprint arXiv:1707.06347.

Pararth Shah, Dilek Hakkani-Tur, and Larry Heck. 2016. *Interactive reinforcement learning for taskoriented dialogue management*.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018a. *Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and online reinforcement learning*. In Proceedings of the 2018

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 41–51.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018b. *Building a conversational agent overnight with dialogue self-play*. arXiv preprint arXiv:1801.04871.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, TsungHsien Wen, and Steve Young. 2016a. *Continuously learning neural dialogue management*. arXiv preprint arXiv:1606.02689.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina RojasBarahona, Stefan Ultes, David Vandyke, TsungHsien Wen, and Steve Young. 2016b. *On-line active reward learning for policy optimisation in spoken dialogue systems*. arXiv preprint arXiv:1605.07669.

Richard S. Sutton and Andrew G Barto. 2018. *Reinforcement learning*. Cambridge, Mass: MIT Press.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. *A networkbased end-to-end trainable task-oriented dialogue system*. arXiv preprint arXiv:1604.04562.

Jason Williams and Steve Young. 2007. *Partially observable Markov decision processes for spoken dialog systems*. Computer Speech & Language,21:393–422.

Ronald J. Williams. 1992. *Simple statistical gradient-following algorithms for connectionist reinforcement learning*. Machine learning, 8(3-4):229–256.

Learning to Describe Editing Activities in Collaborative Environments: A Case Study on GitHub and Wikipedia

Edison Marrese-Taylor[1], Pablo Loyola[2], Jorge A. Balazs[1] and Yutaka Matsuo[1]
Graduate School of Engineering, The University of Tokyo, Japan[1]
{emarrese, jorge, matsuo}@weblab.t.u-tokyo.ac.jp
IBM Research, Tokyo, Japan[2]
e57095@jp.ibm.com

Abstract

We propose to study the automatic generation of descriptions from content editing activities in collaborative environments. We define such task as identifying the changes associated to two consecutive versions of a document, and then producing a message in natural language that explains it, which should provide a compact description of the change while retaining its key informative elements. Our model is based on a sequence to sequence architecture that receives as input the representation of the change, and outputs a message. We propose a framework to conceptualize the problem and two instances for GitHub activity and Wikipedia contributions, two of the most important collaborative systems on the Web. Our results indicate that the proposed approach is able to generate feasible descriptions, which are on average aligned with the semantic purpose of the editing activities.

1 Introduction

One of the positive outcomes of the current pervasiveness of the Web has been the boost of collaboration across several domains. Examples of this are platforms such as Wikipedia and GitHub, where self-organized and voluntary groups of individuals gather based on the common goal of crafting documents and programs (Crowston et al., 2007).

The outcome of these collaborative activities is usually the result of a series of incremental modifications over time. For example, in the case of GitHub, incremental modifications are usually functional changes, which allow to incorporate new features or fix reported bugs. In the case of wiki-based platforms, contributors modify the content of a given article in order to reflect an update on the matter the article is dealing with. The transparency and openness of change management provides complete awareness of the state of the document being crafted, at any point in time (Dabbish et al., 2012).

As collaboration is carried out in a decentralized fashion, the coordination between contributors plays a key role (Von Krogh et al., 2003). While there exist direct ways of communication, such as bug trackers and discussion forums, we are interested in studying the indirect ways in which contributors interact and coordinate.

One of these elements are the short messages that the contributors provide at the time of submitting the change. This short message usually provides a description of the change and serves as a way of broadcasting it to the rest of the community, ideally clarifying the purpose and other technical aspects, and supporting the reviewing process (Guzman et al., 2014).

Therefore, our goal is to use this set of change-message pairs to develop a model able to explain collaborative activities by automatically generating a short passage in natural language. In that sense, we visualize this task as being in-between summarization and translation given the challenges it presents, namely, (i) the length asymmetry between changes and their messages, and (ii) the fact that documents can be written in modalities different from natural language (e.g. source code, art). Our intention is to learn the most salient elements that characterize the change, and then decode them into a description

in natural language. As we will show, this duality has implications on the choices of metrics used for evaluation.

Moreover, rather than describing the content that was changed, we are generating a description of the *action* taken over it, therefore, there is an inherent temporal dimension associated to the generation. Additionally, the action can be seen as the result of an optimization problem: given the state of the file, the agent needs to find the most efficient change that allows him to satisfy the requirement. In other words, the change performed on the file is a function dependent on the current functional state of the file and the given requirement: the change performed was such, only because of the given state of the file. If such state was different, then the change would have been different too.

The usage of models based on deep neural networks in natural language processing has been successful in large part because they learn and use their own continuous numeric representational systems for word and sentences. In particular, distributed representations (Hinton, 1984) applied to words (Mikolov et al., 2013) have meant major breakthroughs allowing networks to parse and represent sentences and phrases using an effective compositional vector grammar. Recurrent neural networks now provide state-of-the-art performance in tasks such as machine translation, sentence-level sentiment analysis, text generation and automatic image captioning.

Moreover, the introduction of the encoder-decoder (Cho et al., 2014) or sequence-to-sequence (Sutskever et al., 2014) architectures presented a successful framework based on neural networks that aims to map highly structured input to highly structured output. Additional improvements on the encoder-decoder architecture came with the addition of attentional components (Bahdanau et al., 2015; Luong et al., 2015), which allowed the decoder to focus on specific information provided by the encoder at a time.

Therefore, to tackle the introduced problem we use a representation learning approach. More concretely,our approach takes inspiration in recurrent neural models, widely used in sequence-to-sequence learning (Bahdanau et al., 2015), but we introduce specific extensions to account for the structural differences in our case. While the Web offers several types of collaborative environments, in this work we focus on GitHub activity and Wikipedia contributions, based on their popularity and data availability. We perform an empirical study based on collected editing activity, and show that the introduced models are able to learn representations from the changes and produce sound descriptions in most cases.

2 Related Work

The analysis of editing activities has been tightly associated with quality assessment tasks. For example, in software engineering, version changes are the basis of regression testing and defect prediction (McIntosh and Kamei, 2017).

In the case of Wikipedia, since one of its core principles is being open for anyone to maintain it, Wikipedia cannot fully ensure the reliability of its articles, and thus sometimes had suffered criticism for containing low-quality information. It is therefore essential to assess the quality of Wikipedia articles automatically. In this context, for example, Su and Liu (2015) approach the problem by using a psycho-lexical resource. On the other hand, Kiesel et al. (2017) aim at automatically detecting vandalism utilizing change information as a primary input. Gandon et al. (2016) also validate the importance of the editing history of Wikipedia pages as a source of information, presenting a new extraction technique which produces a linked data representation for it.

More recently, Yang et al. (2017) proposed an approach for identifying semantic edit intentions from revisions in Wikipedia. Also, Sarkar et al. (2019) and Marrese-Taylor et al. (2019) have focused on the quality assessment issue and proposed approaches that directly produce an edit-level quality label for a given Wikipedia edit. While the former is concerned only with edit-level quality classification of edits, the latter also incorporates a generative part similar to ours but only as an auxiliary task.

Our work is also related to summarization on Wikipedia. Recent work includes Chisholm et al. (2017), where the authors proposed an autoencoder-based model to generate short biographies, and Zhang et al. (2017), where authors present a method to summarize the discussion surrounding a change

in the content, along with a visualization tool to ease comprehension of its evolution.

When it comes to GitHub, we find several papers that perform analyses over the platform, including the work of Batista et al. (2017), who study the correlation among features that measure the strength of social coding collaboration and Nielek et al. (2016), who try to predict which developer will join which project.

In terms of specifically working on code change descriptions, we see that the paradigm is based on the distributional similarities that emerge between natural and programming languages (Hindle et al., 2012). Indeed, both are ways of communication based on sets of defined vocabularies, and their composition is based on structured and sequential instructions. Concretely, Cortes et al. (2014) and Linares et al. (2015) proposed methods based on a set of rules that consider the type and impact of the changes, and Buse and Weimer (2010) combine summarization with symbolic execution.

Moreover, mapping source code to natural language has received special attention in recent years, mainly in the form of summarization. Examples if this are the work of Allamanis et al. (2016) who use a convolutional neural network approach, and Iyer et al. (2016) who used an recurrent neural network architecture capable of learning to summarize Stack Overflow snippets.

In terms of code change description generation, the use of a representation learning paradigm has been proposed Jiang et al. (2017; 2017) and by Loyola et al. (2017; 2018). The authors train an encoder-decoder architecture on a set of commit-message pairs extracted from GitHub open source projects to generate change descriptions. We took that work as a starting point and proposed an extended architecture that considers intra-change comments with an ad-hoc attention mechanism, with the additional feature of generalizing to other data sources such as Wikipedia changes. More recent variations of this include augmenting the model with a pointer network (Liu et al., 2019a), or with abstract syntax trees (Liu et al., 2019b). In contrast, Liu et al. (2018) focused on efficiency and proposed a method that relies on nearest neighbors instead the encoder-decoder.

Finally, our work is also related to Yin et al. (2019), who proposed a general framework for learning edit representations based on a self-supervised approach similar to an auto-encoding task.

3 Proposed Approach

Generative tasks, such as summarization and translation, try to map between source and target sequences ignoring time dependencies across examples. Our main motivation for this work is to explore a task where we can generate a natural language description of a *transition* between states, i.e., adding temporal dimension into the generation by learning to represent the difference between consecutive versions of the changed document.

Web-based collaborative platforms represent a convenient source of indirectly supervised data, as each contributed change is usually required to be submitted along with a short description of its purpose and detail. For this work, we focus on source code changes on GitHub and Wikipedia contributions, based on their availability.

From a broad perspective, a *change* can be seen as the consequence of a requirement, which can be *external*: e.g., the need for a new functionality on a GitHub project, or the need reflect a recent event on someone's biographical article on Wikipedia; or *internal*: e.g., a reported bug or a functionality mismatch on a GitHub project, or the need to revert a vandalism attack on a Wikipedia article. That requirement is internalized by a contributor that identifies which portion of the document should be modified in order to satisfy the requirement. As all proposed changes are expected to be reviewed by a peer, the contributor appends a short description explaining the purpose. Such dual configuration (changes, descriptions) represents our main data source for training.

For both modalities, **GitHub** and **Wikipedia**, we assume the existence of T versions of a given project or article $\{v_1, \ldots, v_T\}$. Given a pair of consecutive versions (v_{t-1}, v_t), we define the tuple (C_t, N_t), where $C_t = \Delta_{t-1}^{t}(v)$ is a representation of the content changes associated to v in time t, and N_t is a representation of its corresponding natural language (NL) description. Let $\mathcal{C}$ be the set of content changes and $\mathcal{N}$ be the set of all descriptions in NL. We con-

sider a training corpus with T content snippets and summary pairs (C_t, N_t), $1 \leq t \leq T$, $C_t \in \mathcal{C}$, $N_t \in \mathcal{N}$. Then, for a given content snippet $C_k \in \mathcal{C}$, the goal of our model is to produce the most likely NL description $N^\star$. The nature of the content snippet $C_k \in \mathcal{C}$ depends of the modality considered.

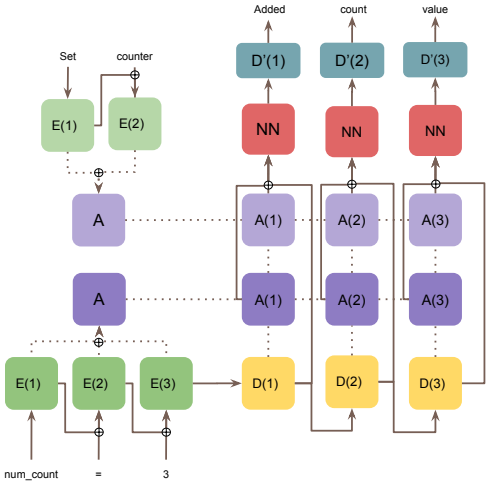


Figure 1: Model architecture with two encoders for the GitHub modality.

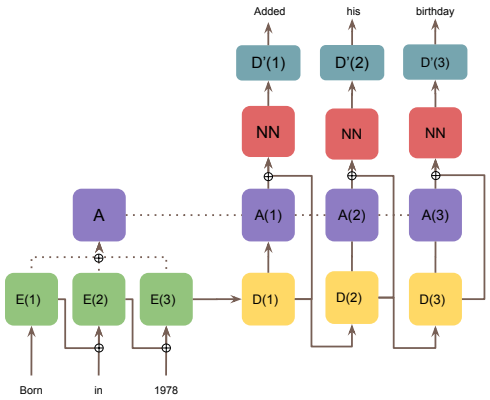


Figure 2: Model architecture based on sequence-to-sequence for the Wikipedia modality.

For both modalities, similarly to Iyer et al. and Loyola et al. (2016; 2017), we use an attention-augmented encoder-decoder architecture. On each case, we assume the existence of an ad-hoc encoder that allows us to obtain a representation for the change we intend to study. The only assumption about our encoders is that their inputs have to be represented as a sequence of tokens, $c_i \in C_t$. With this, our encoders rely on embeddings and bidirectional LSTMs. Let $X_t = x_1, \ldots x_n$ be the embedded input content sequence C_t, using embedding matrix E.

$$\vec{h}_i = LSTM(x_i, \vec{h}_{i-1}) \quad (1)$$

$$\overleftarrow{h}_i = LSTM(x_i, \overleftarrow{h}_{i+1}) \quad (2)$$

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (3)$$

We add special beginning-of-sentence *BOS* and end-of-sentence *EOS* tokens to our output NL sequences, and set the decoder to be an LSTM that reads the representation given by the encoder, generating NL words one at a time based on its current hidden state and guided by a global attention model (Luong et al., 2015). We model the probability of a description as a product of the conditional next-word probabilities. We use an embedding matrix D to encode each NL token $n_i \in N_t$ into a sequence of vectors $Y_t = y_1, \ldots y_m$ and set

$$s_i = LSTM(y_{i-1}, s_{i-1}) \quad (4)$$

$$p(n_i|n_1, \ldots, n_{i-1}) \propto W \tanh(W_1 s_i + W_2 a_i) \quad (5)$$

where $\propto$ denotes a softmax operation, s_i represents the decoder hidden state and a_i is the contribution from the attention model on the input. W, W_1 and W_2 are trainable combination matrices. The decoder repeats the recurrence until a fixed number of words or a special *END* token is generated. The attention contribution a_i is defined as $a_i = \sum_{j=1}^{k} \alpha_{i,j} \cdot h_j$, where $h_j \in H$ is a hidden state associated to the input and $\alpha_{i,j}$ is:

$$t_i = \sum_{j=1}^{k} \alpha_{i,j} \cdot h_j \quad (6)$$

$$\alpha_{i,j} = \frac{\exp(h_i^\top s_i)}{\sum_{h_j \in H} \exp(h_j^\top s_i)} \quad (7)$$

In this way, the decoder is trained as a conditioned language model over the NL vocabulary and on each generation step we let it have full access to the representation of the input as provided by the encoder using the attentional component.

Wikipedia: We set $C_k = x_1, ..., x_L$, as a sequence of L text tokens associated with a change. To encode this sequence we use a bidirectional LSTM.

GitHub: We build C_k based on both code and documentation changes, as extracted from $\Delta_{t-1}^{t}(v)$.

We define the change in source code C_k as having two components: a sequence of source code tokens $SC_k = x_1, ..., x_{L_{SC}}$, and a sequence of documentation tokens $SD_k = z_1, ..., z_{L_{SD}}$. To obtain a vector representation for $\Delta_{t-1}^{t}(v)$, as we model it as two different sequences, we use two bidirectional LSTMs as encoders, one for the source code sequence and one for the documentation sequence. We aggregate each representation using mean pooling and concatenate the resulting vectors. The resulting vector is used to initialize the decoder hidden state.

During training, for both modalities, the decoder iterates until the end of the sentence is reached. For generation, we approximate $N^{\star}$ by performing a beam search on the space of all possible summaries using the model output, with a beam size of 10 and a maximum summary length of equal to the maximum length of the input. For inference, we let the decoder run for this number of steps or until the *EOS* token is generated.

4 Empirical Study

4.1 Wikipedia

We collected historical data dumps from Wikipedia, choosing some of the most edited articles in English and German, in a way analog to the language choice in GitHub. For English, we worked with the articles for *United States* and *World War II*, while for German we chose *Deutschland* (Germany) and *Zweiter Weltkrieg* (World War II). To our eyes, one of the critical differences between our studied modalities is the amount of control users have over the editing activity, which is practically non-existent in the case of Wikipedia. To study this, we also collected the editing history of *Donald Trump*'s article, which exhibited a very dynamic and polarizing editing activity record.

Wikipedia dumps contain every version of a given page in *wikitext*, the official markup-like language, along with metadata for every edit. To obtain the content associated to each $\Delta_{t-1}^{t}(v)$, we sorted the extracted edits chronologically and computed the *diff* of each pair of consecutive versions using the Unix diff tool. Due to the line-based approach of the Unix *diff* tool, small changes in *wikitext* led to big chunks of differences in the resulting *diff* file. To alleviate this problem, we extracted the unique set of sentences that was either added or removed, which gave us a much fine-grained characterization of the edits. For English sentence splitting we used the automatic approach by Kiss et al. (2006), and Somajo (Proisl and Uhrig, 2016), for German.

We found that articles related to controversial topics —such as Donald Trump— exhibited a high proportion of reverting edits, as well as extreme vandalization cases. Since these edits provide no additional information to our model, we filtered them out.

4.2 GitHub

We rely on the concept of code commit, the standard contribution procedure implemented in modern subversion systems (Gousios et al., 2014), which provides both the actual change and a short explanatory paragraph. To model both as a sequence of source code tokens $SC_k = x_1, ..., x_{L_{SC}}$, and a sequence of documentation tokens $SD_k = z_1, ..., z_{L_{SD}}$ we use *diff* files associated to each commit for a given project in GitHub. These *diff* files encode per-line differences between two files or sets of files in a standard format, allowing us to recover source code changes at the line level.

We obtain all the *diff* files for a given project using the GitHub API. However, given the flat structure of the diff file, source code in contiguous lines might not necessarily correspond to originally neighboring code lines. Moreover, they might come from different files in the project. To deal with this issue, we followed Loyola et al. (2017) and only considered the *diff* files of those commits that modify a single file in the project.

To obtain the messages associated to each introduced change, we use the API to download the metadata associated to each commit, which allows us to recover information such as the author and message of each commit.

For this paper, we chose projects for Python and Javascript, as they are among the most widely adopted programming languages. We selected two of the historically most popular projects for each language on GitHub as data sources. For Python, we worked with *Theano* and *youtube-dl*, whereas for Javascript we worked with *angular* and *react*. We parsed the *diff* files using a lexer (Brandl, 2016) to tokenize their contents in a per-line fashion.

Modality	Dataset	Max. Length	Our Model		MOSES
			METEOR	BLEU	BLEU
GitHub	Theano	100	0.319	27.3	5.9
		300	0.220	27.4	5.5
	youtube-dl	100	0.132	18.3	17.6
		300	0.325	12.7	13.0
	angular	100	0.254	21.6	12.7
		300	0.412	20.2	9.7
	react	100	0.330	27.9	10.5
		300	0.263	22.6	7.3
Wikipedia	World War II	100	0.399	14.3	11.8
		300	0.244	14.5	5.2
	Zweiter Weltkrieg	100	0.330	17.5	16.3
		300	0.312	12.1	9.8
	United States	100	0.241	12.6	11.3
		300	0.325	12.8	9.0
	Deutschland	100	0.352	14.2	14.8
		300	0.352	13.9	10.4
	Donald Trump	100	0.610	14.7	10.5
		300	0.581	12.5	7.8

Table 1: Summary of our results on both modalities.

Modality	Max. Length	Mean Ours	Mean MOSES
GitHub	100	23.8	11.7
	300	20.7	8.9
Wikipedia	100	14.7	12.9
	300	13.2	8.4

Table 2: Summary, in terms of BLEU scores, of the impact of increasing the maximum sequence length across modalities.

The extracted commit end edit messages were processed using the Penn Treebank tokenizer (Marcus et al., 1993), which nicely deals with punctuation and other text marks typical of natural language. During experimentation, we found that some excessively repeating patterns on the NL descriptions, such as the phrase *merge pull request*, were misguiding for the learning process so we deleted them from the data, keeping the rest of the content of each sequence, if any. Sequences that solely contained these sequences were discarded.

To evaluate the quality of our generated descriptions we use METEOR (Lavie and Agarwal, 2007) and sentence level BLEU-4 (Papineni et al., 2002). These metrics, popular from automatic machine translation evaluation, are scores calculated for individual translated segments by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. We compute them on our validation set after every epoch and save the intermediate model that maximizes each.

Following previous work on mapping source code to natural language (Loyola et al., 2017; Iyer et al., 2016), we used MOSES (Koehn et al., 2007) as a baseline, which although is designed as a phrase-based machine translation system, was previously used by Iyer et al. (2016) to generate text from source code. Concretely, we treated the tokenized input (only the source code for the case of GitHub) as the source language and the NL description as the target. We trained a 3-gram language model using KenLM (Heafield et al., 2013) and used mGiza to obtain alignments. For validation, we use minimum error rate training (Bertoldi et al., 2009; Och, 2003) in our validation set. To evaluate model capabilities, we generated two versions of each dataset for a maximum input/output sequence length of 100 and 300 tokens.

Data		Reference	Generated
GitHub	Theano	better test error UNK	better error message
		allow to disable the gpu when UNK and UNK	disable the gpu back-end .
		add test case .	added test message for UNK
	youtube-dl	[cbc] skip geo-restricted test case	[generic] add test
		[extractor/generic] add support for onionstudios embeds (closes # NUMBER)	[extractor/generic] handle UNK embeds (closes # NUMBER)
	angular	refactor (UNK) : remove UNK facade (# NUMBER)	refactor (changelog) : add UNK (# NUMBER)
		fix (core) : export dev mode api in UNK closes # NUMBER	fix (UNK) : add UNK UNK closes # NUMBER
	react	clarify tutorial UNK fixes # NUMBER .	clarify tutorial
		add shirtstarter to examples of UNK UNK .	update shirtstarter UNK
Wikipedia	D. Trump	/* Foreign policy */ wiki link	/* Foreign policy */ cite cleanup
		UNK not graduate from Fordham	He did not graduate from Fordham University
	U.S.	/* Economy */ Updated unemployment rate	/* Economy */ Its the US
		/* Economy */ update CPI	/* Economy */ update inflation data
	Deutschland	Änderungen von Benutzer : UNK rückgängig gemacht und letzte Version von Benutzer : UNK wiederhergestellt	Änderungen von Benutzer : UNK rückgängig gemacht und letzte Version von Benutzer : Aka wiederhergestellt
		/ * Von der Bonner zur Berliner Republik (1990 –Gegenwart) * / kor .	/ * Von der Bonner zur Berliner Republik (1990 –Gegenwart) * /

Table 3: Generated v/s original NL descriptions.

5 Results and Discussion

We summarize our results in terms of both METEOR and BLEU metrics on Table 1. Although we think these metrics may not be completely compatible with our task, since it is not exactly translation, results show that they indeed provide a notion of the degree of alignment between the modalities we are mapping. To gain insight into this we analyzed the cross-run correlation between each metric and the validation cross-entropy loss. We found that METEOR is generally more negatively correlated with the loss. Given that this metric uses language-specific resources, we think it may be over-estimating the quality of the generated passages, as in our case they are not regular English phrases. Based on these results, we relied on BLEU to choose the best model each time.

As shown in Table 2, our approach consistently outperforms the baseline. This is even clearer when increasing the maximum length from 100 to 300, which always considerably hinders the baseline's performance but has a comparatively smaller effect on our model. For the particular case of *Theano*, where the increment in length size affected BLEU positively but METEOR negatively for our model, we found that the sizes of both the source vocabulary and the number of training instances increased more compared to other cases —3% and 28% respectively— which could explain the abnormal behavior.

In the case of *youtube-dl*, where MOSES performed better than our approach, we found that the change in maximum length produced a considerable imbalance between the mean lengths of the source and target sequences. Further work is needed to devise a more effective learning strategy in such cases.

In terms of the modalities studied, we see that for GitHub, while the gains of the proposed model against MOSES for both Javascript and Python projects are similar for both sequence length settings –average of 13% and 12% for Javascript, and 11% and 10 % for Python– Python presents higher variance, which is caused by the disparity in performance between *Theano* and *youtube-dl*. In the case of Wikipedia, the model performs consistentl well across articles, always outperforming the baseline.

A more qualitative result is presented in Table 3, where we compare the generated descriptions against the ground truth messages from the test set. In general, we see that the model is able to consistently generate semantically sound descriptions, which are also semantically well correlated to the reference messages. Our results also suggest the emergence of rephrasing capabilities, as the models tend to choose general terms over more specific ones, while also dropping parts of the messages that may seem irrelevant.

An important note is that the model suffers from hallucination, a common problem in sequence-to-

sequence models. Specifically, in the case of GitHub, we see that for those projects whose NL messages exhibit a fixed pattern in their structure, such as in the case of *youtube-dl* where users add a header denoting the file that was edited in the commit, the model tends to more frequently hallucinate the content of the message. In this case, as the content of the "header" section may be too specific for the model to leverage on, we believe this restrains the generation capabilities of the decoder, making it more prone to memorization and therefore less able to correctly generalize.

In the case of Wikipedia, we observed that in most of the cases the model was able to correctly generate the portion of the edit messages that lies between the "/*" symbols, which again can be regarded as a message "header". Compared to the case of GitHub, the nature of the header seems to be different, however. We manually checked the messages and discovered that most of the headers correspond to section titles of the Wikipedia articles. For most of the Wikpedia articles that we worked with, we found that wiki editors tend to add this information as a "header" as a way to more directly communicate with other editors, in a way akin to what we observed in the case of some GitHub projects. In this case, this behavior was more consistent across datasets. As the "header" will probably be highly correlated to the nature of the change introduced, we think in this case the model is indeed able to leverage on this content to correctly generate the message. However, despite the model capabilities in terms of "header" generation, we also observe cases of hallucination in the parts of the messages that lie outside "headers". This is specially apparent in some of the examples for *Donald Trump* and *United States*, as Table 3 shows.

6 Conclusions and Future work

In this paper we proposed to study the automatic generation of descriptions for editing behavior in online content. Concretely, we introduced models based on the encoder-decoder architecture that are able to generate natural language descriptions for editing activities in Wikipedia and GitHub.

We think our results could represent a concrete contribution in improving our understanding of the evolution knowledge bases, in terms of both software and scientific documentation, from a linguistic perspective. We envision this as a tool that could be useful for supporting documentation and quality-related tasks in collaborative environments, where human supervision is insufficient or not always available.

In terms of future work, one of the main lines we intend to explore is the the design of an ad-hoc metric for automatic evaluation of the generated messages. Alongside that, we also intend to do an in-depth human study for a more comprehensive validation and assessing the usefulness of the descriptions we generate. On the other hand, we also intend to improve our models by allowing feature learning from richer inputs, such as abstract syntax trees and also functional such as execution traces in the case of GitHub.

Finally, in this work we have resorted to *diff* files as a primary source of input information, which means our representation contains redundant information and may therefore be inefficient. Although our results showed that this representation works fairly well for the proposed setting, at the same time providing us a model that is language agnostic, we would like to explore other alternatives to model the input. In particular, we are interested in models that directly take a pair of versions of a given document, for example the version before and after a certain introduced change, allowing us to generalize our proposal to different time scales.

Acknowledgments

We are grateful for the support provided by the NVIDIA Corporation, donating two of the GPUs used for this research.

References

[Allamanis et al.2016] Miltiadis Allamanis, Hao Peng, and Charles Sutton. 2016. A convolutional attention network for extreme summarization of source code. In *International Conference on Machine Learning (ICML)*.

[Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*, San Diego, California.

[Batista et al.2017] Natércia A. Batista, Michele A. Brandão, Gabriela B. Alves, Ana Paula Couto da Silva, and Mirella M. Moro. 2017. Collaboration strength metrics and analyses on github. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 170–178, New York, NY, USA. ACM.

[Bertoldi et al.2009] Nicola Bertoldi, Haddow Barry, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, pages 1–11.

[Brandl2016] Georg Brandl. 2016. Pygments: Python syntax highlighter. `http://pygments.org`.

[Buse and Weimer2010] Raymond P.L. Buse and Westley R. Weimer. 2010. Automatically documenting program changes. In *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering*, ASE '10, pages 33–42, New York, NY, USA. ACM.

[Chisholm et al.2017] Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642, Valencia, Spain, April. Association for Computational Linguistics.

[Cho et al.2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

[Cortés-Coy et al.2014] Luis Fernando Cortés-Coy, Mario Linares Vásquez, Jairo Aponte, and Denys Poshyvanyk. 2014. On automatically generating commit messages via summarization of source code changes. In *SCAM*, volume 14, pages 275–284.

[Crowston et al.2007] Kevin Crowston, Qing Li, Kangning Wei, U Yeliz Eseryel, and James Howison. 2007. Self-organization of teams for free/libre open source software development. *Information and software technology*, 49(6):564–575.

[Dabbish et al.2012] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1277–1286. ACM.

[Gandon et al.2016] F. Gandon, R. Boyer, O. Corby, and A. Monnin. 2016. Wikipedia editing history in dbpedia: Extracting and publishing the encyclopedia editing activity as linked data. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 479–482, Oct.

[Gousios et al.2014] Georgios Gousios, Martin Pinzger, and Arie van Deursen. 2014. An exploratory study of the pull-based software development model. In *Proceedings of the 36th International Conference on Software Engineering*, pages 345–355. ACM.

[Guzman et al.2014] Emitza Guzman, David Azócar, and Yang Li. 2014. Sentiment analysis of commit comments in github: An empirical study. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 352–355, New York, NY, USA. ACM.

[Heafield et al.2013] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

[Hindle et al.2012] Abram Hindle, Earl T Barr, Zhendong Su, Mark Gabel, and Premkumar Devanbu. 2012. On the naturalness of software. In *2012 34th International Conference on Software Engineering (ICSE)*, pages 837–847. IEEE.

[Hinton1984] Geoffrey E Hinton. 1984. Distributed representations.

[Iyer et al.2016] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2073–2083, Berlin, Germany, August. Association for Computational Linguistics.

[Jiang and McMillan2017] Siyuan Jiang and Collin McMillan. 2017. Towards Automatic Generation of Short Summaries of Commits. pages 320–323. IEEE, May.

[Jiang et al.2017] Siyuan Jiang, Ameer Armaly, and Collin McMillan. 2017. Automatically generating commit messages from diffs using neural machine translation. pages 135–146. IEEE, October.

[Kiesel et al.2017] Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2017. Spatio-temporal analysis of reverted wikipedia edits.

[Kiss and Strunk2006] Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4):485–525, December.

[Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine

translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. 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*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Linares-Vásquez et al.2015] Mario Linares-Vásquez, Luis Fernando Cortés-Coy, Jairo Aponte, and Denys Poshyvanyk. 2015. Changescribe: A tool for automatically generating commit messages. In *Proceedings of the 37th International Conference on Software Engineering-Volume 2*, pages 709–712. IEEE Press.

[Liu et al.2018] Zhongxin Liu, Xin Xia, Ahmed E. Hassan, David Lo, Zhenchang Xing, and Xinyu Wang. 2018. Neural-machine-translation-based commit message generation: How far are we? In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, ASE 2018, pages 373–384, Montpellier, France, September. Association for Computing Machinery.

[Liu et al.2019a] Qin Liu, Zihe Liu, Hongming Zhu, Hongfei Fan, Bowen Du, and Yu Qian. 2019a. Generating Commit Messages from Diffs Using Pointer-generator Network. In *Proceedings of the 16th International Conference on Mining Software Repositories*, MSR '19, pages 299–309, Piscataway, NJ, USA. IEEE Press.

[Liu et al.2019b] Shangqing Liu, Cuiyun Gao, Sen Chen, Lun Yiu Nie, and Yang Liu. 2019b. ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking. *arXiv:1912.02972 [cs]*, December.

[Loyola et al.2017] Pablo Loyola, Edison Marrese-Taylor, and Yutaka Matsuo. 2017. A neural architecture for generating natural language descriptions from source code changes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–292, Vancouver, Canada, July. Association for Computational Linguistics.

[Loyola et al.2018] Pablo Loyola, Edison Marrese-Taylor, Jorge Balazs, Yutaka Matsuo, and Fumiko Satoh. 2018. Content Aware Source Code Change Description Generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 119–128, Tilburg University, The Netherlands. 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.

[Marcus et al.1993] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

[Marrese-Taylor et al.2019] Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An Edit-centric Approach for Wikipedia Article Quality Assessment. In *Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019)*, pages 381–386, Hong Kong, China, November. Association for Computational Linguistics.

[McIntosh and Kamei2017] Shane McIntosh and Yasutaka Kamei. 2017. Are fix-inducing changes a moving target? a longitudinal case study of just-in-time defect prediction. *IEEE Transactions on Software Engineering*.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

[Nielek et al.2016] R. Nielek, O. Jarczyk, K. Pawlak, L. Bukowski, R. Bartusiak, and A. Wierzbicki. 2016. Choose a job you love: Predicting choices of github developers. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 200–207, Oct.

[Och2003] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

[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 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

[Proisl and Uhrig2016] Thomas Proisl and Peter Uhrig. 2016. Somajo: State-of-the-art tokenization for german web and social media texts. In *Proceedings of the*

9th Web as Corpus Workshop (WaC-X) and the EmpiriST Shared Task, pages 57–62, Berlin, Germany. Association for Computational Linguistics.

[Sarkar et al.2019] Soumya Sarkar, Bhanu Prakash Reddy, Sandipan Sikdar, and Animesh Mukherjee. 2019. StRE: Self Attentive Edit Quality Prediction in Wikipedia. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3962–3972, Florence, Italy, July. Association for Computational Linguistics.

[Su and Liu2015] Q. Su and P. Liu. 2015. A psycholexical approach to the assessment of information quality on wikipedia. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 184–187, Dec.

[Sutskever et al.2014] 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.

[Von Krogh et al.2003] Georg Von Krogh, Sebastian Spaeth, and Karim R Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32(7):1217–1241.

[Yang et al.2017] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying Semantic Edit Intentions from Revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

[Yin et al.2019] Pengcheng Yin, Graham Neubig, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. 2019. Learning to Represent Edits. In *Proceedings of the 7th International Conference on Learning Representations*.

[Zhang et al.2017] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. pages 2082–2096.

A Multilingual Linguistic Domain Ontology

Mariem Neji
MIRACL Laboratory
Sfax, Tunisia
mariem.neji@yahoo.fr

Fatma Ghorbel
CEDRIC Laboratory
Paris, France
fatmaghorbel@gmail.com

Bilel Gragouri
MIRACL Laboratory
Sfax, Tunisia
bilel.gargouri@fsegs.rnu.tn

Nada Mimouni
CEDRIC Laboratory
Paris, France
nada.mimouni@lecnam.net

Elisabeth Métais
CEDRIC Laboratory
Paris, France
elisabeth.metais@cnam.fr

Abstract

Natural Language Processing provides a very significant contribution to various research areas such as the e-health, e-business, education and antiterrorism. However, understanding the meaning, scope and usage of linguistic knowledge is a tedious task for a heterogeneity of potential users. Several approaches have been proposed to represent heterogeneous linguistic knowledge covering specific features of some languages. However, these approaches focused only on the data aspect of linguistic knowledge and neglected the processing one. Moreover, most of them do not support multilingualism and lack of powerful semantic representation and reasoning abilities. In this paper, we propose a multilingual linguistic domain ontology, called LingOnto, that represents and reasons about *(1)* linguistic data, *(2)* linguistic processing functionalities and *(3)* linguistic processing features. Our ontology supports English, French and Arabic languages and can be used by linguistically under-skilled users. In order to evaluate LingOnto and measure its efficiency, we applied it to a framework of identifying valid composition workflows of linguistic web services. Finally, we give the results of the carried-out experiments.

1 Introduction

Natural **L**anguage **P**rocessing (NLP) provides a very significant contribution to various research areas such as the e-health, e-business, education and antiterrorism. It has witnessed over the last years an acceleration in progress on a wide range of different applications such as sentiment analysis, knowledge mining and reasoning and search engine (Zhou et al., 2020).

However, understanding the meaning, scope and usage of linguistic knowledge is a tedious task. This complexity is mainly due to three reasons. First, NLP domain's potential users are heterogeneous and a considerable number of them are linguistically under-skilled. Second, the language is always changing, evolving, and adapting to its user's needs. For instance, words can acquire new meanings over time (e.g., the meaning of **"apple"** is a fruit but the meaning of **"Apple"** is a company). Finally, every language has its own specificities. Indeed, each language has its own structures and ways of interpreting. For example, Arabic has verbal and nominal sentences; but English has only verbal sentences.

In NLP, two types of approaches have been proposed to represent linguistic knowledge : *(1)* online registries-based approach such as the ISOcat registry[1], the SIL Glossary of linguistic terms[2] and the CLARIN Concept Registry (Schuurman et al., 2016) and *(2)* ontologies-based approach such as the General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, 2010), OntoTag ontologies (De Cea et al., 2004) and WordNet (Gangemi et al., 2002). However, these approaches present only linguistic data (e.g., word, noun, verb and adjective) and do not focus on modeling linguistic processing functionalities (e.g., tokenization, stemming and part of speech tagging) and linguistic processing features (e.g., treatment type, formalism and process-

[1] http://www.isocat.org/

[2] http://www-01.sil.org/linguistics/GlossaryOflinguisticTerms

ing level). Moreover, most of these approaches lack of powerful semantic representation and reasoning abilities. Finally, most of them do not support multilingualism.

In this paper, we propose a multilingual linguistic domain ontology, called LingOnto. It represents and reasons about linguistic processing functionalities, features and their linking with linguistic data rather than merely representing linguistic data. It supports English, French and Arabic languages. This ontology can be used by linguistically under-skilled users. In order to evaluate LingOnto, we choose to experiment it in the context of lingware engineering (Baklouti et al., 2010). Particularly, it is applied to a framework of identifying valid composition workflows of **Ling**uistic **W**eb **S**ervices (LingWS).

The current paper is organized as follows. Section 2 presents the related work. In Section 3, we detail the proposed ontology. Section 4 describes the carried-out experiments and the obtained results. Finally, Section 5 draws conclusions and future research directions.

2 Related Work

A considerable number of approaches for representing linguistic knowledge are available in the literature. We categorize them into two categories ***"online registers-based approach"*** and ***"ontologies-based approach"***.

2.1 Online Registers-Based Linguistic Knowledge Representation Approach

The SIL Glossary of linguistic terms (Eugene et al., 2004) provides information in the form of glossaries and bibliographies designed to support linguistic research. However, only 900 linguistic terms are covered in this glossary. Moreover, this latter supports only English and French languages. In addition, the usage of the SIL glossary is limited to search a defined term whose relation to other terms is unspecified. Consequently, it is not suitable for gaining comprehensive knowledge about a linguistic term in the NLP field.

In an attempt to provide a more comprehensive registry, (Kemps-Snijders et al., 2009) proposed ISOcat Data Category Registry (DCR). This registry aims at representing data categories at different linguistic levels such as syntactic, morphosyntactic, terminological, lexical and so on. However, ISOcat provides a wide range of different "views" and "groups" which makes navigating through it a very hard task. Moreover, it has no data model representing linguistic terminology in an interrelating holistic structure. Besides, its semantic structure provides definitions and very unspecific superordinate and subordinate concept relations such as "is_a" or "has_kinds".

Trying to define linguistic data in a stricter manner, the CLARIN Concept Registry, has taken over the work of ISOcat. Although, it still provides very limited structural and relational information.

2.2 Ontologies-Based Linguistic Knowledge Representation Approaches

(Farrar and Langendoen, 2010) proposed the GOLD ontology. It is based on the principles of knowledge engineering. It provides a taxonomy of nearly 600 linguistic concepts and formalizes 83 objects properties. These latter are very complex, specific and interrelate mostly only two concepts, which leaves the majority of the concepts unrelated. In addition, GOLD originates from the language documentation community and do not focus on NLP and corpus interoperability. Therefore, a number of data categories commonly assumed in NLP were not originally represented in GOLD. For example, gold:CommonNoun was added only recently following a suggestion by the author. Moreover, it do not cover all the linguistic knowledge. It defines only linguistic data and do not focus on modelling linguistic processing functionalities and features. A more fundamental problem is that this ontology is a very inefficient model for linguistic terminology. GOLD conflate both semantic and syntactic roles. The development of GOLD process has been stopped in 2010.

Focusing on the interoperability and language understanding, (De Cea et al., 2004) proposed OntoTag ontologies. These ontologies are applied to develop NLP applications on the basis of ontological representations of linguistic annotations. However, they consider only Iberian Romance languages (in particular Spanish). Moreover, they cover only linguistic data. They are not publicly available at the moment.

(Chiarcos and Sukhreva, 2015) proposed OLiA

ontologies which are closer related to the OntoTag ontologies. They introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology. However, these ontologies do not represent and reason about linguistic processing and features.

WordNet is a lexical resource that is rich enough to be considered alongside actual ontologies. It contains an extensive taxonomic and mereological structure which could be regarded as a kind of proto-ontology. However, it focuses on representing only some linguistic data. Moreover, WordNet object properties are not used in a consistent way, sometimes they are broken or present redundancy. (Gangemi et al., 2002) demonstrated that a substantial transform of WordNet's upper categories is needed in order to be used directly as an ontology.

It is worth mentioning that, in all of the above mentioned published works, the authors focused only on the data aspect of linguistic knowledge and neglected the processing one. Moreover, most of them do not support multilingualism and lack of powerful semantic representation and reasoning abilities.

3 LingOnto: a Multilingual Linguistic Domain Ontology

We propose a multilingual linguistic domain ontology, called LingOnto. It represents and reasons about linguistic knowledge. It handles *(1)* linguistic data *(2)* linguistic processing functionalities and *(3)* linguistic processing features. It supports English, French and Arabic languages. LingOnto can be used by linguistically under-skilled users. The current version of LingOnto includes 216 classes, 136 object properties and 326 **S**emantic **W**eb **R**ule **L**anguage (SWRL) rules. LingOnto is never frozen, which means that we can add other linguistic knowledge.

3.1 Overview

We are based on the design principles presented by Gruber (1995), which are objective criteria for guiding and evaluating ontology designs, such as clarity, coherence, minimal encoding bias and minimal ontological commitments. Following these principles, we define the following top-level concepts of our ontology as shown in Figure 1.

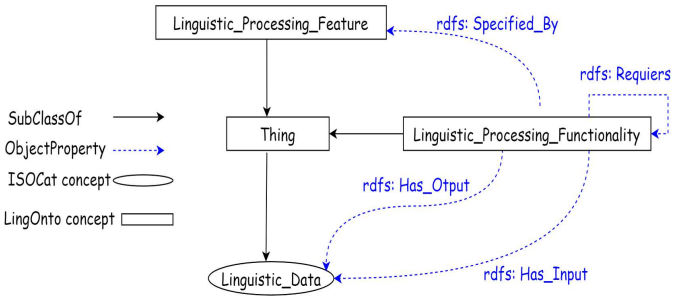


Figure 1: The top level concepts of LingOnto.

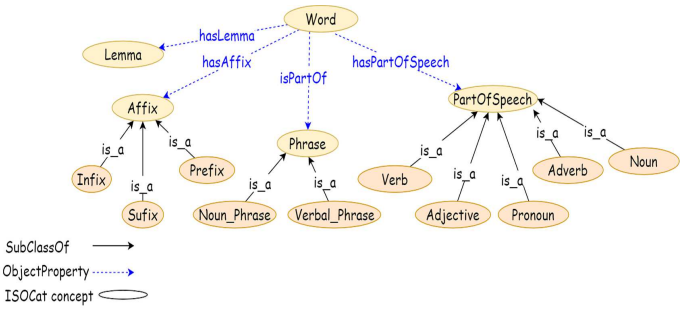


Figure 2: The Classification of some linguistic data.

3.2 Linguistic Data Classification

Referring to the ISOcat standard, we identify a set of linguistic data concepts. We choose ISOcat as it covers more terms of linguistic resources and linguistic levels compared to WordNet and GOLD. Figure 2 shows a part of LingOnto, illustrating the classification of some linguistic data. In fact, a word "Word" has a part of the speech "PartOfSpeech" which can be a noun "Noun", verb "Verb", adjective "Adjective", and so on. For this, we have defined an "is_a" object property between these latter and the "PartOfSpeech" class. Similarly, a word "Word" has an affix "Affix" which can be a prefix "Prefix", infix "Infix" or suffix "Suffix". In addition, a word "Word" is a part of a sentence "Phrase". As a consequence, an "isPartOf " object property has been established between the "Word" class (domain) and the "Phrase" class (range).

3.3 Linguistic Processing Functionality Classification

In order to identify a set of linguistic processing functionalities, a manageable selection of language

processing platforms (e.g., Grid (Ishida, 2011) and Weblicht (Hinrichs et al., 2010)) and NLP toolkits (e.g., Apache OpenNLP [3], Standford CoreNlP [4], FreeLing [5] and LingPipe[6]) is examined. The list is restricted to toolkits supporting English, French and Arabic languages. Moreover, we focus on linguistic processing functionalities, leaving out other functionalities provided by some of the toolkits. For example, FreeLing provides a variety of processors, including modules for performing tasks of statistical machine learning. In the following, we present some of the standard linguistic processors that we extract.

Linguistic Processors = {*Language Identifier, Sentence Splitter, Tokenizer, POS Tagger, Lemmatizer, Sense Tagger, Morphological Analyzer, Chunker, NE Recognizer, Coreference Resolver, Dependency Parser, Phrase Structure Parser, Speech Recognizer, TextTo-Speech Converter, Translator, Paraphraser*}

A linguistic processor implements often one or two linguistic processing functionalities. As a first example, "Morphological Analyzer" implements "Tokenization", "POS Tagging" and "Lemmatization" functionalities. As a second example, the "NE Recognizer" implements "Chunking" and "NE Classification" functionalities.

In the herein work, we intend to construct an ontology involving both lower and higher level processing functionalities in order to satisfy variable granularity user's need. Hence, we propose a set of linguistic processing functionalities as following:

Linguistic Processing Functionalities = {*Language Identification, Sentence Splitting, Tokenization, POS Tagging, Lemmatization, Sense Tagging, Morphological Analyzing, NE Recognizing, Chunking, NE Classification, Coreference Resolution, Dependency Parsing, Phrase Structure Parsing, Speech Recognition Text-To-Speech Conversion, Translation, Paraphrasing*}

After identifying a set of linguistic processing functionalities, we identify the relationships that may exist between them. There are a hierarchical interdependencies between the different linguistic processing functionalities (Hayashi and Narawa , 2012).

[3]http://opennlp.apache.org
[4]http://nlp.stanford.edu/software/ corenlp.shtml
[5]http://nlp.lsi.upc.edu/freeling/
[6]http://alias-i.com/lingpipe/

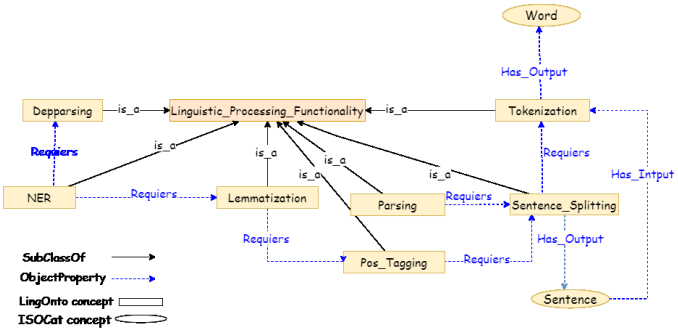


Figure 3: The Classification of some linguistic processing functionalities.

Indeed, a linguistic processing functionality used to perform analysis at one level may require, as input, the results of an analysis of a lower level. For example, syntactic analysis like parsing usually requires words to be clearly delineated and part-of-speech tagging or morphological analysis to be performed first. In the annotation process for example, the texts must be tokenized, their sentences clearly separated from each other and their morphological properties analyzed before starting the parsing functionality . Hence, we identify the object property "Requires". As shown in Figure 3, the "Tokenization" class is in relation with the "Sentence_Splitting" class through this object property. Moreover, each linguistic processing functionality manipulates various linguistic data as inputs and others as outputs. Hence, we propose the objects properties "Has_Input" and "Has_Output". For instance, as shown in Figure 3, the "Tokenization" class is in relation with the "Sentence" class through "Has_Input" object property. It is also in relation with the "Word" class through "Has_Output" object property. To reason about linguistic processing functionalities, we propose a set of SWRL rules. For example, the SWRL rules allow deducing the object property "Requires":

Linguistic_Processing_Functionality (A) $\wedge$ *Linguistic_Processing_Functionality (B)* $\wedge$ *Linguistic_data (I)* $\wedge$ *Has_Output (A,I)* $\wedge$ *Has_Input (B,I)* $\rightarrow$ *Requires (A, B).*

Linguistic_Processing_Functionality (A) $\wedge$ *Linguistic_Processing_Functionality (B)* $\wedge$ *Linguistic_Processing_Functionality (C)* $\wedge$ *Requires (B,A)* $\wedge$ *Requires (C,B)* $\rightarrow$ *Requires (C,A).*

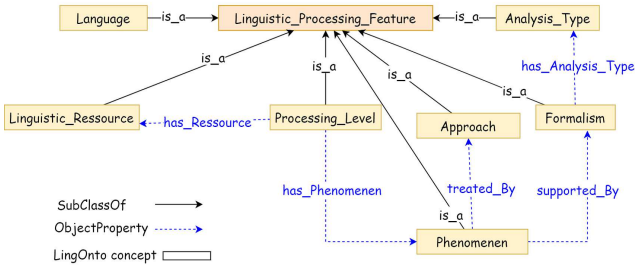


Figure 4: The Classification of some linguistic processing features.

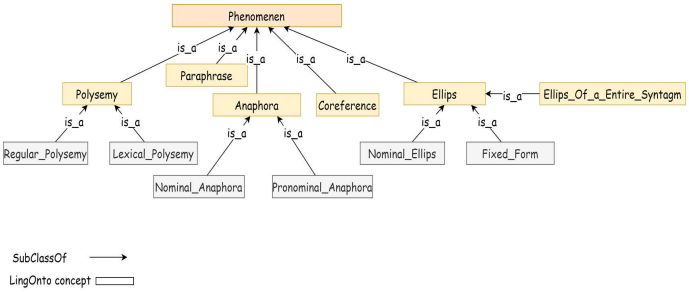


Figure 5: The Classification of some linguistic phenomenon.

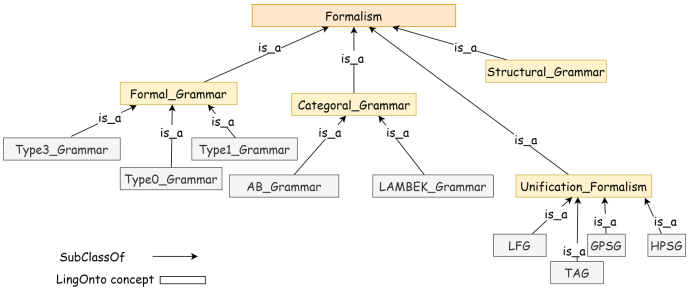


Figure 6: The Classification of some linguistic formalism.

3.4 Linguistic Processing Feature Classification

Each linguistic processing functionality is characterized by a set of linguistic processing features. We propose that the "Linguistic_Processing_Feature" class includes the following sub-classes:

- "Processing level": it represents the processing level of a linguistic processing functionality. In NLP, we distinguish mainly four processing levels: lexical, morphological, syntactic, and semantic. Each processing level is characterized by both its resources (e.g., dictionaries, tree bank and corpus) and phenomena (e.g., ellipsis, anaphora and accord). For that, we propose the object properties "has_Resource" and "has_Phenomenon".

- "Phenomenon": it is the linguistic phenomenon treated by a processing level. It has the "refined_into" object property, since each phenomenon has its subPhenomena. For example, in the ellipsis phenomenon we distinguish the nominal ellipsis (the omission of the essential part of a nominal phrase: the head) and an ellipsis of a whole phrase (e.g., subject ellipsis, verb ellipsis, both verb and complement ellipsis). The "Phenomenon" class also has a "treated_By" object property in relation with the "Approach" class and "supported_By" object property with the "Formalism" class.

- "Approach": it is the linguistic approach treated by a phenomenon. It can be a statistical, linguistic or hybrid approach (linguistic and statistical).

- "Formalism": it is the linguistic formalism that supports a phenomenon. There are several types of formalisms such as HPSG and LFG for syntactic grammars (Kahane, 2006). A formalism can concern a main phenomenon or a subphrenomen. Also, it can have a type of analysis. So, we propose the "has_Analysis_Type" object property with the "Analysis type" class.

- "Analysis type": it is the type of analysis that characterizes a linguistic formalism, namely, bottom-up analysis, top-down analysis, surface analysis and so on.

- "Language": it is important to learn about the specificity and the structure of each language to deal with its complexity. For example, Arabic language is a very rich language with complex morphology, with different and difficult structure than other languages. LingOnto focus on English, French and Arabic languages.

4 Experimentation

We apply the proposed ontology to a framework of identifying valid composition workflows of LingWS

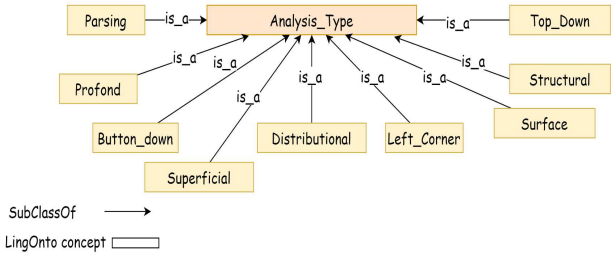


Figure 7: The Classification of some linguistic analysis type.

(Neji et al., 2018). Then, we evaluate its efficiency in the context of the last one.

4.1 Application to LingWS Composition Workflows Identification Framework

The LingOnto is applied to a framework of identifying valid composition workflows of LingWS. It targets under-skilled users in the lingware engineering area.

First, the user generates a dynamic ontological view from LingOnto based on a set of selection criteria. This is done thanks to a user friendly ontology visualization tool called LingGraph (Neji et al., 2019). The choice of one or more criteria is made to show only the components corresponding to the user's need. For instance, Figure 8 shows LingOnto's components representing the different linguistic processing functionalities related to the morphological level of English language. Second, the user starts the identification of a workflow of linguistic processing functionalities based on the generated ontological view. If the user selects a functionality, LingOnto proposes a set of possible functionalities choices that can be added to the workflow. This step is done thanks to the aforementioned LingOnto SWRL rules. For instance, as shown in Figure 8, a Part-of-Speech Tagging functionality can be added to the workflow only after a Tokenization functionality. Finally, the corresponding LingWS(s) to each selected linguistic processing functionality is discovered taking into account the set of linguistic processing features presented in LingOnto. Indeed, the discovery process performs a matching between the linguistic processing features of each selected linguistic processing functionality and the description of each required LingWS. This step explores the LingWS registry (Baklouti et al., 2015).

4.2 Evaluation

A total of 30 users were recruited to participate in this evaluation study. They are researcher members of NLP Research Group of MIRACL laboratory (Tunisia, Sfax) and CEDRIC laboratory (France, Paris). The selected users have the same NLP and languages competences. Before beginning the experiment, they were asked to fill a pre-questionnaire about their prior knowledge and expertise in NLP research field and languages. These users are equally allocated into three groups where each group focus on only one language i.e., English, French or Arabic. Each user identified a set of composition workflows from LingOnto related to the morphological level. An NLP domain expert participated in this evaluation in order to identify the number of valid possible composition workflows corresponding to the morphological level of each language.

For each group of users working on a given language, we note:

- All_User_W : the total number of composition workflows identified by all the users of the group without redundancy.
- V_All_User_W : the total number of **valid** composition workflows identified by all the users of the group without redundancy.
- User_W : the total number of composition workflows identified by a given user of the group.
- V_User_W : the total number of **valid** composition workflows identified by a given user of the group.
- Exp_W : the total number of **valid** composition workflows identified by the NLP domain expert for the concerned language.

We use the Recall and Precision evaluation metric as follow:

- The Recall associated to a given language R_L = *(V_All_User_W / Exp_W)*.
- The Precision associated to a given language P_L = *(V_All_User_W / All_User_W)*.

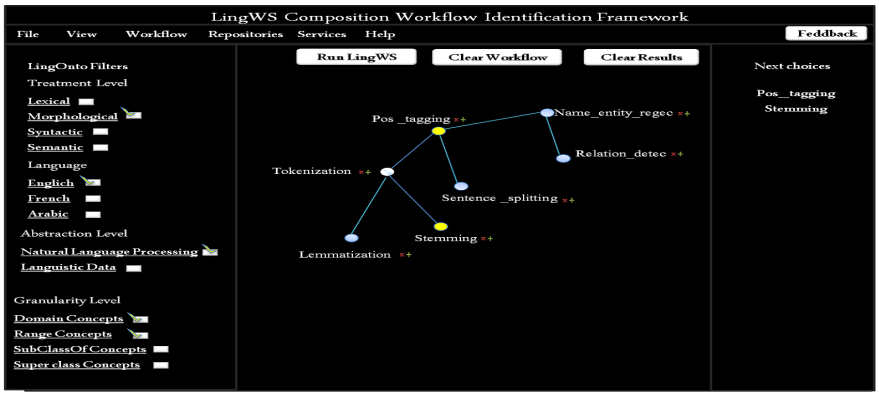


Figure 8: Screenshot of LingWS composition workflows identification framework showing a part of LingOnto.

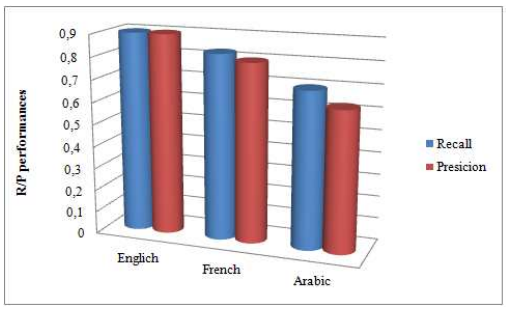


Figure 9: R/P performances of the three languages.

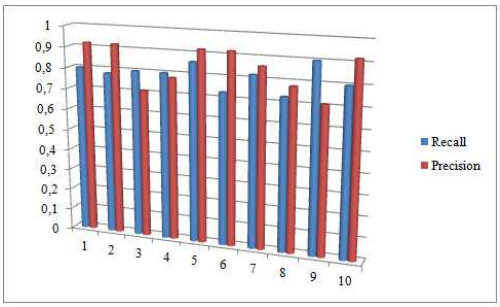


Figure 10: R/P performances of English users.

- The Recall associated to a given user R_U = *(V_User_W / Exp_W)*.
- The Precision associated to a given user P_U = *(V_User_W / User_W)*.

As shown in Figure 9, the mean of the R/P performances measures indicates that our ontology is efficient in identifying valid composition workflows. Besides, the overall mean of the precision associated to the English and French language is better than the overall mean of the precision associated to the Arabic language. This gap of R/P performances is explained by the fact that the Arabic language is characterised by a complicated morphology compared to English and French languages. Figure 10 shows the R/P performances of the English language users. We note that these performances are not the same for all the users. While the users engaged for the English language have similar levels of competences, then the difference in their performance could reflect a complexity in exploiting the visualization tool LingGraph.

5 Conclusion and Future Work

In this paper, we proposed a multilingual linguistic domain ontology, called LingOnto, for representing and reasoning about linguistic knowledge. It helps linguistically under-skilled users in understanding the meaning, scope and usage of linguistic knowledge.

At the beginning, we elaborated a state of the art focusing on approaches that are proposed to represent linguistic knowledge. This study showed that existing works consider only linguistic data.To the best of our knowledge, there is no approach that allows representing and reasoning about linguistic processing functionalities and features and their linking with linguistic data. Moreover, most of them do not support multilingualism. Compared to related work, LingOnto allows representing and reasoning about linguistic data and linguistic processing functionalities and features. It supports English, French and Arabic languages. We applied LingOnto to a framework of identifying valid composition work-

flows of LingWS.

Currently, we are applying LingOnto to a smart memory prosthesis for Alzheimers patients called CAPTAIN MEMO. It is proposed in the context of VIVA project *(Vivre a Paris avec Alzheimer en 2030 grâce aux nouvelles technologies)*. LingOnto is used to identify composition workflows of a nature language-interrogation system that aims to facilitate the communication between the prosthesis and the Alzheimer's patients.

We plan to allow the LingOnto ontology to be referenced by the Linked Open Vocabularies (LOV) platform. Moreover, we plan to exploit the NLP domain expert's feedback to improve the proposed ontology.

References

Baklouti, N., Bouaziz, S., Gargouri, B., Aloulou, C., and Jmael, M. 2010. *Towards the reuse of lingware systems: a proposed approach with a practical experiment*. International Conference on Information Integration and Web-based Applications and Services, pages 566572.

Baklouti, N., Gargouri, B., and Jmaiel, M. 2015. *Semantic-based approach to improve the description and the discovery of linguistic web services* . Engineering Applications of Artificial Intelligence, 46, 154-165.

Chiarcos, C. and Sukhreva, M. 2015 . *OLIA - Ontologies of Linguistic Annotation*. Semantic Web Journal, 518:379386.

De Cea, G. A., de Mon, I. A., Gomez-Perez, A., and Pareja-Lora, A. 2004. *Ontotags linguistic ontologies: improving semantic web annotations for a better language understanding in machines*. International Conference on Information Technology: Coding and Computing(ITCC). (Vol. 2, pp. 124-128). IEEE.

Eugene E. Loos, Susan Anderson, Dwight H. Day, Jr., Jordan Paul. 2004. *Glossary of linguistic terms, volume 29*. SIL International.

Fellbaum, C. 2012. *WordNet*. The encyclopedia of applied linguistics.

Farrar, S., and Langendoen, D. T. 2010. *An owl-dl implementation of gold: An ontology for the semantic web* . Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology.

Gangemi A., Guarino, N., Oltramari, A. and Borgo, S. 2002. *Cleaning-up wordnets top-level*. Proceedings of the 1st International WordNet Conference.

Gruber, T. R. 1995. *Toward principles for the design of ontologies used for knowledge sharing?* . International journal of human-computer studies, 43(5-6), 907-928.

Hayashi, Y., and Narawa, C. 2012 . *Classifying Standard Linguistic Processing Functionalities based on Fundamental Data Operation Types*. LREC (pp. 1169-1173).

Hinrichs, M., Zastrow, T., and Hinrichs, E. W. 2010. *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. International Conference on Language Resources and Evaluation, pages 489493,Valletta, Malta.

Ishida, T. (Ed.). 2011. *The language grid: Service-oriented collective intelligence for language resource interoperability*. Springer Science and Business Media.

Kahane, S. . *Polarized unification grammars* . International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 137-144.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. 2009. *Isocat: Remodeling metadata for language resources*. International Journal of Metadata, Semantics and Ontologies, 4(4), 261-276.

Neji, M., Gargouri, B., and Jmaiel, M.. 2018. *A semantic approach for constructing valid composition scenarios of linguistic Web services*. Procedia Computer Science, 126, 685-694.

Neji, M., Ghorbel, F., and Gargouri, B. 2019. *A smart search-based ontology visualization tool using sparql patterns*. International Conference on Knowledge Science, Engineering and Management (pp. 33-44). Springer, Cham.

Schuurman, I., Windhouwer, M., Ohren, O., and Zeman, D. 2016. *Clarin concept registry: the new semantic registry*. Selected Papers from the CLARIN Annual Conference 2015, pages 6270.

Zhou, M., Duan, N., Liu, S., and Shum, H. Y.. 2020. *Progress in neural nlp: Modeling, learning, and reasoning*. Engineering, 6(3), 275-290.

Iterative Multilingual Neural Machine Translation for Less-Common and Zero-Resource Language Pairs

Minh-Thuan Nguyen, Phuong-Thai Nguyen, Van-Vinh Nguyen, Minh-Cong Nguyen Hoang
Department of Computer Science, University of Engineering and Technology, VNU Hanoi
{thuannm, thainp, vinhnv}@vnu.edu.vn
minhcongnguyen1508@gmail.com

Abstract

Research on providing machine translation systems for unseen language pairs is gaining increasing attention in recent years. However, the quality of their systems is poor for most language pairs, especially for less-common pairs such as Khmer-Vietnamese. In this paper, we show a simple iterative training-generating-filtering-training process that utilizes all available pivot parallel data to generate synthetic data for unseen directions. In addition, we propose a filtering method based on word alignments and the longest parallel phrase to filter out noise sentence pairs in the synthetic data. Experiment results on zero-shot Khmer→Vietnamese and Indonesian→Vietnamese directions show that our proposed model outperforms some strong baselines and achieves a promising result under the zero-resource condition on ALT benchmarks. Besides, the results also indicate that our model can easily improve their quality with a small amount of real parallel data.

1 Introduction

Neural Machine Translation (NMT) has recently achieved impressive performance on high-resource language pairs which have large amounts of parallel training data (Wu et al., 2016) (Vaswani et al., 2017). However, these systems still work poorly when the parallel data is low or unavailable. Research on zero-resource language pairs is gaining much attention in recent years, and it has been found to use pivot language, zero-shot NMT, or zero-resource NMT approaches to deal with the translation of unseen language pairs.

In pivot language approaches, sentences are first translated from the source language into the pivot language through a source-pivot system, and then from the pivot language into the target language by using a pivot-target system. Although this simple process has shown strong translation performance (Johnson et al., 2017), it has a few limitations. The pivoting translation process at least doubles decoding time during inference because more than one pivot language may be required to translate from the source to the target language. Additionally, translation errors compound in a pipeline.

Zero-shot NMT approaches are inspirited from multilingual NMT (multi-NMT) systems that use only one encoder and one decoder to represent multiple languages in the same vector space, hence it should be possible to take advantage of data from high-resource language pairs to improve the translation of low-resource language pairs. (Ha et al., 2016; Johnson et al., 2017) showed that the zero-shot systems are able to generate reasonable output at the target language by adding the desired output language's language tag at the beginning of the source sentence. Note that there is no direct parallel data between the source and target languages during training. However, the performance of these approaches is still poor when the source and target languages are unrelated or the observed language pairs are not enough to capture the relation of unseen language pairs.

Similar to the above approaches, zero-resource NMT approaches do not use any direct source-target parallel corpus, but the approaches focus on generating pseudo-parallel corpus by using back-translation

to translate sentences in the pivot language of the pivot-target parallel corpus to the source language (Lakew et al., 2017; Gu et al., 2019). One of the main limitations of these approaches is that the source between training and testing scenarios are different since the source in training is synthetic. However, the approaches still outperform pivot language and zero-shot NMT approaches because they can potentially utilize all available parallel and monolingual corpus (Currey and Heafield, 2019).

In this work, our main contributions are (1) improving the quality of zero-resource NMT by introducing a simple iterative *training-generating-filtering-training* process and (2) proposing a noise filtering method. Especially, we evaluate our approach on less-common and low-resource language pairs such as Khmer-Vietnamese. In this scenario, source-pivot (Khmer-English) and pivot-target (English-Vietnamese) pairs are also low-resource (pivot is often English). Our approach starts from a multilingual NMT system that is trained on source-pivot and pivot-language pairs, the system then generates source-target synthetic corpus by back-translating the pivot side of the pivot-target corpus to the source language. Next, We filter out poor translations in the generated translations by applying our proposed data filtering method based on word alignments and the longest parallel phrase. After that, the multilingual NMT system is continuously trained on both the filtered synthesis data and the original training data, we repeat this *training-generating-filtering-training* cycle for a few iterations. As a result, our experiments showed that by adding the filtered synthetic corpus, our model outperformed the pivot, zero-shot, and zero-resource baselines over zero-shot Khmer→Vietnamese and Indonesian→Vietnamese directions on the Asian Language Treebank (ALT) Parallel Corpus (Riza et al., 2016). Moreover, the experiment results indicate that our model can easily improve their quality with a small amount of real parallel data.

The rest of this paper is organized as follows. We first review relevant works on translation for zero-resource language pairs in Section 2, then introduce some background and related formulas in Section 3. Next, we show our approach in Section 4. After that, we illustrate our experiments and results in Section 5. Finally, our conclusion is presented in Section 6.

2 Related Work

Training a machine translation system for translating unseen language pairs has received much interest from researchers in recent years. This section discusses relevant works on zero-shot and zero-resource NMT, which are related to our approach.

Zero-shot NMT

(Ha et al., 2016; Johnson et al., 2017) showed that using a single NMT can learn to translate between language pairs it has never seen during training (zero-shot translation). Their solution does not require any changes to the traditional NMT model architecture. Instead, they add an artificial token at the beginning of the source sentence to specify the required target language. Although this approach illustrated promising results for some untrained language pairs such as from Portuguese to Spanish, its performance is often not good enough to be useful and lags behind pivoting. In our work, we use this system as an initial multi-NMT system.

(Arivazhagan et al., 2019) pointed out that the success of zero-shot translation depends on the ability of the model to capture language invariant features for cross-lingual transfer. Therefore, they proposed two classes of auxiliary losses to align the source and pivot vector spaces. The first minimizes the discrepancy between the feature distributions by minimizing a domain adversarial loss (Gani et al., 2015) that trains a discriminator to distinguish between different encoder languages using representations from an adversarial encoder. The second takes advantage of available parallel data to enforce alignment between the source and the pivot language at the instance level. However, this approach does not work for less-common language pairs such as Khmer-Vietnamese since the size of multilingual training data including source-pivot and pivot-target is low, so it is not enough to capture the language invariant features.

Zero-resource NMT

(Lakew et al., 2017) used a multilingual NMT system to generate zero-shot translations on some portion of the training data, then re-start the training process on both the multilingual data and the generated translations. By adding the synthetic cor-

pus, the model can alleviate the spurious correlation problem. This work is similar to our work but they did not filter out noise sentence pairs in the synthetic corpus.

(Currey and Heafield, 2019) augmented zero-resource NMT with monolingual data from the pivot language. The authors pointed out that the pivot language is often high-resource language and more high-quality than the monolingual source or target language (pivot language is often English), so leveraging the monolingual pivot language data is worthwhile to enhance the quality of zero-resource NMT systems.

3 Background

3.1 Neural Machine Translation

The standard NMT architecture contains an encoder, a decoder and an attention-mechanism, which are trained with maximum likelihood in an end-to-end system (Bahdanau et al., 2014). Assume the source sentence and its translation are $x = \{x_1, ..., x_{T_x}\}$ and $y = \{y_1, ..., y_{T_y}\}$ respectively.

Encoder is a bidirectional Recurrent Neural Network (RNN) (Schuster and Paliwal, 1997) that encodes the source sentence into a sequence of hidden state vectors, the hidden state vector of word x_i is $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$, where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are forward and backward hidden state respectively.

$$\overrightarrow{h_i} = f(e_{x_i}, \overrightarrow{h}_{i-1}) \tag{1}$$

$$\overleftarrow{h_i} = f(e_{x_i}, \overleftarrow{h}_{i+1}) \tag{2}$$

Note that e_{x_i} is the vector of word x_i, f is a nonlinear function such as Long Short-term Memory (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (Cho et al., 2014).

Attention is a mechanism used to compute a context vector by searching through the source sentence at each decoding step (Bahdanau et al., 2014). At the j-th step, the score between the target word y_j and the i-th source word is computed and normalized as below:

$$e_{i,j} = v_a^T \tanh(W_a s_{j-1} + U_a h_i) \tag{3}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{i'=1}^{T_x} \exp(e_{i'j})} \tag{4}$$

The context vector c_j is computed as a weighted sum of all source hidden states:

$$c_j = \sum_{i=1}^{T_x} \alpha_{ij} h_i \tag{5}$$

Decoder is a unidirectional RNN which uses the representation of the encoder and the context vector to predict words in the target language. At the j-th step, the target hidden state s_j is computed by:

$$s_j = f(e_{y_{j-1}}, s_{j-1}, c_j) \tag{6}$$

Given the previous predicted words $y_{<j} = \{y_1, ..., y_{j-1}\}$, the context vector c_j and the target hidden state s_j, the decoder is trained to predict the next word y_j as follows:

$$p(y_j | y_{<j}, s_j, c_j) = \text{softmax}(W_o t_j) \tag{7}$$

$$t_j = g(e_{y_{j-1}}, c_j, s_j) \tag{8}$$

where g is a nonlinear function, W_o is used to output a vocabulary-sized vector.

3.2 Multilingual NMT

(Ha et al., 2016; Johnson et al., 2017) indicated a simple approach to use a standard NMT system to translate between multiple languages. This system leverages the knowledge from translation between multiple languages and is referred to as a multilingual NMT system. In order to make use of multilingual data containing multiple language pairs into the standard NMT system, authors proposed one simple modification to the input data, which is to add an artificial token at the beginning of the input sentence to indicate the desired target language. After adding the token to the input data, over-sampling or under-sampling techniques are applied to balance the ratio of language pairs in the multilingual data, and the model is trained with all the multilingual data at once. Besides, a shared wordpiece model (Sennrich et al., 2015) across all the source and target data is used to address the problem of translation of unknown words and limitation of the vocabulary for computational efficiency, usually with 32,000 word pieces.

4 Approach

This paper concentrates on improving the quality of zero-resource NMT between two languages X and Y given a pivot language Z. We assume that we have $X \leftrightarrow Z$ and $Z \leftrightarrow Y$ parallel data, but no direct $X \leftrightarrow Y$ data. Algorithm 1 represents our proposed training process. Notably, our experiments focus on less-common and low-resource language pairs such as Khmer-Vietnamese, Indonesian-Vietnamese, so the amount of $X \leftrightarrow Z$ and $Z \leftrightarrow Y$ parallel data is quite small. Therefore, in order to build a good initial multi-NMT model, the first step of our work is to augment the multilingual training data that is shown in Section 4.1. Take a look at the Algorithm 1, given an initial training data D including $X \leftrightarrow Y$ and $Y \leftrightarrow Z$ parallel data, our training process contains four main steps which are iterated for multiple times.

Algorithm 1: Iterative Multi-NMT with Data Filtering Procedure

1: D = (X $\leftrightarrow$ Z, Z $\leftrightarrow$ Y)
2: **repeat**
3: Multi-NMT $\leftarrow$ training using dataset D
4: **for each** Z in (Z $\leftrightarrow$ Y) **do**
5: X* $\leftarrow$ Multi-NMT(Z), generating
6: **end for**
7: S $\leftarrow$ (X* $\leftrightarrow$ Y), synthetic data
8: F $\leftarrow$ Filter(S), filtering synthetic data
9: D $\leftarrow$ D $\cup$ F
10: **until** Multi-NMT converges

Figure 1: Algorithm of the proposed approach using iterative multi-NMT with data filtering.

- Step 1 (line 3): Train a multilingual NMT by using the training dataset D.
- Step 2 (line 4, 5, 6): Generate $(X^* \rightarrow Y)$ synthetic parallel data by using the trained multi-NMT model to translate sentences from pivot language Z in $(Z \leftrightarrow Y)$ to language X. We can obtain more synthetic data $(X \leftrightarrow Y)$ by translating sentences from pivot language Z in $(X \leftrightarrow Z)$ to language Y.
- Step 3 (line 8): Filter the synthetic data to eliminate bad parallel sentence pairs by using data selection techniques (See Section 4.2).
- Step 4 (line 9): Expand the multilingual training data by adding the filtered synthetic data F to the original training data D.

In our *training-generating-filtering-training* cycle, new synthetic $X \leftrightarrow Y$ data is generated at each iteration. We expect that by adding this synthetic data, the multi-NMT model not only improves the translation of zero-shot directions between X and Y but also boosts other directions such as between X and Z, Y and Z. Therefore, round after round, we can build a better multi-NMT system with the synthetic data. Use this better system in order to generate new synthetic data, then use this data with the original training data to build an even better system. Finally, this cycle continues until the model converges.

4.1 Data Augmentation

As mentioned above, if the amount of multilingual training data is too small, the multi-NMT system is unable to learn to translate between zero-shot directions. Hence, in our work, to augment the parallel data for $(X \leftrightarrow Z)$ and $(Z \leftrightarrow Y)$, we leverage monolingual data in both target and source side by using back-translation (Sennrich et al., 2016) and self-training (Zhang and Zong, 2016). Given a parallel data $(X \leftrightarrow Z)$ and monolingual data M_X, M_Z in language X, Z respectively, we denote by $\overrightarrow{f}$ and $\overleftarrow{g}$ the forward (from X to Z) and the backward (from Z to X) NMT systems.

Back-translation is a popular data augmentation method utilizing target side monolingual data. To perform back-translation, given the parallel data $(X \leftrightarrow Z)$, a base backward NMT system $\overleftarrow{g}$ is trained and use it to translate M_Z to language X, denoted by $\overleftarrow{g}(M_Z)$. The original parallel data $(X \leftrightarrow Z)$ is then concatenated with the back-translated data $(\overleftarrow{g}(M_Z) \leftrightarrow M_Z)$ to obtain a new training data.

Self-Training augments the original training data by first training a base forward NMT system $\overrightarrow{f}$ on $(X \leftrightarrow Z)$ data, then use this trained model to translate M_X to language Z, denoted by $\overrightarrow{f}(M_X)$. The new synthetic data $(M_X \leftrightarrow \overrightarrow{f}(M_X))$ is also combined with the original training data to obtain a new training dataset.

In our work, we augment parallel data by using both these two methods because they are complementary to each other. The original training data

is combined with back-translated and self-trained data to obtained the augmented parallel data, $(X \leftrightarrow Z) \cup (\overleftarrow{g}(M_Z) \leftrightarrow M_Z) \cup (M_X \leftrightarrow \overrightarrow{f}(M_X))$.

4.2 Data Filtering

Combining synthetic data with the multilingual training data is a simple and effective way to boost the quality of zero-shot directions in zero-shot NMT and zero-resource NMT systems (Lakew et al., 2017; Currey and Heafield, 2019). However, the synthetic data potentially contains a lot of noise—translation errors, since it is often generated by using back-translation or self-training. Therefore, in this section, we show our proposed method to filter noise sentence pairs from synthetic data based on sentence semantic similarity. As described in Section 4, a synthetic sentence pair (x_i, y_i) is generated by translating z_i in $(Z \leftrightarrow Y)$ data to x_i. We consider that (x_i, y_i) is good synthetic sentence pair if x_i is both semantically similar to y_i and z_i. A semantic score for each synthetic sentence x_i is computed as below:

$$\text{score}(x_i) = \frac{\text{sim}(x_i, y_i) + \text{sim}(x_i, z_i)}{2} \quad (9)$$

where $sim(x_i, y_i)$ and $sim(x_i, z_i)$ are the semantic similarity of (x_i, y_i) and (x_i, z_i) sentence pair respectively.

To compute the semantic similarity of two sentences in different languages, (Xu et al., 2019) relies on cosine similarities of sentence embedding vectors in a common vector space such as bilingual word embedding (Luong et al., 2015b). Our method first also embeds words in different languages into a common vector space as work in (Conneau et al., 2017), then calculate the sentence similarity based on *word alignment scores* and the *longest parallel phrase* of the candidate sentence pairs. In order to acquire word alignments of a sentence pair (x, y), we iterate sentence x from left to right and greedily align each word in x to the most similar word in y which was not already aligned. For measuring the similarity of words we use cosine similarity of word embeddings. Afterward, given a set of word alignments A, we can easily extract parallel phrases of (x, y) by using the phrase extraction algorithm in the Statistical Machine Translation System (Koehn et al., 2003). Finally, the semantic similarity score of the sentence pair (x, y) is computed by averaging word alignment scores and weighting it with the ratio of the length of the longest parallel phrase p and the length of the sentence x as follows:

$$\text{sim}(x, y) = \frac{|p|}{|x|} \times \frac{\sum_{a \subset A} \text{score}(a)}{|A|} \quad (10)$$

where $|p|$ and $|x|$ are the length of longest parallel phrase and sentence x respectively, $|A|$ is the number of word alignments, a is a word alignment candidate and $score(a)$ is word alignment score that is computed by using cosine similarity of two words in the alignment a.

5 Experiments

5.1 Dataset

In this work, we evaluate our approach on zero-resource Khmer-Vietnamese (km-vi) and Indonesian-Vietnamese (id-vi) language pairs with English is the pivot language. The parallel datasets for Khmer-English (km-en) and Indonesian-English (id-en) are from the Asian Language Treebank (ALT) Parallel Corpus (Riza et al., 2016) and for English-Vietnamese is from the UET dataset (Vu Huy et al., 2013) (see Table 1 for details). All testing datasets are from the ALT corpus with size of 1,018 sentences. In addition, we used monolingual data released in Wikipedia[1] for Vietnamese, English and Indonesia and data from WMT2020[2] for Khmer. After de-duplication and removing too short (<5 tokens) or too long (>100 tokens) sentences, we obtained approximately 11 million, 5 million, 2 million and 3 million unique sentences for English, Vietnamese, Khmer, and Indonesian respectively. Moreover, as mentioned in Section 4.1, before training models, we augmented the multilingual training data by using back-translation and self-training. In order to choose the right ratio between real and synthetic parallel data, we experimented on different real-to-synthetic ratios. We found that 1:4 real-to-synthetic ratio is the best ratio for both Khmer-English and Indonesian-English pairs as shown in Table 2. Finally, we acquired the

[1] https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[2] http://www.statmt.org/wmt20/parallel-corpus-filtering.html

Direction	Training	
	real	real+BT+ST
Khmer-English	18,088	162,792
English-Vietnamese	233,000	-
Indonesian-English	18,088	162,792

Table 1: Number of sentences used for training. $real$ column show the size of original data and $real+BT+ST$ column illustrates the size of the augmented data.

real:syntheic ratio	km → en		id → en	
	BT	ST	BT	ST
1:0	14.19	13.7	21.57	20.52
1:1	15.32	15.72	22.26	21.18
1:2	16.87	17.58	24.06	22.10
1:3	17.25	18.21	24.37	21.99
1:4	**18.3**	**18.62**	**24.79**	**22.64**
1:5	18.1	17.93	24.02	21.60
1:6	17.54	17.01	23.70	21.42

Table 2: Experiment results on BLEU score to choose the right real:synthetic ratios for Khmer→English (km→en) and Indonesian→English (id→en) using back-translation (BT) and self-training (ST).

final augmented data by combining the original data with back-translated and self-trained data as shown in Table 1. Note that, to prevent imbalances between language pairs in the multilingual training data, we did not augment for the English-Vietnamese pair since the size of this pair is much larger other pairs.

5.2 Preprocessing

To learn a shared vocabulary for training multi-NMT, we used SentencePiece (Kudo and Richardson, 2018) with size 32,000 over the combined English, Vietnamese, Khmer, and Indonesian monolingual data. Besides, we added target language tags at both the beginning and end of the source sentences in the multilingual training data.

The multilingual word embedding model used in our filtering method was acquired by using the unsupervised method in MUSE library[3]. The word embeddings for English, Vietnamese, Khmer, and Indonesian are trained with fastText toolkit[4] on corresponding monolingual data.

[3]https://github.com/facebookresearch/MUSE

[4]https://fasttext.cc/

All translation results shown in our work were computed in terms of BLEU score (Papineni et al., 2002) measured with *multi-bleu.perl* script[5]

5.3 Models

All models in our experiments are based on the encoder-decoder with attention architecture (Luong et al., 2015a). We used OpenNMT-py[6] to run all experiments with the configuration as follows. We used the Gradient Descent optimizer with a learning rate of 1.0 that decayed exponentially in the last 80% of the training duration, training batch is 64, maximum sentence length is 100, beam width is 10, label smoothing is 0.2, dropout is 0.3 and is applied on top of various process, all models variables are initialized uniformly in range (-0.1, 0.1).

In this paper, we evaluate our proposed method on two direct (zero-shot) translations, Khmer → Vietnamese (km → vi) and Indonesian → Vietnamese (id → vi). Notably, the setting of experiments for these 2 directions is the same, so in the following, we only describe experiments for evaluating the Khmer-Vietnamese language pair.

Firstly, We compare our models to three baselines as follows:

- **zero-shot NMT**: This model is trained on the Khmer ↔ English and English ↔ Vietnamese parallel data.
- **zero-resource NMT**: This model is trained on the synthetic data Khmer ↔ Vietnamese created by using the above zero-shot NMT model to translate English sentences in (English ↔ Vietnamese) to Khmer sentences.
- **pivot language**: use the above zero-shot NMT to translate Khmer sentences into English then from English to Vietnamese.

Our proposed models are designated as below:

- **Iterative multi-NMT**: This model is trained by iterating *training-generating-training* schema for several rounds. We use the above zero-shot NMT as an initial multi-NMT model for this training process.

[5]https://github.com/moses-smt/mosesdecoder

[6]https://github.com/OpenNMT/OpenNMT-py

- **Iterative multi-NMT + Xu's data filtering**: This model is trained by iterating *training-generating-filtering-training* schema for several rounds as shown in Section 4. We also use the above zero-shot NMT as an initial multi-NMT model and use the method of (Xu et al., 2019) in the data filtering step.
- **Iterative multi-NMT + our data filtering**: This model is trained by using the *training-generating-filtering-training* process and our proposed method for data filtering.

Note that, in the last two models, we use a similarity threshold of 0.4 achieved the best result (see Table 4 for details), to filter out poor synthetic sentence pairs.

5.4 Results and Analysis

Table 3 shows our results for the $km \rightarrow vi$ and $id \rightarrow vi$ zero-resource translation experiments. Experiments (1), (2), and (3) indicate the performance of the three baseline models. It can be seen that zero-shot NMT performed the worst result while the two other models illustrate promising results. The explanation for this results is that the amount of multilingual training data is not enough for enabling zero-shot translation on the multi-NMT system. Experiment (4) outperforms all three baseline models since it is benefit from both zero-shot and zero-resource NMT system. In addition, Experiments (5) and (6) show the effect of our *training-generating-filtering-training* process. By eliminating poor synthetic sentence pairs before re-training, the systems perform better results. Especially, the results on experiment (5) and (6) indicate that our proposed filtering method is more effective than the method of (Xu et al., 2019) for filtering noises in synthetic data.

Table 4 shows the effect of different filtering threshold on translation performance. All models are trained similar to the model *Iterative multi-NMT + our data filtering*, the only different is the filtering threshold to eliminate poor sentence pairs. Notably, a threshold of 0.0 means that all synthetic data is kept to re-train in the next iteration. The results illustrate that the threshold of 0.4 achieved the best result, outperforming the baseline (threshold is 0.0) by +1.64 and +1.69 BLEU for $km \rightarrow vi$ and $id \rightarrow vi$ directions respectively.

	model	km→vi	id→vi
(1)	zero-shot NMT	3.43	6.75
(2)	zero-resource NMT	13.82	14.26
(3)	pivot language	12.59	12.99
(4)	Iterative multi-NMT	15.23	17.24
(5)	Iterative multi-NMT + Xu's data filtering	16.02	18.51
(6)	Iterative multi-NMT + our data filtering	**16.87**	**18.93**

Table 3: BLEU scores for our proposed models compared with strong baselines.

Threshold	km → vi	id → vi
0.0	15.23	17.24
0.1	15.81	17.75
0.2	16.02	17.96
0.3	16.25	18.29
0.4	**16.87 (+1.64)**	**18.93 (+1.69)**
0.5	16.62	18.58
0.6	16.37	18.34

Table 4: The effect of the quality of filtered synthethic data with different filtering thresholds in terms of BLEU sore.

On the other hand, Table 5 shows that if we finetune our proposed model *Iterative multi-NMT + our data filtering* on a small amount of real parallel data, the model performs a significant improvement by +9.26 and +4.76 over the baselines (models are only trained on real parallel data). The real datasets for $km \rightarrow vi$ and $id \rightarrow vi$ are from the ALT corpus with size of 18,088 sentence pairs. This results prove that our proposed model work well on both zero-resource and low-resource language pairs.

model	km → vi	id → vi
direct	13.39	16.81
Iterative multi-NMT + our data filtering + incremental training	**22.65 (+9.26)**	**21.57 (+4.76)**

Table 5: Translation performance (BLEU) when finetuning our proposed model on a small amount of real parallel data.

6 Conclusion

In this paper, we have shown a *training-generating-filtering-training* cycle to build a model for translating zero-resource language pairs. In addition, we proposed a simple filtering method based on word alignments and the longest parallel phrase to filter out poor quality sentence pairs from the synthetic data. Experiment results show that our proposed methods outperformed some strong baselines and achieve a promising result under zero-resource conditions for the Khmer→Vietnamese and Indonesian→Vietnamese directions. Specially, our proposed model can easily improve their quality with a small amount of real parallel data.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation, 03.

Dzmitry Bahdanau, Kyunghyun Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409, 09.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *CoRR*, abs/1710.04087.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. pages 99–107, 01.

Yaroslav Gani, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-adversarial training of neural networks. 05.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. pages 1258–1268, 01.

Thanh Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. 11.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. 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 the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Surafel Melaku Lakew, Quintino Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving zero-shot translation of low-resource languages. 12.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective Approaches to Attention-based Neural Machine Translation. *CoRR*, abs/1508.04025.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015b. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, 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 the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681, 12.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. 08.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

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. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Hien Vu Huy, Phuong-Thai Nguyen, Tung-Lam Nguyen, and M.L Nguyen. 2013. Bootstrapping Phrase-based Statistical Machine Translation via WSD Integration. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1042–1046, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. 09.

Guanghao Xu, Youngjoong Ko, and Jungyun Seo. 2019. Improving Neural Machine Translation by Filtering Synthetic Parallel Data. *Entropy*, 21(12):1213, Dec.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas, November. Association for Computational Linguistics.

Enhancing Quality of Corpus Annotation: Construction of the Multi-Layer Corpus Annotation and Simplified Validation of the Corpus Annotation

Youngbin Noh[1], Kuntae Kim[1], Minho Lee[1], Cheolhun Heo[1], Yongbin Jeong[1], Yoosung Jeong[1], Younggyun Hahm[1], Taehwan Oh[2],Hyonsu Choe[2], Seokwon Park[2], Jin-Dong Kim[3] and Key-Sun Choi[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Yonsei University, South Korea, [3]Database Center for Life Science, Japan
{vincenoh, kuntaek, pathmaker, fairy_of_9, kuonom, wjd1004109, hahmyg, kschoi}@kaist.ac.kr
{ghksl0604, choehyonsu, sseokhon}@yonsei.ac.kr
jdkim@dbcls.rois.ac.jp

Abstract

In this paper, we construct simultaneously multi-layer corpus annotation of 7 linguistic layers as a gold set. And we design the validation procedure using the gold set and report the results of a validation procedure for other large-scale corpus annotation of 7 linguistic layers. Furthermore, we present a simplified validation method without a gold set using annotation models learned by the gold set. Based on the validation results, the tendency of annotation across the entire corpus is observed macroscopically, and the corpus annotation validation results are analyzed microscopically to verify the validation methodology to address the case study.

1 Introduction

As a resource for natural language processing, the corpus is annotated with additional information for various purposes. To annotate such various information to the raw corpus, corpus annotation project must be designed elaborately considering the requirements of the annotation procedure, annotation units, annotation tools, human annotators, and so on. A reliable annotation design makes the corpus annotation better quality(Finlayson and Erjavec, 2017). Also, the design suitability of the corpus annotation project needs to be proved empirically, so the design of the corpus annotation project must be revised and supplemented iteratively(Pustejovsky and Stubbs, 2012).

In this paper, we construct 7 linguistic layers[1](Ide, 2017) of multi-layered corpus annotation as gold set (210K *Eojeol*[2]s) to validate large-scale corpus annotation by the 7 layers as evaluation set (3M *Eojeol*s). The gold set is annotated on the subset of the raw corpus of the evaluation set. The annotator groups of gold set by each layer, who annotated gold set assisted by auto-labeling are groups of experts separated from the annotator groups of evaluation set.

We have designed and applied a corpus annotation method that uses the simple inter-annotator agreement to construct the gold sets at the same time under limited time and human resources. To do this, we assigned one annotation unit to two annotators. According to annotation results from two annotators, we conducted the cross-checking process to determine the final annotation result.

In this paper, validation of evaluation sets by layers is performed by comparing two corpus annotations (gold set - evaluation set)[3] using the gold set as a criterion. Comparative validation of the two corpus annotations using a gold set can only be performed on a limited part of the evaluation set sharing the same part of raw corpus to be annotated.

[1]We construct and evaluate the 7 linguistic layer corpus annotation: morphological analysis, lexical sense analysis, named entity analysis, subject anaphora resolution, co-reference resolution, dependency analysis, and semantic roles analysis. The evaluation sets are also constructed by the same layers.

[2]In Korean, the word segment divided by white space is called "*Eojeol*", this is composed of a noun or verb stem combined with a postposition ("*Josa*") or ending ("*Eomi*") that function as inflectional and derivational particles. (Noh et al., 2018)

[3]In this project, we are constructed 7 linguistic layers of corpus annotations as gold set to validate 7 linguistic layers of corpus annotation (evaluation set) constructed by other project groups. The evaluation sets after validation can be downloaded at https://corpus.korean.go.kr/.

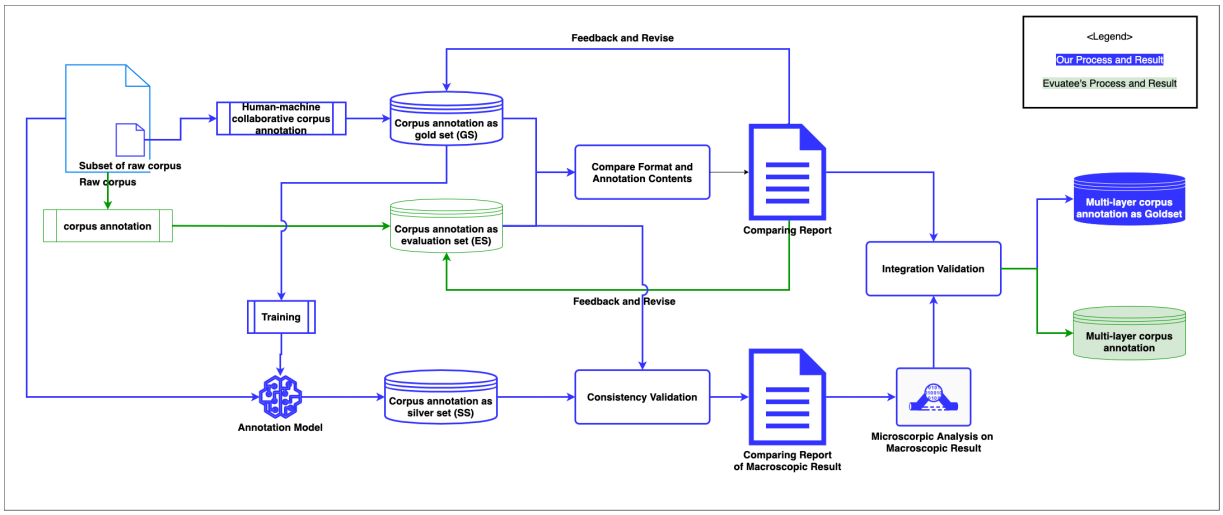


Figure 1: The flow of the corpus annotation and validation process in this paper. Blue-coloureds indicate our process and result, and green-coloureds indicate evaluatee group's process and result.

That's why we present an additional method to validate in the range of evaluation set without a gold set. Thus, we constructed another corpus annotation (silver set) to observe the consistency of annotation across the entire evaluation set by annotating an annotation model using the gold set of each layer as training data. The silver set is compared with the evaluation set to observe the tendency of annotation across the entire evaluation set, and a data-driven statistical analysis is performed to evaluate the appropriateness of the validation results.

In following section 2, we introduce the related works to the design and annotation of corpus annotation projects and corpus validation. In section 3, we introduce the design and annotation process for the construction of the gold set performed in this study. In section 4, we address the validation process for the format and annotation contents of the corpus annotation with an evaluation set using a gold set and validate the annotation consistency of the evaluation set without a gold set. In sections 5 and 6, we report the validation results in section 4 and issues related to these results.

2 Related Works

The corpus annotation project should be modeled according to the goal of the project to reflect appropriate specifications. This is because the corpus annotation, which is a result of corpus annotation, is learning data for the machine learning algorithm to learn specific phenomena that are not only linguistic but also non-linguistic. Therefore, the corpus annotation project design needs to appropriately reflect the characteristics of the phenomena to be learned in the machine learning algorithm. A broadly used framework is the MATTER cycle(Pustejovsky and Stubbs, 2012). The MATTER cycle is a cyclical process in which a corpus annotation project design ensures that the corpus annotation produces machine learning results that are appropriate for the goal of the corpus annotation project.

The OntoNotes(Hovy et al., 2006; Weischedel et al., 2011; Weischedel et al., 2013) is a multi-layer corpus annotation constructed over six years of texts in various genres in three languages (English, Chinese, and Arabic). It is a multilingual, multi-layer corpus annotation that annotated the structural information of the text as well as the semantic information to understand the meaning of the context based on a syntactic structure derived from Penn Treebank corpus, and the predicate-argument structure derived from Penn PropBank. It includes annotations such as word sense disambiguation for verbs and nouns, entity names annotation, the ontology of each word, and coreference relations. OntoNotes had tried to secure at least 90% of inter-annotator agreement in each corpus annotation, improving the quality of corpus annotation.

		Gold Set (Ours)	Evaluation Set	**The OntoNotes 5.0** (more details in (Weischedel et al., 2013))
Language		Korean		English, Chinese, Arabic
Domain of Raw Corpus	**written**	Newspaper		newswire, broadcast news, broadcast conversation and, web data in English and Chinese, a pivot corpus in English (Old and New Testaments and Arabic (Newswire)
	Spoken	Transcripts of recording files (public conversation, public monologue, private conversation)		
Linguistic layer	**written**	Morphological analysis, lexical sense analysis, named entity analysis, subject zero anaphora resolution, coreference resolution, dependency analysis, semantic role analysis		Penn Treebank, Penn PropBank, word sence disambiguation for nouns and verbs, word senses connected to an ontology, and coreference
	Spoken	Morphological analysis, lexical sense analysis, named entity analysis, subject zero anaphora resolution, coreference resolution		
Quantity	**written**	140K *Eojeol*s	2M *Eojeol*s	2.9M words
	spoken	70K *Eojeol*s	1M *Eojeol*s	
Annotator groups		7 annotator groups by the layers different from evaluatee groups	7 annotator groups by the layers	
Constructing time		about 6 months per entire gold set (7 layers)	about 4-6 month per an evaluation set	about 6 years released to The OntoNotes 5.0 from The OntoNotes1.0

Table 1: Comparison of corpus specification of gold set, evaluation set and The OntoNotes 5.0. *Eojeol* is a unit of word segmentation of Korean.

In NLP area, various evaluation and annotation methodologies have been used to enhance and manage corpus quality as a natural language processing resource. As a corpus annotation quality control methods, inter-annotator agreement(Pustejovsky and Stubbs, 2012) has been generally used to control the result of corpus annotation. Checking inter-annotator agreement among annotators is widely used not only for evaluating the results of annotations from an assigned group of annotators, but also for evaluating the quality of data collected from an unspecified number of annotator, such as crowdsourcing methodology(Nowak and Rüger, 2010; Dumitrache, 2015; Dumitrache et al., 2018; Poesio et al., 2019).

3 Construction of Corpus Annotation

In this section, we address the overall procedures of constructing corpus annotation as a gold set. The gold set and the evaluation set share a raw corpus, and the gold set is a corpus annotation of 7 layers constructed by sampling 7% of the raw corpus. Therefore, in this paper, the corpus annotation of 7 layers is simultaneously constructed to build multi-layer corpus annotation.

3.1 Annotation Specification

To establish the corpus annotation guidelines by layers, the existing corpus annotation guidelines are revised and used according to the project purpose[4]. To make sure that the revised guidelines are not ambiguous or lacking in reflecting actual linguistic phenomena, three different annotator groups[5] constructed sample corpus annotation layer by layer on the same range of raw corpus. These sample corpus annotation also assessed the completeness of the corpus annotation guidelines, but were used as an annotation example in the annotation process. Through this process, it is possible to supplement

[4]The annotation guidelines referenced in this project refer to the annotation guideline for each layer from the 21st century Sejong project (morphological analysis, lexical sense analysis), and the guidelines made by the Electronics and Telecommunications Research Institute (ETRI; named entities analysis, subject zero anaphora resolution, co-reference resolution, and semantic role analysis) and Telecommunications Technology Association (TTA; dependency analysis). These guidelines do not refer to literature information in this paper, because it also includes non-public materials. For inquiries about these guidelines, please contact NIKL, ETRI, TTA.

[5]The annotation results of the three groups - ours, evaluatee groups, and expert group of National Institute of Korean Language (NIKL) annotated on the same range of raw corpus on the seven layers. The disagreement among the corpus annotation results of the three groups was decided by the NIKL expert group and the annotation guidelines were reflected in the results of the decision by NIKL.

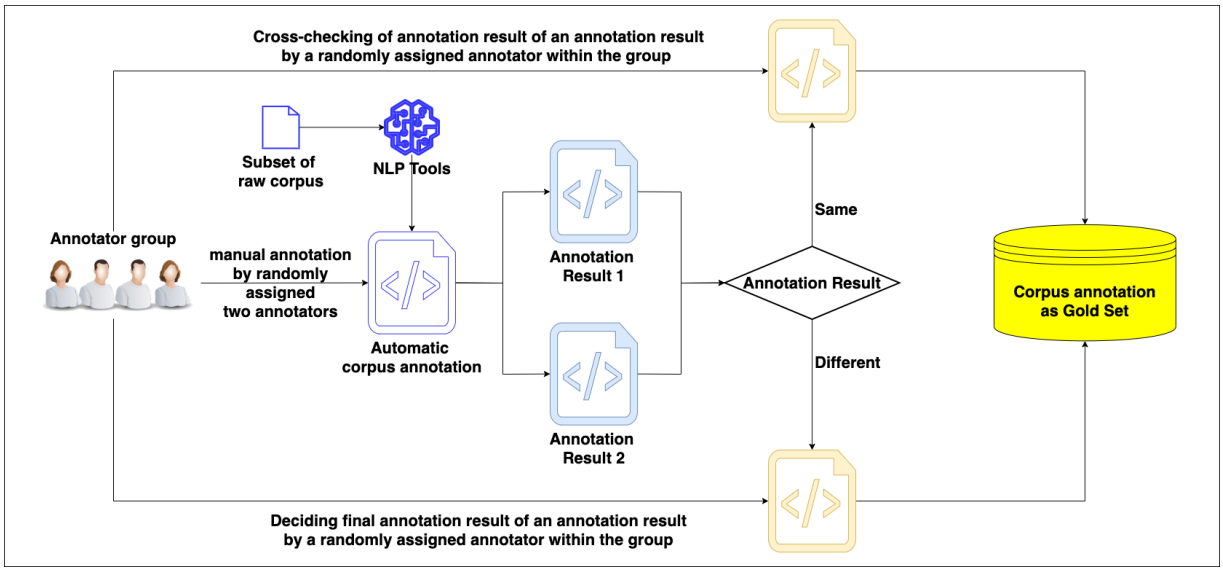


Figure 2: The flow of construction procedure. This figure is a detailed representation of the part of the Human-machine collaborative corpus annotation in Figure 1.

the process of evaluating the repetitive annotation scheme through this process.

In this project, the seven annotator groups[6] were recruited by layers, consisting of experts in Korean or linguistics, who can fully understand annotation guideline and apply this to corpus annotation. They studied and analyzed the existing guidelines and constructed sample corpus annotation. And they annotated the gold sets using the revised guidelines reflecting the results of the annotation of the sample corpus annotation. To ensure validation results of corpus annotation between two annotator groups, they were performed corpus annotation independently not only within their group but also from the annotator group of the evaluatee group. Also, they are completely separated from the annotation group of the evaluatee group and are independently annotated to provide conditions for evaluating the results of the annotation by inter-annotator agreement.

3.2 Annotation Environments

As mentioned above, the annotators of our project annotated independently of each other. For the annotators to be separated, a workbench with a personal workspace is required. We also designed a web-based annotation environment to reduce the time and location constraints of annotation work.

This project used a web-based workbench. This annotation environment was designed for this project and was developed to reflects the annotation schemes of the seven layers. This workbench could only be accessed by annotators and administrator designated for each layer. Multiple annotators can annotate one annotation unit, and it is possible to grasp the annotator's annotation status on an online browser.

Annotators can use only the set of tag sets defined in the annotation schemes, and the annotation contents can also annotate only those in the annotation schemes. If the annotation guidelines have updated and annotators need to add new annotation content, they can ask the administrator to remove the restrictions for annotation contents. Also, the annotators annotated referring the results of the automatic preprocessing of the raw corpus by layers.

This workbench is equipped with a function of randomly assigning a defined annotation unit to an annotator. This allows an annotator not to annotate the entire document at once, but to annotate parts so that one document is annotated by multiple annotators.

[6]The annotators by layers is a group of experts with a master's degree or Ph.D. of the Korean language or linguistics. They were qualified as annotators by NIKL before corpus annotation.

3.3 Corpus Annotation Procedure

In this paper, corpus annotation as a gold set was constructed using a human-machine collaborative annotation method (Figure 2). First, an automatic corpus annotation is automatically annotated using NLP Tools[7] to the subset of the raw corpus. This includes the annotation results of the Korean morphemes, dependency parsing, named entity recognition, and semantic role labeling (recognizing predicate and argument), and so on.

Next, the annotator in the annotator group manually re-annotates the annotation unit by referring to the annotation result of the NLP tool. One annotation unit is annotated by randomly assigned two annotators in the group. If the annotation results of an annotation unit by two annotators are same, these annotation units are cross-checked by a randomly assigned annotator within the group. Else if the annotation results of an annotation unit by randomly assigned two annotators are different, these annotation units are decided final annotation result by a randomly assigned annotator within the group. Through these processes, an annotation unit is checked by annotators at least two times. After two parallel processes, those results of two-stage are made corpus annotation gold set.

4 Validation of Corpus Annotation

After constructing the gold sets, using these corpora, we validate the evaluation sets. The validation of the evaluation sets validates the format of corpus annotation, the annotation contents, and the annotation consistency. After that, the integration validation is performed to be used the evaluation sets as a multi-layer corpus annotation (refer in Figure 1).

4.1 Format Validation

The format validation is a process of confirming whether the corpus has been constructed by the defined annotation format, and also is a stage of confirming whether the corpus can be used as electronic data. At this stage, corpus annotation is validated about the standard format and data structure. In addition to the corpus format, in this stage, it is checked whether or not a label defined for each layer is used, and whether other content is included in addition to the specified annotation content. When a format error is detected in this stage, the evaluation set does not proceed to the annotation contents validation stage, and correction and supplementation are required.

4.2 Annotation Contents Validation

The annotation contents validation performs data-oriented validation that compares the gold set and evaluation set. At this validation, we compare the annotation contents of the two corpora and report the different annotation content to the evaluatee group. The annotation content validation items are selected based on the annotation guidelines for each layer, and the annotation contents defined in the layer are selected as validation items and shared in the evaluatee groups. Based on this report, the evaluatee groups can modify and supplement their corpus annotation. This process is a method of evaluating based on the inter-annotator agreement between annotation groups, and it is judged that the correct annotation is performed when the annotation of the two groups matched.

4.3 Consistency Validation for Evaluation Set using Silver Set

The annotation consistency of the evaluation set is evaluated indirectly by confirming that the tendency of the annotation of the gold set is similar to the evaluation set. To do this, we create an automatic annotation model for each layer using a gold set as training data and construct an automatic corpus annotation (silver set) that annotates automatic annotation on a raw corpus in range of without a gold set. By comparing the silver set and evaluation set, the consistency of the annotation content is analyzed to evaluate the annotation consistency.

To validate the annotation consistency of the evaluation set, we were divided into 10 sections to analyze the tendency of the agreement between the silver set and evaluation set. The average of the agreement of corpus annotations between two corpora in 10 sections was averaged, and when the observed agreement rate of each section deviated from the 99% confidence interval ($\alpha = 0.01$) compared to the

[7]We were supported by the Exobrain Korean Language Analysis Toolkit v3.0 developed by the Electronics and Telecommunications Research Institute (ETRI) to automatically annotate the raw corpus.

mean value, the corresponding section was evaluated to have relatively lower annotation consistency than other sections.

4.4 Integration Validation

As a final stage in constructing a multi-layer corpus annotation of seven layers, it is necessary to check whether the raw corpus of evaluation sets are preserved and whether the annotation schemes have been observed. To make a multi-layer corpus annotation by combining the seven sets, we compared the statistics of the number of documents, paragraphs, sentences, and *Eojeol*s in each corpus, and confirm that the defined ID assignment rules are followed.

5 Results

5.1 Annotation Contents Validation

Table 2 shows the results of annotation agreement between the gold set and the evaluation set. Annotation contents validation of morphological corpus annotation was validated to match the *Eojeol* segment and morphological label annotation. In Korean *Eojeol*, morphemes such as stems (*Eogan*), postposition (*josa*), ending (*Eomi*) are combined to form a single *Eojeol*. Thus, to analyze morphemes, it is necessary to check whether the *Eojeol* segmentation is same and whether the same label is annotated to the segmented morpheme. Written corpus annotation was relatively consistent with both segmentation and label annotation in spoken corpus annotation. Segmenting concordance was lower than label annotation concordance, indicating that there was a difference in the morpheme semantic analysis of *Eojeol* between annotators.

When comparing the annotation agreement between written corpus annotation and spoken corpus annotation, the tendency of the written corpus annotation shows a higher annotation agreement than spoken corpus annotation. It is because a spoken raw corpus transcribed public monologues (news), public conversation (broadcast conversation, interview, lecture, and so on), and private conversation recording. In the case of public monologue, it was refined to a level similar to that of the written raw corpus with well-refined text. In the case of private conversations or broadcast interviews, however, many features of spoken language (i.e., speech break, blur, reduction, slang, and so on) appeared, making it difficult for annotators to analyze text.

Layer	Validation contents	Measures	Written	Spoken
MP	*Eojeol* segmentation	Accuracy	98.6	93.84
	Morpheme label	F1	99.22	97.84
LS	Lexical sense ID	F1	92.47	92.49
NE	Named entity annotation	F1	86.02	94.48
ZA	Predicate Identification (PI)	F1	88.93	88.48
	Subject anaphora resolution	Accuracy	79.20	65.71
CR	MUC	F1	68.20	59.44
DP	Dependency head and label	LAS	87.45	n/a
SR	Predicate Identification (PI)	F1	87.82	n/a
	Argument Identification (AI)	F1	73.86	n/a

Table 2: Results of annotation contents validation.

5.2 Consistency Validation

Table 3 shows the results of annotation consistency validation. Annotation consistency validation was performed separately for written and spoken corpus annotation. To validate the consistency of the annotation, some of the measures used to validate the annotation contents for each layer were used, and the consistency was evaluated through the consistency of the indicator. The 99% confidence interval compared to the average value of the annotated agreement of the silver set and evaluation set for each section was shorter than that of the majority of written corpus annotation (Avg. of CI length: Written = 3.603, Spoken = 3.193). In the case of written corpus annotation, sections 1 and 2 of ZA corpus annotation showed a markedly low agreement, affecting the average CI length of ZA increase.

The corpus showing the shortest 99% confidence interval is the named entity annotated written corpus with the smallest difference in the annotation content with the silver set of 10 sections (99% CI length = .422 ($21.98 \leq CI \leq 22.430$), confidence($\alpha = 0.01$)= 0.202, σ=0.248). Also, one section out of the 99% confidence interval was analyzed, and it was evaluated that the consistency of the annotation of the named entity written corpus annotation was relatively higher than that of the other corpora.

Compared to the annotation content validation result (Table 2), when the gold set and the evaluation set match 80% or more (MP (Written, Spoken), LS (Written, Spoken), NE (Written, Spoken), ZA (Written, Spoken), DP, SR), the length of the 99% confidence interval, except for the ZA written and MP spoken corpus annotation, is all 1.5 or less. There-

Layer		Measures	1	2	3	4	5	6	7	8	9	10	Avg.	CI Length
MP	**Written**	**Accuracy**	**85.87**	**85.87**	86.38	86.09	86.51	**86.60**	86.44	86.33	86.52	86.49	86.298	0.533
MP	**Spoken**	**Accuracy**	**75.35**	**73.75**	76.98	79.30	78.28	79.24	77.40	78.99	**80.16**	79.32	77.877	3.681
LS	**Written**	**Accuracy**	87.64	**86.81**	87.4	87.45	87.81	**87.99**	87.47	**87.95**	87.41	**88.14**	87.607	0.703
LS	**Spoken**	**Accuracy**	**77.72**	79.31	79.51	79.86	78.89	79.2	79.66	**78.37**	79.09	**80.75**	79.236	1.498
NE	**Written**	**Accuracy**	22.04	22.10	21.99	22.58	22.1	22.18	22.26	**21.97**	22.37	22.69	22.228	0.449
NE	**Spoken**	**Accuracy**	**27.45**	20.03	20.10	20.54	21.62	20.55	22.54	**19.15**	22.06	21.20	21.524	4.206
ZA	**Written**	**F1**	**13.98**	**8.55**	20.84	**38.75**	29.3	33.35	31.41	33.31	22.82	31.34	26.365	17.353
ZA	**Spoken**	**F1**	23.70	**22.78**	**22.72**	**24.01**	23.71	**24.08**	22.80	23.07	22.81	**24.02**	23.370	1.058
CR	**Written**	**F1 (MUC)**	**51.32**	50.08	50.34	**49.66**	51.13	**51.23**	**51.36**	50.91	**49.77**	51.09	50.689	1.201
CR	**Spoken**	**F1 (MUC)**	36.39	33.93	33.71	**38.42**	**42.07**	33.57	**38.25**	34.20	33.89	**32.34**	35.677	5.524
DP	**Written**	**UAS**	68.37	68.77	68.11	68.68	68.43	69.00	68.92	69.03	**69.73**	**67.18**	68.622	1.224
SR	**Spoken**	**F1 (AI)**	**61.91**	61.30	61.14	61.34	61.06	61.10	**60.78**	61.15	61.25	61.18	61.221	0.522

Table 3: Results of consistency validation. The bold indicates the agreement rate between the silver set and evaluation set outside the 99% confidence interval. The following models were used for the automatic annotation model for annotation consistency validation: (Ma and Hovy, 2016) - MP (Written), (Joshi et al., 2019) - ZA (Written, Spoken), CR (Written, Spoken), (Straka et al., 2016) - DP (Written), (Lee et al., 2015; Bae et al., 2017) - SR (Written). The annotation models of layers that have no reference was developed and used as a statistical-based learning model.

fore, it was evaluated that it showed high annotation consistency.

The annotation consistency validation by creating a learning model using a gold set started from the hypothesis that the balanced composition of the raw corpus represents a language phenomenon. In the case of the written raw corpus, it is constructed only newspaper articles, so there is relatively little bias in language phenomena according to genres or domains of text. Therefore, it can be said that the written corpus represents more representative of the language phenomenon than the spoken corpus composed of public dialogue, public monologue, and private dialogue.

The quality of the silver set automatically annotated to a well-balanced corpus does not significantly affect the result of the annotation consistency validation. The goal of annotation consistency validation is to verify that the evaluation set shows the annotation characteristics of the annotation model learned by the gold set. As long as it is annotated silver set as a model that properly trains the gold set, it is evaluated that it does not matter if the agreement rate between the silver set and the valuation set is low. However, when comparing the measured value of the agreement divided into 10 sections with the average value, by setting the confidence interval of the average value to 99% (confidence $\alpha = 0.01$), the evaluation standard of annotation consistency for each section was generously set. Also, if the section is further subdivided into 10 or more, more accurate annotation consistency validation will be possible.

As an example, when the results of semantic role corpus annotation annotation consistency validation were analyzed in detail, each section showed a maximum difference of 0.69 from the average (in Table 3, $60.933 \leq CI \leq 61.455$, $60.78 \leq AgreementRate \leq 61.91$). Although it was recorded that it was out of the 99% confidence interval in two sections, the length of the CI was 0.522, which was shorter after the written named entity corpus annotation, and could be evaluated as showing stable annotation consistency. Also, the agreement between the silver set and the evaluation set constructed with the automatic annotation model trained with the gold set is 61.221, but when comparing the sample annotation corpus and the silver set, the consistency was 66.43. It could be judge indirectly as having no significant effect on annotation consistency.

5.3 Case study

A typical inconsistency was mis-annotation on exceptions (Table 4). In semantic role annotation, it consists of the cases on the adverbs of Korean that share a root with a specific verb, auxiliary verbs that composes a predicate in combination with the main verb, and verbs that are a part of periphrastic constructions or tagmeme equivalents. Some Korean verbs function as a marker of aspect, modal, and negation in predication or used as an element to form

	imi already	*jinan* has passed	*ile* thing	*daehae* about	*geuleohge* in that way	*malhaneun* saying	*geoseun* is	*olhji* right	*anhda.* not
Gold				ARG1	ARGM-MNR	*malha*.01	ARG1	*olh*.01	ARGM-NEG
E1		*jinan*.01	ARG1						
E2			ARG2	*daeha*.01					
E3					*geuleoh*.01				
E4								ARG1	*anh*.01

Table 4: SR example of mis-annotation on a sentence which means 'It is not right to say so about what has already passed.'

multi-word periphrastic construction. In particular, the verbs included in the periphrastic constructions are characterized by: 1) do not affect the content of the proposition, only play grammatical functions, 2) the use is morphologically fixed, and 3) it cannot form a sentential predicate or is not related to the argument structure. E1 is a disagreement caused by the annotator mistaken complementation for relativization, E2 is a disagreement on multi-word periphrastic construction *'-e daeha-'* (about/on that), E3 is a disagreement on adverbs that share a root with a specific verb, E4 is a disagreement on an auxiliary verb for negation.

6 Discussions and Conclusions

In this paper, we propose and implement a methodology for constructing language resources for NLP tasks quickly and efficiently for goals of annotation project, but also try to achieve an appropriate level of corpus annotation quality assurance.

We designed the constructing process of gold set to consider agreement within annotators when the results from two annotators for one annotation unit match or not, the annotation contents in the annotator group were once again annotated so that the annotation contents were cross-checked and confirmed. This method is a simple and reliable method to check the difference in the subjectivity of the annotator in a short time. In particular, because Korean has the properties of agglutinative language, it has the possibility that a single annotation unit can be interpreted in multiple meanings, so it is necessary to carefully consider the context information surrounding the annotation units. Even though the annotators have annotated deliberately in compliance with the annotation guideline, there are many possibilities for annotation due to semantic diversity, ambiguity, or subjectivity of annotators.

To construct a gold set in a short time and use it to validate the evaluation set, the quality and authority of the gold set are always important. That is why we designed the process of determining the appropriate annotation minimizing annotation bias while comparing the annotation results within or among annotator groups. Furthermore, we use the gold set as training data of the annotation model to annotate the silver set. If this method is more elaborately, it could be an alternative to evaluating the quality of a corpus annotation when all gold set corresponding to the evaluation set could not be made.

It is difficult for everyone to interpret identically the same linguistic phenomenon due to environmental or individual aspect. Also, it is hard to say that the gold set is always correct. Therefore, this paper tried to aim to present a method to reduce individual and group bias when constructing corpus annotation. In future research, we try to further generalize the methodology for constructing and validating corpus annotation.

Acknowledgments

This work is written based on the results of the 'Corpus Integration Validation'(NIKL 2019-01-61) project of the National Institute of Korean Language.

This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01780, The technology development for event recognition/relational reason-

ing and learning knowledge based system for video understanding)

This work was supported by Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) [2016-0-00562(R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly]

References

Jangseong Bae, Changki Lee, and Hyunki Kim. 2017. Korean semantic role labeling with highway bilstm-crfs. In *Annual Conference on Human and Language Technology*, pages 159–162. Human and Language Technology.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing semantic label propagation in relation classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium, November. Association for Computational Linguistics.

Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *European Semantic Web Conference*, pages 701–710. Springer.

Mark A Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Nancy Ide. 2017. Introduction: The handbook of linguistic annotation. In *Handbook of Linguistic Annotation*, pages 1–18. Springer.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Changki Lee, Soojong Lim, and Hyunki Kim. 2015. Korean semantic role labeling using structured svm. *Journal of KIISE*, 42(2):220–226.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Youngbin Noh, Jiyoon Han, Tae Hwan Oh, and Hansaem Kim. 2018. Enhancing universal dependencies for korean. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 108–116.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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 1778–1789.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees

Masanori Oya
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This study proposes the idea that the difference between the syntactic structures of a sentence and its translation pair in another language can be numerically represented by their Euclidean distance, calculated on the basis of the degree centralities and closeness centralities of the syntactic dependency trees of the sentences. The mean distances thus calculated for a set of translation pairs of two languages can be used as a measure of the similarity/difference between these two languages. A corpus analysis using a multi-lingual parallel corpus reveals that mean Euclidean distances thus calculated seem to reflect the typological tendencies of the differences between languages within and between language families.

1 Introduction

Sentence similarity measuring has recently attracted the attention of many researchers because it is required for various natural language applications, such as those related to question answering (De Boni and Manandhar, 2003), plagiarism detection (Alzahrani et al., 2012), and semantic searching (Farouk et al., 2018). Sentence similarity measuring has been conducted through a variety of techniques, yet the majority of them emphasize lexico-semantic similarity between sentences. Syntactic similarity measures in this context are typically employed only to augment the accuracy of semantic similarity measurement (e.g., Batanovic and Bojic, 2015; Lee et al., 2014; Ma and Suel, 2016), possibly based on the observation that the same meaning can be expressed by a variety of sentences with different syntactic structures and, inversely, that sentences with the same syntactic structure but different words can have completely different meanings. This one-to-many correspondence between sentential meaning and syntactic structures situates lexical or semantic similarity between sentences as equivalent to sentence similarity and syntactic similarity as playing a secondary, supplemental role in sentence similarity.

In spite of the success of sentence similarity measuring, its trend with an emphasis on lexico-semantic similarity should not distract us from investigating the purely syntactic similarities or differences between sentences, which remains worthy of extensive linguistics research with high quality data. The most appropriate data for this purpose is multilingual parallel corpora that consist of a large number of sentences and their translations in several languages. Since these translation pairs are semantically equivalent and the lexico-semantic differences between them are somewhat controlled, we can focus on their syntactic similarities/differences.

In analyzing syntactic similarities/differences between translation pairs in a multilingual parallel corpus, it is important to take a quantitative approach for the following reasons. First, quantifying the syntactic similarity of a given translation pair of two languages included in a multilingual parallel corpus allows it to represent the pure syntactic similarity of the translation pair, and thus to be applied to the types of natural language processing applications mentioned above as an auxiliary measure for sentence similarity. Second, and more importantly, many existing similarity analyses were subjectively conducted by individual researchers, such as those focused on

the differences and similarities in the syntactic structures of sentences in languages of different branches or families. In this context, quantitative approaches to syntactic structure can bring greater objectivity to linguistic analyses (Oya, 2014).

Addressing the need for more quantitative work in this area of linguistics, this study proposes (1) that the syntactic-structural property of a sentence can be numerically represented as the graph centralities of the dependency tree of the sentence, (2) that the difference in the syntactic structures between a sentence and its translation can be numerically represented by their Euclidean distance, and (3) that the average of the distances between a set of translation pairs can be used as a measure of similarity or difference between the two languages. To this end, this study makes a number of assumptions. First, we assume that the structural setting of the syntactic dependency structure for a sentence can be represented by two unique centrality values of the dependency relationships among the words in the sentence: degree centrality and closeness centrality. A dependency tree, which is a formalism of syntactic structure, has one unique degree centrality and one unique closeness centrality. Therefore, these two unique values can be used as unique coordinates, based on which it is possible to calculate the Euclidean distance between them. Second, we assume that the syntactic dependency structures of translation pairs are identical when the Euclidean distance between them is zero, and that their similarity is inversely proportional to the Euclidean distance between them (i.e., the more distant they are from each other, the less similar they are to each other). Since translation-pair sentences have the same meaning, the semantic difference between them is controlled as a minimum; thus, we can presume that the Euclidean distance between these syntactic structures represents the purely syntactic difference between these two sentences of a translation pair. Third, we assume that the mean Euclidean distance thus calculated between translation pairs from two languages of the same language branch (or family) is shorter than that between two languages of different branches. This assumption is based on the insight that the translation pairs of two languages of the same language branch may share similar structural settings because they are semantic equivalents. To verify this assumption, the study (1) used sentences from a multi-lingual parallel corpus that includes translation pairs of sentences from Indo-European languages such as Germanic, Romance, and Slavic as well as from non-Indo-European languages such as Chinese, Japanese, and Finnish; (2) calculated the degree centralities and closeness centralities of these sentences; (3) calculated the Euclidean distances between the translation pairs of these languages; and (4) compared these distances across languages of the same linguistic branch and of different linguistic branches.

2 Dependency grammar and typed-dependency trees

A recent trend in syntactic analysis is the emergence of numerous dependency-based frameworks, such as Word Grammar (Hudson, 2010), the Extensible Dependency Grammar (Debusmann and Kuhlmann, 2007), the Stanford Dependency (De Marneffe and Manning, 2012), and Universal Dependency (McDonald et al., 2013; Nivre et al., 2016; Tsarfaty, 2013; De Marneffe et al., 2014; Zeman, 2015). These modern developments in dependency grammar frameworks originate from Tesnière (1959). Tesnière's notion of dependency grammar can be summarized as follows (Oya, 2020): (1) each word in a sentence is dependent on another word, (2) no word in a sentence is independent, and (3) the dependency relationship between words is directed from a head to a tail. For example, Figure 1 outlines the dependency tree for the sentence "David has written this article."

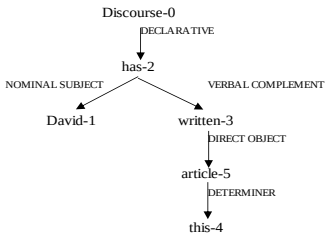


Figure 1. The dependency tree for "David has written this article."

The formalism and dependency types chosen in this study are based on Universal Dependency (UD), which makes explicit the connections among the words in a sentence.

The characteristics of such a network can be quantified in several ways based on methods from

the field of graph theory (Oya, 2014). In other words, the structural properties of networks of words in sentences can be made explicit in dependency grammar and then quantified using graph theory.

In particular, the representation of dependency relationships among words can be interpreted as a directed acyclic graph (DAG) in the UD framework. Therefore, the dependency tree for a sentence is a DAG and represents the formal syntactic property of the sentence.

3 Graph centralities and the Euclidean distances of syntactic dependency trees

The characteristics of a given graph are defined by various measures in graph theory (Freeman, 1979; Wasserman and Faust 1994). One such measure is centrality. The centrality of a node in a graph represents its relative importance within the graph. Two types of graph centrality are employed in this study: degree centrality and closeness centrality.

3.1 Degree centrality

Degree centrality is defined by the degree of a given node, that is, the number of edges a given node has (Freeman, 1979; Wasserman and Faust, 1994). The degree centrality of a node in a graph is formally represented as follows:

$$C'_D(n_i) = \frac{d(n_i)}{g-1} \quad (1)$$

where $C'_D(n_i)$ is the degree centrality of the graph, $d(n_i)$ is the degree of the ith node of the graph n, and g is the number of the nodes in the graph (Wasserman and Faust, 1994).

The degree centrality of the whole graph C_D is the sum of the maximum degree in the graph minus the degree of each of all the other nodes, divided by the largest possible sum of the maximum degree of the graph minus the degree of all the other nodes (Wasserman and Faust, 1994). The degree centrality ranges from 0 to 1:

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]}{max\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]} \quad (2)$$

The degree centrality of a given typed-dependency tree of a sentence indicates the extent to which the words in the sentence are dependent on one particular word (Oya 2014). Thus, the degree centrality of a syntactic dependency tree can be interpreted as its flatness. Degree centrality equals 1 when a node is adjacent (connected by one edge) to all the other nodes in a graph. In terms of the dependency among words in a sentence, the degree centrality of a syntactic dependency tree is 1 when a particular word is the dependency head of all the other words in a sentence.

Degree centrality decreases as the edges connecting nodes in a graph become less concentrated on one particular node; in terms of dependency among words in a sentence, the degree centrality of a syntactic dependency tree decreases as the dependencies among words become less concentrated on one particular word in a sentence.

3.2 Closeness centrality

The distance from one node to another is represented by the number of edges between them. Freeman (1979) and Wasserman and Faust (1994) define closeness centrality as the reciprocal of the sum of the length of a path from one node to another in a graph. The closeness centrality of a graph is calculated as follows (Sabidussi, 1966; Wasserman and Faust, 1994; Beauchamp, 1965):

$$C_c(n_i)=\frac{g-1}{\sum_{j=1}^{g} d(n_i,n_j)} \quad (3)$$

where g means the number of nodes and $d(ni,nj)$ is the shortest path (geodesic distance) between the nodes n_i and n_j. Closeness centrality thus calculated can be viewed as the inverse average distance between node i and all the other nodes in the graph, ranging from 0 to 1 (Wasserman and Faust, 1994).

As in the case of degree centrality, closeness centrality equals 1 when a node is adjacent to all the other nodes in a graph. In terms of the dependency among words in a sentence, the closeness centrality of a syntactic dependency tree is 1 when one particular word is the dependency head of all the other words in a sentence.

Closeness centrality decreases as the nodes are aligned further away from each other in a graph; meanwhile, in terms of dependency among words in a sentence, closeness centrality of a dependency tree represents the extent to which its words are close to each other along its dependencies, and thus numerically indicates the embeddedness of the words in the tree; greater closeness centralities mean less embedded dependency trees (Oya, 2014).

3.3 The Euclidean distance between syntactic dependency structures

The degree centrality and closeness centrality of one syntactic dependency tree can be used as unique coordinates to indicate the structural setting of the tree, because one syntactic dependency tree contains unique degree and closeness centralities. On the basis of these coordinates of the dependency trees of translation-pair sentences, it is possible to calculate the Euclidean distance, that is, the structural similarity, between them.

The Euclidean distance is calculated as follows. Let there be two points in the Cartesian coordinates, represented as $p = (p_1, p_2,..., p_n)$ and $q = (q_1, q_2,..., q_n)$, respectively; then, the Euclidean distance from p to q (or from q to p) is calculated by the following formula:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_1 - p_1)^2} \quad (4)$$

The Euclidean distance between a sentence and its translation pair can be calculated with their degree centralities on the x-axis and their closeness centralities on the y-axis on the assumption that degree centralities and closeness centralities are orthogonal coordinates. For example, the Euclidean distance between a sentence in one language s_1 and its translation pair in another language s_2 is calculated as follows. Suppose sentence s_1 has a degree centrality 0.22 and a closeness centrality 0.33 and the sentence s_2 has a degree centrality 0.14 and a closeness centrality 0.41. These sentences emerge as two vectors, the first of which is s_1 (0.22, 0.33) and the second is s_2 (0.14, 0.41). The Euclidean distance between them is calculated as follows:

$$d(s_1, s_2) = \sqrt{(0.22 - 0.14)^2 + (0.33 - 0.41)^2} \approx 0.113 \quad (5)$$

The Euclidean distance between a sentence and its translation pair in one language allows us to calculate the distance between the sentence and its translation pairs from other languages. For example, the Euclidean distance between sentences s_1 and s_3, which is its translation pair from a third language, can be calculated by the same procedure described above. Suppose that sentence s_1 has a degree centrality 0.22 and a closeness centrality 0.33, and sentence s_3 has a degree centrality 0.20 and a closeness centrality 0.35. Then, these sentences emerge as two vectors, the first of which is s_1 (0.22, 0.33) and the second is s_3 (0.2, 0.35). The Euclidean distance between them is calculated by the following formula below:

$$d(s_1, s_3) = \sqrt{(0.22 - 0.2)^2 + (0.33 - 0.35)^2} \approx 0.028 \quad (6)$$

Thus, the Euclidean distance between s_1 and s_2 (approximately 0.113) is greater than that between s_1 and s_3 (approximately 0.028). In other words, s_1 is closer to s_3 than to s_2 with respect to their structural settings, which are hierarchically represented by dependency among the words, and numerically by their degree centralities and closeness centralities. This means that the syntactic structure of s_1 is more similar to that of s_3 than that of s_2.

3.4 Comparisons of the Euclidean distances calculated from a parallel-corpus

Notice that the idea of Euclidean distance between a sentence from Language A and a sentence from Language B described in the previous section can further be applied collectively by contrasting the Euclidean distances between sentences from Language A and sentences from Language B and further contrasting the Euclidean distances between sentences from Language A and sentences from Language C. If it is found that the frequencies of shorter distances between Languages A and B are significantly larger than those between Language A and Language C, then it can be concluded that Language A is structurally closer to Language B than to Language C.

The overall Euclidean difference between Language A and Language B can be represented as the distribution of the frequencies of the Euclidean distances between the translation pairs of these two languages along with a certain interval. If Language A and Language B are structurally similar, then the frequencies of shorter Euclidean distances between them are expected to be higher than those of longer ones; hence, their distribution will be skewed to the left. On the other hand, if Language C is less similar to Language A than Language B, then the frequencies of shorter Euclidean distances are expected to be smaller than those between Language A and B; here, the peak of the distribution goes to the right. If the difference between these two distributions (i.e., that for Languages A and B and that for Languages

A and C) is statistically significantly large, then the structural difference between Language A and Language B is statistically significantly larger than that between Language A and Language C. (cf. Section 5.2).

The Euclidean distances between translation-pair sentences in a multilingual parallel corpus can indicate the dependency-structure settings of these languages, especially the similarity/difference between their syntactic structures. Since these translation pair sentences have the same meaning, their semantic similarity or difference is appropriately controlled for purely syntactic-structural comparisons or contrasts between these languages.

4 Corpus analysis: the Euclidean distance between dependency trees in terms of centrality measures

4.1 Purpose

As discussed in the Introduction, the purpose of this analysis is to examine the assumption that the distance between syntactic structures, or syntactic dependency trees, is expected to represent a purely syntactic similarity or difference between two languages from either the same language family or from different ones.

4.2 Data: The Parallel Universal Dependency corpus

This study used Parallel Universal Dependency (PUD) treebanks as the data for calculating the Euclidean distance between translation pairs based on the degree centrality and closeness centrality of their syntactic dependency trees.

PUD treebanks were created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to UDs. PUD treebanks contain 19,000 sentences from 19 different languages (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish). The subcorpus of PUD treebanks for each language contains 1,000 sentences, in a fixed order across languages. The sentences are aligned one-to-one, although some sentences are translated into two sentences. Of the 1,000 sentences, 750 are translated from English texts, and the remaining 250 sentences are translated from German, French, Italian, or Spanish, which were translated into English and then translated into other languages (their ID numbers indicate the original language). The translation was done by professional translators and annotated with morphological and syntactic tags by Google. They were then converted by UD community members to match UD Version 2 guidelines. For further details on PUD treebanks, refer to the UD webpage (https://universaldependencies.org/).

In summary, PUD treebanks is a set of subcorpora such that (1) each subcorpus of a language contains 1,000 sentences of that language, (2) these 1,000 sentences have their translation counterparts in 18 other languages (1,000 multiplied by 18 equals 18,000 translation pairs for one language), and (3) PUD treebanks ultimately contain 342,000 translation pairs (18,000 multiplied by 19).

4.3 Method

The word count, degree centrality, and closeness centrality of each sentence in PUD treebanks was calculated by an original Ruby script. Another Ruby script was created to calculate the Euclidean distance of the syntactic dependency trees of all the translation pairs in PUD treebanks.

Then, a spreadsheet application was used to calculate the frequencies of Euclidean distances (cf. Section 3.4) between the translation pairs in all the unique combinations of 19 languages (19 languages multiplied by 18 other languages minus non-unique combinations equal 171 combinations).

Translation pairs from different languages show different distributions of Euclidean distances. For example, 1,000 translation pairs exist between English and Japanese in PUD treebanks, and the 1,000 Euclidean distances between them are distributed from 0 to 0.85, with the most frequent one 0.1 with an interval of 0.01 (65 pairs). There are also 1,000 translation pairs between English and German, and their 1,000 Euclidean distances are distributed from 0 to 0.7, with the most frequent 0.04 (90 pairs).

The distributions of these frequencies in each of these unique combinations were compared by a Wilcoxon's signed-rank test, in order to check whether the difference between these distributions

was wider than that caused by chance. This test was chosen because the frequencies of the Euclidean distance between translation pairs did not seem to be normally distributed.

For reasons of space, we focus here on our comparisons of the Euclidean distance between English and Japanese, and the Euclidean distance between English and other languages (18 comparisons overall). The English and Japanese languages were chosen as the focus of comparison (Language A and B in Section 3.4) because they do not belong to the same language family; English belongs to the Indo-European family, while Japanese does not; thus, comparing them provides a starting point for comparisons of other language pairs, which will be conducted in future studies.

5 Results

5.1 The Euclidean distances between different languages in PUD treebanks

The first row in Table 1 summarizes the descriptive statistics *of* the Euclidean distances among the syntactic dependency trees of all the translation pairs in the PUD treebanks in terms of degree centralities and closeness centralities. All the other rows show the Euclidean distances between the syntactic dependency trees of each language and those of all the other 18 languages in the PUD treebanks in terms of their degree centralities and closeness centralities.

Finnish (a non-Indo-European language) has the longest mean Euclidean distance from all the other languages (0.139) while Portuguese (a Romance language, Indo-European family) has the shortest (0.100). The mean Euclidean distances of 9 out of 11 Indo-European languages in PUD treebanks are less than 0.11, while those of 6 out of 8 non-Indo-European languages in PUD treebanks are more than 0.11.

Finnish also has the largest SD (0.139) while German has the smallest (0.075). The SDs of 9 out of 11 Indo-European languages in the PUD treebanks are less than 0.8, and those of 7 out of 8 non-Indo-European languages are more than 0.8.

	Euclidian distance						
	Mean	Median	Mode	Max.	Min.	SD	N
All	0.112	0.095	0	1.005	0	0.083	342,000
Arabic	0.123	0.109	0	1.002	0	0.082	18,000
Chinese	0.118	0.103	0	1.002	0	0.082	18,000
Czech	0.110	0.093	0	1.002	0	0.083	18,000
English	0.102	0.086	0	0.853	0	0.077	18,000
Finnish	0.139	0.115	0	0.869	0	0.105	18,000
French	0.104	0.088	0	0.835	0	0.079	18,000
German	0.104	0.089	0	0.869	0	0.075	18,000
Hindi	0.115	0.101	0	1.005	0	0.079	18,000
Indonesian	0.103	0.088	0	0.793	0	0.078	18,000
Italian	0.102	0.087	0	0.835	0	0.078	18,000
Japanese	0.121	0.106	0	0.856	0	0.081	18,000
Korean	0.137	0.121	0	1.002	0	0.091	18,000
Polish	0.107	0.092	0	1.002	0	0.079	18,000
Portuguese	0.100	0.084	0	0.856	0	0.078	18,000
Russian	0.102	0.087	0	0.865	0	0.077	18,000
Spanish	0.101	0.085	0	0.775	0	0.078	18,000
Swedish	0.105	0.087	0	1.002	0	0.082	18,000
Thai	0.110	0.093	0	1.005	0	0.086	18,000
Turkish	0.125	0.109	0	1.002	0	0.087	18,000

Table 1. The mean Euclidean distances of the syntactic dependency trees of each language to those of all the other languages.

5.2 Comparing the Euclidean distances between different language pairs

This section reports the comparisons of the frequencies of the Euclidean distances between English and Japanese, and those between English and other languages, all in PUD treebanks. These comparisons are intended to show the structural differences between languages in terms of the different distributions of their Euclidean distances based on their degree centralities and closeness centralities.

Of all the 18 comparisons, only one pair (English-Japanese and English-Swedish; Swedish is the Language C in Section 3.4) showed a significant difference between the distributions of the frequencies of the distances, and three others show slightly significant differences (those between English–Japanese and English–German, English–Polish, and English–Spanish).

Figure 2 describes one of the instances in which the two distributions show a significant difference (English-Japanese and English-Swedish).

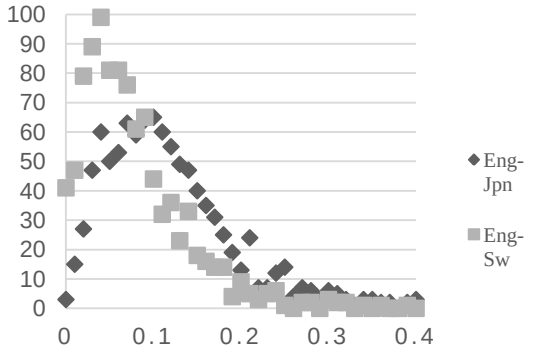


Figure 2. The frequencies of the Euclidean distances between the English sentences and their Swedish translations (Eng-Sw; N = 1,000), and those between the English sentences and their Japanese translations (Eng-Jpn; N = 1,000); x axis: Euclidean distances (interval: 0.01; max: 0.4); y axis: frequencies

A Wilcoxon Signed-Ranks Test indicated that the number of short Euclidean distances between English and Swedish translation pairs was statistically significantly larger than the number of short Euclidean distances between English and Japanese translation pairs ($Z = 2.01$, $p = 0.044$). This means that a significantly larger number of translation pairs of English and Swedish in the PUD treebanks are structurally closer to each other than the English and Japanese pairs.

Figure 3 describes one of the instances in which the two distributions do not show any significant difference; one distribution includes the frequencies of the Euclidean distances between the English sentences and their Japanese translations, and the other those between the English sentences and their Chinese translations.

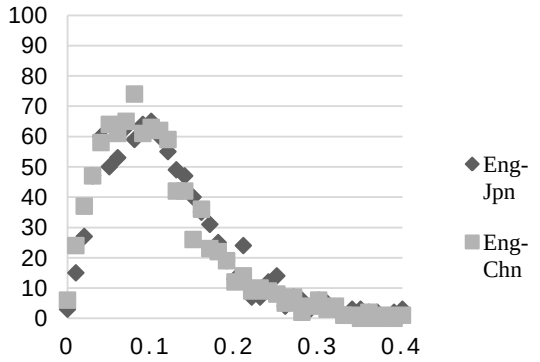


Figure 3. The frequencies of the Euclidean distances between the English sentences and their Chinese translations (Eng-Chn; N = 1,000), and those between the English sentences and their Japanese translations (Eng-Jpn; N = 1,000); x axis: Euclidean distances (interval: 0.01; max: 0.4); y axis: frequencies

A Wilcoxon Signed-Ranks Test indicated that the frequency of short Euclidean distances between English and Japanese translation pairs was not larger than the frequency of short Euclidean distances between English and Chinese translation pairs ($Z = 0.5$, $p = 0.617$). This means that there is no statistically significant difference between the distribution of frequencies of English-Japanese Euclidean distances and those of English-Chinese Euclidean distances.

6 Discussion

The results of the comparisons of the Euclidean distances between the two pairs of languages does not seem to contradict the assumption that the distance between the syntactic dependency trees calculated by the two graph centrality measures represents a purely syntactic difference between two languages. It was found that the mean Euclidean distances of syntactic dependency trees of each language to all the others are divided approximately into two groups. The first group has shorter Euclidean distances and includes Indo-European languages, and the second group has longer Euclidean distances and includes non-Indo-European languages. It was also found that the distribution of frequencies of the Euclidean distances between English (a Germanic language, Indo-European family) and Japanese (a non-Indo-European language) shows no significant difference with that between English and Chinese (a non-Indo-European language), a slightly significant difference with that between English and Spanish (a Romance language, Indo-European family), and a significant difference with that between English and Swedish (a Germanic language, Indo-European family). These results might lead us to assume that the Euclidean distances among syntactic dependency trees calculated above seem to represent the structural similarity of the languages of the same language family/branch and structural difference/diversity of languages of different language families/branches. To verify this assumption, further comparisons must be made between more language pairs in the PUD treebanks or other parallel corpus data with a

larger variety of languages, which is our research goal for future studies.

7 Conclusion

This study has proposed the idea that the difference of syntactic structures of a sentence and its translation pair in another language can be numerically represented by their Euclidean distance, calculated on the basis of the degree centralities and closeness centralities of the syntactic dependency trees of sentences, and that the mean distances thus calculated for a set of translation pairs of two languages can be used as a measure to show similarity/difference between these two languages. The corpus analysis using a multi-lingual parallel corpus revealed that along with some interesting properties of the graph centrality measures of syntactic dependency trees, mean Euclidean distances between the syntactic dependency trees of translation-pair sentences of a variety of languages seem to reveal their typological tendencies. Further comparisons are needed between more language pairs in PUD treebanks or other multi-lingual parallel corpus data.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20K00583.

References

Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, 42(2):133-149. https://doi.org/10.1109/TSMCC.2011.2134847.

Vuk Batanovic and Dragan Bojic. 2015. Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity. Computer Science and Information Systems, 12(1):1-31. https://doi.org/10.2298/CSIS131127082B.

Murray A. Beauchamp. 1965. An improved index of centrality. Behavioral Science, 10:161-163.

Marco De Boni and Suresh Manandhar. 2003. The use of sentence similarity as a semantic relevance metric for question answering. Proceedings of the AAAI Symposium on New Directions in Question Answering, SS-03-07.

Marie-Catherine de Marneffe and Christopher D. Manning. 2012. Stanford Typed Dependency Manual Revised for the Stanford Parser v.2.0.4. Retrieved May 30, 2020 from http://nlp.stanford.edu/software/dependencies_manual.pdf.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. Proceedings of LREC14.

Ralph Debusmann and Marco Kuhlmann. 2007. Dependency Grammar: Classification and exploration. Project report (CHORUS, SFB 378). Retrieved May 30, 2020 from http://www.ps.uni-saarland.de/~rade/papers/sfb.pdf

Mamdouh Farouk, Mitsuru Ishizuka, and Danushka Bollegala. 2018. Graph matching based semantic search engine. Proceedings of 12th International Conference on Metadata and Semantics Research, Cyprus. https://doi.org/10.1007/978-3-030-14401-2_8.

Linton C. Freeman. 1979. Centrality in social networks. Social Networks, 1:215-239.

Richard Hudson. 2010. An Introduction to Word Grammar. Cambridge University Press, Cambridge.

Ming Che Lee, Jia Wei Chang, Tung Cheng Hsieh. 2014. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal. https://doi.org/10.1155/2014/437162.

Weicheng Ma and Torsten Suel. 2016. Structural sentence similarity estimation for short texts. 29th International Florida Artificial Intelligence Research Society Conference. Key Largo, United States. 232-237.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 92-97.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation, 1659-1666.

Masanori Oya. 2014. A Study of Syntactic Typed-Dependency Trees for English and Japanese and Graph-centrality Measures [Doctoral dissertation] Waseda University, Tokyo, Japan.

Masanori Oya. 2020. Structural divergence between root elements in English-Japanese translation pairs. Journal of Global Japanese Studies, 12:107-126.

Gert Sabidussi. 1966. The centrality index of a graph. Psychometrika, 31: 581−603.

Lucien Tesnière. 1959. Éléments de syntaxe structural. Klincksieck, Paris.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 578-584.

Stanley Wasserman and Katherine A. Faust. 1994. *Social network analysis.* Cambridge University Press.

Daniel Zeman. 2015. Slavic Languages in Universal Dependencies. In Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning. Bratislava, Slovakia.

Plausibility and Well-formedness Acceptability Test on Deep Neural Nativeness Classification

Kwonsik Park
Korea University
Department of Linguistics
Oneiric66@korea.ac.kr

Sanghoun Song
Korea University
Department of Linguistics
sanghoun@korea.ac.kr

Abstract

The present work compares performance among several deep learning models, of which task is to clasify nativeness of English sentences. The current study constructs 4 models, each using different deep learning networks: RNN, LSTM, BERT and XLNet. We use 3 test suites to evaluate the four models: (i) 8 test sets composed of 4 native- and non-native-written data, (ii) a supplemented version of well-formedness and plausibility test set consisting of 120 sentences from Park et al. (2020), and (iii) a test set of 196 sentences consisting of 11 types (27 subtypes) for grammaticality judgment test (DeKeyser, 2000). The results show that the more up-to-date models, BERT and XLNet outdo relatively out-of-date models, RNN and LSTM. The latest model among the 4 models is XLNet, but it does not outperform BERT in every aspect. Presuming that the ways deep learning learns language are, to some extent, similar to the strategies of L2 learners, the current work trains the models with data consisting of native and learner English sentences to compare nativeness judgments between deep learning models and humans for investigating if it is the case. This paper concludes that there are few learnability problems shared by the two agents.

1 Introduction

Deep learning is no longer an unfamiliar word to NLP researchers. It has already been used in various tasks in NLP such as the ways to resolve long-distance filler-gap dependencies (Da costa & Chaves, 2020; Chaves, 2020; Wilcox et al, 2019), number agreement (Linzen & Leonard, 2018), reflexive anaphora (Goldberg, 2019), to name a few, and shown lots of striking results. However, there are very few works attempting to train deep learning with learner corpora.

It seems that the ways deep learning learns language are similar to L2 learners' language learning strategies in that both artificial and natural intelligence generalize data, extract meaningful features, solve problems, check errors, modify what has been their knowledge and memorize what they have learned from this procedure. This is in line with the Fundamental Difference Hypothesis (Bley-Vroman, 1988). The hypothesis argues that adult learners learn language with analytical, problem-solving mechanisms. In addition, they are also alike in that their language learning is limited by the poverty-of-the-stimulus, i.e., they depend mainly on input data from the outside and cannot learn a language completely without rich enough data whereas children can.

To check if it is the case, this work makes four language models built up with four different artificial neural network, Recurrent Neural Network (RNN, Mikolov et al., 2010), Long-Short Term Memory RNN (LSTM, Sundermeyer et al., 2012), Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2018), and XLNet (Yang et al., 2019). The task of the models is nativeness classification. Prior works such as Warsdadt et al. (2019) construct their classification models trained with L1 data labeled with a binary grammaticality value. To some extent, their models learn unacceptability of sentences because they are trained with ungrammatical sentences, but we

cannot say they learn *learners*. Instead, we train our models with L1 and L2 data to learn L2 learners. Each of the four models is then evaluated by classifying nativeness of every sentence in 3 test suites to examine the deep learning models in various angles.

Type	Test Data	Number of Sentences	Average Sentence Length
Native (diplomatic, news)	The English Gigaword	3,000	24.83
	The Europarl Corpus	3,000	23.32
Native (novel)	The Speckled Band	572	17.16
	The Little Prince	1,835	9.12
Non-Native (elaborated)	The Three English Speeches of President Kim Dae-Jung	484	17.51
	The Tanaka Corpus	3,000	7.79
Non-Native (not elaborated)	INUMLC (spoken)	697	4.55
	INUMLC (written)	613	14.52

Table 1: The 8 test sets

2 Nativeness Classification

The task assigned to the four models is to identify nativeness of sentences, i.e., to predict whether a sentence is written by a native or non-native speaker. Pawley and Syder (1983) observe that nativelike sentences be 'institutionalized' and 'lexicalized' patterns. They also note that two essentials are required for an expression to be a 'native selection'; not only should it be grammatically well-formed but sounds plausible. An ill-formed sentence refers to the one that has syntactic violations, and an implausible sentence is the one that is syntactically well-formed but sounds awkward to native speakers, e.g., "*I wish to be wedded to you."* (Pawley and Syder, 1983) does not sound natural despite its well-formedness.

3 Model Construction

The present work constructs deep learning models which are trained with native and non-native data to predict nativeness of sentences using four types of deep learning networks: RNN, LSTM, BERT and XLNet. As mentioned above, the task is binary classification of nativeness. The models are designed to output '1' when they predict a sentence as native one, and '0' when predict it as a non-native one. RNN and LSTM models were trained for 10 epochs, and among the 10 epochs, the model with the highest validation accuracy was the highest was used (BERT and XLNet for 4 epochs).

3.1 Data

The entire data are composed of 651,665 sentences, which consist of two parts, training and validation data, and the test data consists of three test suites.

Training and Validation Data

Training data is made up of native- and learner-written sentences, the total size of which is 586,501 sentences (7,852,306 words). The native data used in this paper is extracted from the Corpus of Contemporary American English (COCA). The learner data is excerpted from Yonsei English Learner Corpus (YELC) and Gacheon Learner Corpus (GLC), both of which were made by undergraduate students in Yonsei University and Gacheon University in Korea, respectively. Validation data is one-tenth of the entire data (65,164 sentences), which is used for evaluating classification accuracy of the deep learning models.

Test data

Test data consists of three test suites: (i) 8 test sets, (ii) a supplemented version of well-formedness and plausibility test items consisting of 120 sentences from Park et al. (2020) and (iii) grammaticality judgment test items composed of 196 test items from DeKeyser (2000).

The 8 test sets are made up of 4 native- and 4 non-native-written test sets. As shown in Figure 1, Each of The English Gigaword and The Europarl

Corpus is a highly elaborated version of native randomly excerpted from original data, the former being news and the latter being diplomatic sentences. The Speckled Band and The Little Prince are novel data. English Speeches of President Kim Dae-Jung (hereafter, KDJ) was revised by Korean diplomatic experts, so it has no syntactic violation. The Tanaka Corpus is composed of 3,000 sentences randomly extracted from original data. This was edited by researchers in the process of corpus construction, so this corpus also does not have any syntactic violation. In contrast, both the written version of Incheon National University Multilingual Learners Corpus (hereafter, INUMLC (written)) and the spoken version of Incheon National University Multilingual Learners Corpus (hereafter, INUMLC (spoken)) have lots of syntactic errors in them.

The second test suite English test items from Park et al. (2020), which originally were designed to compare well-formedness and plausibility judgments of native English subjects to those of a deep learning model (the model used in the paper was RNN). The test items of the article consist of controlled and filler items. We use only the controlled items because the filler set was made for the language experiment. The controlled items are composed of 60 well-formedness and 50 plausibility test items, so we add 10 plausibility test sentences to balance between them. Consequently, the test suite consists of 120 sentences.

The last test suite is test items from Dekeyser (2000), which are made up of 196 well-formedness test items categorized into 11 types (27 subtypes). The test items are originally made for examining English grammaticality judgments of immigrants living in the United States (Dekeyser, 2000). The second and third test suites are made up of pairwise sentences and each is labeled in a binary way: ill-formed sentence is labeled '0' and well-formed is '1'. The second suite is not composed of minimal pairs while the third one is. Test items from Dekeyser (2000) are used to investigate whether each model shows similar judgment patterns to those of L2 learners (immigrants).

3.2 Four types of networks

The present study constructs four models, each of which is made from the four different deep learning networks.

Firstly, RNN is a type of artificial neural network in which information from a previous step data. Each of them is composed of 3,000 sentences is updated in the current step but has a limitation of gradient vanishing, which refers to the problem that the longer steps information is carried over, the more information the model loses.

Secondly, LSTM is an elaborated version of RNN. It resolves the problem of gradient vanishing to some extent by updating information from previous steps selectively; important one is memorized and not important one is discarded.

Thirdly, BERT is a relatively up-to-date neural network that 'is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers' (Devlin et al., 2018), which means it can entail bidirectional contextual information on each word by using a special noise token, [MASK], which should be predicted by the model, resulting in representing rich information. BERT is so powerful that it outdoes performance of many previous NLP models, but it has a limitation that 'BERT assumes the predicted tokens are independent of each other given the unmasked tokens, which is oversimplified as high-order, long-range dependency is prevalent in natural language' (Yang et al., 2019).

Lastly, to solve this problem, Yang et al. elaborated BERT to build up the XLNet network, which is capable of doing both autoregressive and autoencoding methods by '[maximizing] the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order'.

The present work constructs four models using those four kinds of deep learning networks and investigates which model is better to detect nativeness of English sentences.

4 Results

The 4 models are evaluated on 3 test suites by classifying nativeness of every sentence in each test suite.

4.1 Test Suite I

Table 2 shows the results of evaluation on the 8 test sets. Numbers in the table are proportions of sentences which are predicted as native sentences in each test set. Left 4 test sets are native data, so a higher score means better performance whereas for 4 test sets in the right side, a lower score means better performance.

	Native Data				Non-native Data			
	The English Gigaword	**The Europarl Corpus**	**The Speckled Band**	**The Little Prince**	**KDJ**	**The Tanaka Corpus**	**INUMLC (spoken)**	**INUMLC (written)**
RNN	97.9	95.7	95.8	91	95.8	83.4	82	44.6
LSTM	95.8	92.4	88.1	76.8	86.7	66.4	76.7	30.8
BERT	95.5	90.1	92.6	80.1	**73.5**	**53.2**	**57.8**	12
XLNet	97.1	94.6	96.1	85.8	76.4	59.8	64.4	**7.6**

Table 2: Nativeness judgment results on the 8 test sets
(proportions of sentences which are predicted as native sentences)

The 4 models predict the nativeness of native test sets well in general except for The Little Prince. The low accuracy on The Little Prince is probably caused by two reasons: the first one is its sentence length is relatively short (see Table 1, the average sentence length is 9.12), which means each sentence in the novel has relatively few clues for the models to make use of to predict nativeness, and the other one is that it has somewhat learner-like sentences which do not contain difficult words or complex clauses. For example, (1a) and (1b) are the ones that every model predicts as non-native sentences. We do not say, of course, sentence length is not a necessary condition of nativeness. But the short sentence length must have influenced the models' judgments.

(1) a. It is unnecessary.
b. This is a ram.

With respect to 4 non-native test sets, there are remarkable differences among test sets. Regarding KDJ and The Tanaka Corpus, the 4 models give relatively high scores to them although they are non-native data. This is probably because, as mentioned before, they are manually edited learner corpora that have no syntactic violation. The reason that KDJ is given higher scores than The Tanaka Corpus could be explained in terms of plausibility; the editing on learner sentences in The Tanaka Corpus was focused only on their syntactic well-formedness, whereas KDJ was sophisticatedly edited on both grammaticality and its content, i.e., its well-formedness and plausibility. Furthermore, The Tanaka Corpus has more learner-like sentences than KDJ because KDJ is such a diplomatic document that it has few learner-like sentences that L2 learners use in everyday conversation. (2a) and (2b) are the ones in The Tanaka Corpus that every model classifies as non-native sentences, partially because of its awkwardness or learner-likeness.

(2) a. My father is proud of my being handsome.
b. My uncle gave me a book.

(2a) is syntactically well-formed but not made in a frequently used pattern. (2b) has no syntactic violation, too, but it is the one that usually appears in learner textbooks. To sum up, KDJ gets higher score as it was elaborated in terms of plausibility as well as well-formedness of its sentences. Nevertheless, the nativeness of KDJ is not fully satisfied; it does not get the score as high as that of native data from models except for RNN, which indicates, however elaborated a text is, it's almost impossible for non-natives to reach the level of native speakers (Park et al. 2019). The scores of 4 models show that BERT and XLNet can detect the nativeness of KDJ and The Tanaka Corpus better than RNN and LSTM.

This better performance of BERT and XLNet is clearly found in the results of INUMLC (written). INUMLC (written) is the most learner-like text which has lots of ill-formed and implausible sentences. These are instantiated in (3a) ~ (3c).

(3) a. *I worried and tired because a lot of people.
b. *I think I could wrote various articles.
c. In my life I like to read books

(3a) has syntactic violations of omitting *be*-verb and using *because* instead of *because of*. (3b) also has a syntactic violation of using a past verb *wrote* after a modal verb *could*. Although (3c) is syntactically well-formed, it is predicted as a non-

	RNN	LSTM	BERT	XLNet
Well-formedness Judgment items	56.6%	56.6%	70.0%	**73.3%**
Plausibility Judgment Items	55.0%	60.0%	**63.3%**	56.6%
Well-formedness + Plausibility	55.8%	58.3%	**66.6%**	65.0%

Table 3: Nativeness judgment results on test items from Test Suite II

native sentence by all the models, partially because *in my life* is not plausible; not only does the expression not go well with the context, but it also usually occurs in the last position of a clause. In this sense, the performance of the four models is reflected the most in the scores of INUMLC (written) because models' judgments on this pure learner data show how correctly they can discriminate non-native sentences from native ones. The scores of BERT and XLNet show a drastic decrease from those of RNN and LSTM, and the latest model, XLNet, gives the lowest score to INUMLC (written), indicating it has the highest performance on the test set.

INUMLC (spoken), on the other hand, gains scores from the models that do not accord with our intuition; although syntactic violations and awkward expressions are more common in spoken data, the scores are quite high, which means all the four models cannot correctly classify the nativeness of the test set. An explanation is that the average sentence length of INUMLC (written), 4.55 words, is too short for the models to detect clues to use for sentences classification. This phenomenon is also found in Park et al. (2019), where an RNN model is used for nativeness classification. (4a) ~ (4c) are the examples of short sentences that all models classify as native sentences.

(4) a. Uh, okay.
b. Oh, yeah.
c. Okay.

Whether sentences above are made by natives or non-natives is probably hard to predict even for humans. We are not arguing, of course, that short sentences are the only factor that causes the models to incorrectly judge the test set; it is just one variable that influences the predictions.

In sum, the results on the 8 test sets indicate that XLNet and BERT have better performance than RNN and LSTM, and all the models seem to consider plausibility (i.e., no awkwardness) as well as well-formedness (i.e., no syntactic violation) of sentences. This demonstrates that deep learning can learn syntactic and, by extension, beyond syntactic information from native and learner data. This is reasonable because the four models learn information of a word by considering the words surrounding it, which is like the way Firth (1961) put forward: 'You shall know a word by the company it keeps'. This proposes the possibility of investigating nativeness that has been tricky and hard to prove.

One limitation of the analyses on Test Suite I is we cannot confirm why the models do not exactly classify INUMLC (spoken) and The Little Prince, both of which have relatively short sentences. To investigate short sentence length really confuses deep learning's judgments, in the future research, it is needed to exclude short sentences by establishing a certain threshold of length and compare results to those of this experiment.

4.2 Test Suite II

Table 3 shows the nativeness classification results on a supplemented version of test items from Park et al. (2020). As shown in the table, the accuracy of BERT and XLNet (66.6% and 65%, respectively) is again higher than RNN and LSTM (55.8% and 58.3%, respectively) on the entire test items.

The well-formedness judgment test items consist of 60 sentences that are subcategorized into (i) negative frequency adverb, (ii) the use of *hardly*, (iii) the collocation of the and same, (iv) overpassivization, (v) the use of middle verb, and (iv) be-insertion. Ill-formed sentences of each subcategory are instantiated in (5a) ~ (5f).

(5) a. *You hardly can breathe.
b. *John finds it hardly to talk with strangers.
c. *Mary felt same way about the incident.

		Well-formedness Test Items	Plausibility Test Items
Human Judgements (Park et al., 2020)		55/60 (91.6%)	35/50 (70.0%)
Deep Learning Judgments	RNN	34/60 (56.6%)	26/50 (52.0%)
	LSTM	34/60 (56.6%)	29/50 (58.0%)
	BERT	42/60 (70.0%)	**33/50 (66.0%)**
	XLNet	**44/60 (73.3%)**	30/50 (60.0%)

Table 4: Humans judgments in Park et al. (2020) and deep learning judgments in this paper

d. *A table was appeared.
e. *The articles are translating easily.
f. *Mary is drink water.

XLNet performs better than any other model on the well-formedness test items. If we exclude the well-formedness judgments on adverbs, the accuracy of every model increases: RNN/LSTM: 62.5%, BERT: 80%, XLNet: 85%, which seems to be caused by the difficulty of learning adverb positions; adverbs in English are relatively free in word order, and some adverbs such as negative frequency adverbs are strictly restricted to use while some are not, so deep learning probably feels hard to learn the proper usage of adverbs.

The plausibility judgment items are composed of 60 sentences that are subcategorized into (i) semantic prosody, (ii) semantic preference, (iii) the position of adverbs, (iv) the position of actually, (v) overcomplexity and (vi) collocation of words, each of which consists of 10 sentences. The sixth one is added in the current work to balance the number of plausibility test items with that of well-formedness test items. Implausible sentences of each category are exemplified in (6).

(6) a. A fantastic feast broken out.
b. The company is undergoing customer praise.
c. Also, I hope you're coming to our party tonight.
d. I made my car repaired, actually.
e. That I meet you makes me so happy.
f. Tom ate the pill.

XLNet is not the best but ranks third among the four models, which means XLNet tends to focus mainly on well-formedness of sentences rather than plausibility when predicting nativeness. BERT, On the other hand, gains the highest accuracy among the models, which shows BERT seems to have a more comprehensive view that considers both well-formedness and plausibility of sentences. The results can explain why BERT classifies KDJ and The Tanaka Corpus slightly better than XLNet (see Table 2): KDJ and The Tanaka Corpus are both syntactically well-formed, so from the perspective of XLNet, they deserve higher scores.

Notably, of plausibility test items, the ones for checking the knowledge about overcomplexity are particularly hard for the models to predict them correctly; the implausible items that all four models wrongly classify as plausible are just five sentences, three of which are overcomplex sentences. The three are instantiated in (7)

(7) a. It's the day before Monday.
b. John's becoming Mary's spouse is what he wants.
c. It's one-half of ten dollars.

Every model classifies them as native sentences, but native speakers feel awkward when reading them. An explanation for their wrong prediction is such that deep learning lacks a sense of economy. It is probably true that learners are not capable of making such sentences due to its syntactic complexity, so deep learning is likely to judge such syntactic complexity as a standard of native sentences.

As shown in Table 4 that compares human judgments in Park et al. (2020) to deep learning judgments in this paper, the results show that the accuracy of human judgments overwhelms that of deep learning judgments on well-formedness test items. XLNet is the closest one to native speakers, but there still be a big gap between them. Notably, on the other hand, there is not such a big difference between them on plausibility items; BERT, which is the most sensitive to plausibility as mentioned before, almost reaches the correct rate of humans (the difference of the number of correctly classified sentences is just two items).

	RNN	LSTM	BERT	XLNet
Individual Sentences (196 items)	50.5%	46.4%	54.5%	**56.1%**
Minimal Pairs (98 pairs)	4 pairs	4 pairs	**16 pairs**	14 pairs

Table 5: Nativeness judgment results on test items from DeKeyser (2000)

The participants in the paper are native English speakers. The current study, by extension, compares judgments of English learners to that of deep learning to investigate if there are any shared learnability problems. The next chapter is where this comparison is carried out.

4.3 Test Suite III

The grammaticality judgment test items from DeKeyser (2000) are pairwise minimal pair items consisting of 196 sentences (98 minimal pairs). Categorized into 11 types (27 subtypes), the test set is a highly refined set of items to measure test takers' knowledge in a wide variety of grammatical aspects. The term 'grammaticality' in the article is compatible with 'syntactic well-formedness' in the current paper, that is, this test suite is designed to consider only syntactic well-formedness rather than plausibility of sentences. In this sense, the most critical difference between Test Suite II and Test Suite III is whether they include implausible sentences or not. This test set is chosen for two reasons: (i) to examine deep learning's syntactic knowledge from a more integrated view and (ii) to investigate if there exist common learnability problems that both deep learning and L2 learners experience.

As shown in Table 5, the accuracies of every model are lower than what the models have on the test set from Park et al. (2020), which reveals the limitation that our models haven't learned various syntactic information. (8a) ~ (8k) are examples of 11 types for testing the knowledge of well-formedness.

(8) a. *Last night the old lady die in her sleep. (past tense)
b. *Three boy played on the swings in the park. (plural noun)
c. *John's dog always wait for him at the corner. (third-person singular)
d. *The little boy is speak to a policeman. (present progressive)
e. *Tom is reading book in the bathtub. (determiners)
f. *Peter made out the check but didn't sign. (pronominalization)
g. *The man climbed the ladder up carefully. (particle movement)
h. *George says much too softly. (subcategorization)
i. *Will be Harry blamed for the accident? (yes-no questions)
j. *What Martha is bringing to the party? (wh-question)
k. *The dinner the man burned. (word order)

Suppose the model had failed to learn the grammatical taxonomy of syntax, it would resort to simple heuristics that return probability value of plausibility when judging nativeness of sentences. The test set is largely composed of plausible sentences, so without knowledge of well-formedness, models are likely to classify them as native sentences. This is reflected in the number of sentences that each model predicts as native ones: RNN (159/196), LSTM (125/196), BERT (91/196), and XLNet (134/196). XLNet, which is relatively weak to capture plausibility, predicts the test items as native sentences even more frequently than LSTM (the accuracy of LSTM is also higher than XLNet on plausibility test items of Test Suite II, see Table 2). The results of RNN indicate that it is quite biased to classifying sentences as native ones compared to those of the other models, which is also shown in the 8 test sets (see Table 2); the gap between the highest and the lowest score that RNN gives to the 8 test sets is smaller than any other model. BERT, on the other hand, is the only model that the number of sentences predicted as native ones is lower than half of the test set. This indicates BERT does not have a biased judgement standard compared to the other models.

Minimal sentence pairs that the models classify both correctly are considerably low: RNN, LSTM, BERT and XLNet correctly predict just 4, 4, 16 and 14 pairs out of 98 pairs, respectively. Although the sentences correctly classified by BERT and XLNet are almost 4 times more than those by RNN

and LSTM, BERT and XLNet in the current work are still not good classifiers.

In DeKeyser (2000), the author sorts the test items into three groups into high (difficult), marginal (middle) and low (easy) groups based on the test results from the participants; if the difference of answers to a question among participants is big, the question is distinguished into the difficult group, and if answers to a question from most participant are similar, it is classified into the low group. We investigate if there are any common learnability problems between the subjects and our models; if what has been learned by deep learning and L2 learners are similar, learnability problems of the two agents are also likely to be shared. Presuming minimal pairs that the models predict both correctly are the ones that each model certainly has learned, the current study examines how the pairs are spread among the three levels of difficulty. The number of correctly predicted minimal pairs of 4 models is shown in Table 6.

	RNN	**LSTM**	**BERT**	**XLNet**
High	3/4	2/4	5/16	4/14
Marginal	0/4	0/4	6/16	3/14
Low	1/4	2/4	5/16	7/14

Table 6: Number of correctly classified minimal pairs

It seems that there are few learnability problems that deep learning and learners share because if there are shared problems, the correctly predicted pairs should have been the most in the easy group. From this result, we can assume that the aspects of learning of deep learning and learners are rather different.

5 Discussion and Conclusion

Much of related literature focuses only on syntactic well-formedness. This is partially because plausibility or nativelikeness is hard to define, and numerous contextual factors intervene clear categorization of them. However, for a sentence to be literally *well-formed*, satisfying syntactic rules only is not enough. In this sense, we attempt to examine sentences considering both syntactic well-formedness and plausibility.

The four models in this paper are evaluated from three angles. Firstly, the current work examined whether (and how) the models classify nativeness of sentences. The results of the first test suite, the 8 test sets, show that the deep learning models can correctly classify nativeness in a reasonable way, and we find with the results that deep learning can obtain knowledge of not just syntactic well-formedness but plausibility of sentences. Among the models, BERT and XLNet are more correct at nativeness judgments. BERT has more strength in capturing plausibility and XLNet is better at judging well-formedness.

Secondly, the models are investigated in terms of plausibility and well-formedness, and compared with nativeness judgments of English native speakers in a prior paper. The results reveal that XLNet is the best well-formedness classifier, but the native judgments overwhelm it. On plausibility items, however, we find that BERT almost reaches the level of native judgments.

Lastly, this paper evaluates the models with test items from DeKeyser (2000). The results indicate that our models are vulnerable to cover various kinds of syntactic violations. Learnability problems regarding obtaining syntactic information are not shared between deep learning and L2 learners, which means the learning strategies of deep learning and learners are quite different. DeKeyser (2000) explains the learnability problems of learners in terms of salience. The author observes that the more salient a grammatical factor, the easily and faster the factor is learned. For example, gender error is 'perceptually salient' because '[p]ronoun gender errors are so irritating to native speakers that they will almost always correct them when their nonnative interlocutors make such mistakes, […]' (DeKeyser, 2000). This case does not occur to deep learning.

In this paper, some similarities and differences between deep learning and humans are discovered. However, still the argumentation on the cognitive underpinning that integrates learning process of L2 learners and deep learning is not clearly developed. It follows that visualization of internal vector representation is required to endorse the assumption for the future research.

References

Bley-Vroman, R. (1988). The fundamental character of foreign language learning. Grammar and second language teaching: A book of readings, 19-30.

Chaves, R. P. (2020). What Don't RNN Language Models Learn About Filler-Gap Dependencies?. Proceedings of the Society for Computation in Linguistics, 3(1), 20-30.

Da Costa, J. K., & Chaves, R. P. (2020). Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. Proceedings of the Society for Computation in Linguistics, 3(1), 189-198.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. Studies in second language acquisition, 22(4), 499-533.

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.

Firth, J. R. (1961). Papers in Linguistics 1934-1951: Repr. Oxford University Press.

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities. arXiv preprint arXiv:1901.05287.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Linzen, T., & Baroni, M. (2020). Syntactic Structure from Deep Learning. arXiv preprint arXiv:2004.10827.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. arXiv preprint arXiv:1807.06882.

Mikolov, T., Karafiát, M., Burget, L., Č ernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association.

Park, K., You, S., & Song, S. (2019). Using the Deep Learning Techniques for Understanding the nativelikeness of Korean EFL Learners. Language Facts and Perspectives, 48, 195-227

Park, K., You, S., & Song, S. (2020). Not Yet as Native as Native Speakers: Comparing Deep Learning Predictions and Human Judgments. English Language and Linguistics, 26, 199-228.

Pawley, A., & Syder, F. H. (2014). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Language and communication (pp. 203-239). Routledge.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 7, 625-641.

Wilcox, E., Levy, R., & Futrell, R. (2019). What Syntactic Structures block Dependencies in RNN Language Models?. arXiv preprint arXiv:1905.10431.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).

A Simple Disaster-Related Knowledge Base for Intelligent Agents

Clark Emmanuel Paulo, Arvin Ken Ramirez, David Clarence Reducindo
Rannie Mark Mateo, Joseph Marvin Imperial
National University
Manila, Philippines
jrimperial@national-u.edu.ph

Abstract

In this paper, we describe our efforts in establishing a simple knowledge base by building a semantic network composed of concepts and word relationships in the context of disasters in the Philippines. Our primary source of data is a collection of news articles scraped from various Philippine news websites. Using word embeddings, we extract semantically similar and co-occurring words from an initial seed words list. We arrive at an expanded ontology with a total of 450 word assertions. We let experts from the fields of linguistics, disasters, and weather science evaluate our knowledge base and arrived at an agreeability rate of 64%. We then perform a time-based analysis of the assertions to identify important semantic changes captured by the knowledge base such as the (a) trend of roles played by human entities, (b) memberships of human entities, and (c) common association of disaster-related words. The context-specific knowledge base developed from this study can be adapted by intelligent agents such as chat bots integrated in platforms such as Facebook Messenger for answering disaster-related queries.

1 Introduction

The Philippines is a common ground for natural disasters such as typhoons and flooding. According to the latest statistics of Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAG-ASA)[1], there are more tropical cyclones entering the vicinity of the Philippines than anywhere else in the world. In a year, almost 20 tropical cyclones enter with 70% chance of developing into a full-blown typhoon. As possible result, catastrophic aftermath such as economic failure, destruction of infrastructures, and loss of lives may be imminent without proper information dissemination and preparedness. Thus, directing all research and technological efforts to disaster preparedness and disaster risk reduction to mitigate the tremendous impacts of natural calamities have been prioritized by the country for years.

According to Statista[2], the Philippines has 44 million active Facebook users in 2019 and is predicted to reach approximately 50 million by 2023. Thus, taking in consideration the disaster-prone situation of the country, we note the importance of establishing context-specific knowledge bases where it can be integrated to commonly used digital platforms such as Facebook Messenger to answer context-specific questions on disasters for possible information dissemination and awareness. To make this possible, we present our initial efforts in building a simple knowledge base composed of word concepts joined by semantic word relationships extracted from word embeddings trained from a large online news corpus on disasters.

The overview of the study is as follows: First, we compared and contrasted previous works done in field of ontology learning and mining information from word embeddings for building knowledge

[1]Statistics on annual tropical cyclones in the Philippines: http://bagong.pagasa.dost.gov.ph/climate/tropical-cyclone-information

[2]Number of Facebook users in the Philippines: www.statista.com/statistics/490455/number-of-philippines-facebook-users/

bases. Next, we discussed the method of extraction of concepts and entities from our news article dataset as well as the semantic labels used for building concept relationships. In the Results section, we detail the outcome of using word embeddings trained from our collected dataset for extracting semantically similar concepts using an initial concept seedwords list. In addition, we also discussed the integrity of the disaster ontology from expert validation. We wrap up the study by extensively discussing the semantic changes exhibited by concepts present in the knowledge base over time.

2 Previous Work

In this section we highlight significant works in using knowledge bases (KBs) as the main source of intelligence for virtual agents as well as current trends using word embeddings as one way of extracting information and semantic word relationships.

2.1 Building Knowledge-Based Intelligent Agents

Over the years, the use of powerful knowledge bases are seen on a wide variety of intelligent agents such as chat bots, storytelling agents, and recommender systems. The work of (Ong et al., 2018) focused on building a commonsense knowledge base for a storytelling agent for children using assertions extracted from ConceptNet (Speer et al., 2017). ConceptNet is a large semantic network of word relationships that can be adapted by intelligent agents for commonsense reasoning and identifying object relationships. The study developed a knowledge base by filtering out concepts and concept relationships that are not used in the context of children storytelling from the ConceptNet.

Similarly, Han et al. (2015) developed a natural language dialog agent that utilizes a knowledge base to generate diverse but meaningful responses to the user. The system extracts related information from knowledge base, which was adapted from FreeBase (Bollacker et al., 2008), via an information extraction module. The use of a large, external knowledge base allows the dialog system to expand on the information from the users response for a more detailed and interactive reply.

On the other hand, the work of Wang et al. (2010) focused on the development of a system composed of three ontology-based sub-agents for personal knowledge, fuzzy inference, and semantic generation for evaluating a person's health through his/her diet. The system makes use of an ontology with an embedded knowledge base considering the persons health statistics such as BMI, Caloric Difference, Health Diet Status combined with rules laid down by domain experts. Results shows that the proposed system exhibits an intelligent behavior in helping the dietary patterns of users based on their information from the constructed ontology.

2.2 Knowledge from Word Embeddings

The advent of word embeddings as one of the modern approaches in extracting semantic relationships of words has fueled research works to use its potential to build more powerful knowledge bases. Sarkar et al. (2018) used a supervised approach, similar to text classification, for predicting the taxonomic relationship via similarity of two concepts using a Word2Vec embedding (Mikolov et al., 2013b). Results showed that combining the word embedding with an SVM classifier outperformed baseline approaches for taxonomic relationship extraction such as using the Jaccard similarity formula and naive string matching.

The study of Luu et al. (2016) went as far as building a custom neural network architecture with dynamic weighting to significantly increase the performance of statistical and linguistic approaches in extraction word relationships from word embeddings. The neural network considers not only the word relationship such as hypernym and hyponym but also the contextual information between the terms. The proposed approach exhibits generalizability for unseen word pairs and has obtained 9% to 13% additional accuracy score using a general and domain-specific datasets.

Likewise, the work of Pocostales (2016) submitted to the SemEval-2016 Task for Taxonomy Extraction Evaluation (Bordea et al., 2016) focused on using GloVe word embedding model (Pennington et al., 2014) with an offset feature to extract hypernym candidates from a sample word list. Results showed that a vector offset cannot completely capture the hypernym-hyponym relationship of words due to complexity.

2017	Tag	2018	Tag	2019	Tag
family	noun	*act*	verb	*bulletin*	noun
fire	noun	*announced*	verb	*quake*	noun
flood	noun	*ashfall*	noun	*typhoon*	noun
update	verb	*police*	noun	*weakened*	adjective
tricycle	noun	*lava*	noun	*affected*	verb

Table 1: Top 5 seed words per year.

3 Data

For this study, we scraped over 4,500 Filipino disaster-related news articles from years 2017 to 2019 (1,500 articles per year) from Philippine news websites using Octoparse Webscraping Tool as our primary dataset. The corpus covers a wide range of natural disasters that transpired in the Philippines such as typhoons, earthquakes, landslides and also includes statistics from damages and casualty reports. This large collection of news articles will be the groundwork of the knowledge base as it contains disaster-related context words as concepts. We partitioned the dataset into three by year (2017 to 2019) for the word embedding model generation. The purpose of partitioning the dataset will allow us to analyze the temporal changes of sematic relationships of concepts. More on this is discussed in the succeeding sections. To note, all concepts presented in this document are translated to English for the international audience.

4 Building the Knowledge Base

4.1 Extracting Concept Seedwords

Building a knowledge base starts with establishing concepts or words that refer to real world entities such as **nouns**, **adjectives**, and **verbs** that represent everyday objects such as *apple, spoon, cake*, people like *mother, police, mayor*, description of objects such as *beautiful, red, big*, and action words that signify an activity such as *walks, eating* and *jumped* (Ong, 2010).

To identify the grammatical category of each concept to know whether it is a noun, an adjective, or a verb to aid the semantic relationship labelling, we used a Filipino parts-of-speech (POS) tagger[3] developed by Go and Nocon (2017) which is currently integrated in the Stanford CoreNLP package (Manning et al., 2014). We extracted the top 50 high-frequency concepts per year from the collected news dataset, having a total of 150 initial seed words. Table 1 shows the top 5 seed words per year from the initial word list. Both common words such as *family*, *tricycle*, *police* and *update* as well as disaster-related words such as *flood*, *ashfall*, *typhoon*, and *quake* to name a few are present. These concepts are then paired with other concepts to form a meaningful representation of knowledge called a **binary assertion** described in the next section.

4.2 Semantic Relationship Labelling

Once the set of context-specific words are obtained from the dataset, in the case of this study, disaster-related concepts, the next process to establish the correct semantic relationship of words. These semantic relationships can be structured in the form of a binary assertion as previously stated. Ong et al. (2018) stated that the binary assertions of concepts are needed by virtual agents such as a story-telling agent or a chatbot to be able to generate responses from a commonsense knowledge base. Semantic relationships can also be used for other tasks such query expansion of words as well as information retrieval (Attia et al., 2016). For this study, we adapt the binary assertion format used by ConceptNet (Speer et al., 2017) shown below:

`[concept1 semantic-rel concept2]`

where `semantic-rel` stands for the specified semantic relationship of the two concepts (`concept1` and `concept2`) that contains meaning. We adapted the six semantic relation labels from the CogALex-2016 Shared Task on Corpus-Based Identification of Semantic Relations (Santus et al., 2016) which are **Synonym (SYN) Antonym (ANT)**, **Hypernym (HYP)**, **Membership (PartOf)**,

[3] github.com/matthewgo/FilipinoStanfordPOSTagger

Relation	Tag	Rule
Synonym	SYN	If `concept1` has the same meaning with `concept2`
Antonym	ANT	If `concept1` has the opposite meaning with `concept2`
Hypernym	HYP	If `concept1` has a broader meaning compared to `concept2`
Performs	DO	If `concept1` is the actor/does of `concept2`
Membership	PartOf	If `concept1` is a member of `concept2`
Adjective	IS	If `concept1` describes `concept2`
Cause	CAUSE	If `concept1` is the cause of event `concept2`
Effect	dueTo	If `concept1` the resulting effect of event `concept2`
Random	RAND	If `concept1` has no direct relationship `concept2`

Table 2: Semantic relations with its corresponding rules and examples.

and **Random (RAND)** as shown in Table 2. In addition, we also added a few semantic labels of our own with respect to the concepts that we will be working on this study which are in the field of disasters. Thus, we added **Performs (DO)** to signify action, **Adjective (IS)** to signify description, and **Cause (CAUSE)** and **Effect (dueTo)** to signify consequences of events in a disaster setting.

4.3 Concept Expansion using Word Embeddings

Word embeddings are representations of an entire document vocabulary in a mathematical vector space. Each word is represented with a set of real-valued numbers given a specific set of dimensions. Word embedding architectures such as **Word2Vec** (Mikolov et al., 2013a; Mikolov et al., 2013b) and **Global Vectors** or **GloVe** (Pennington et al., 2014) capture various relationships of words in a corpus such as the **semantic similarity**, **syntactic similarity**, and **word co-occurrence** to name a few. Thus, if two words are commonly used together in the same context, such as in disaster-related articles, we can expect them to be close together when represented in the vector space. For example, finding the most similar terms using the query word *warning* from a word embedding model trained on a disaster-related dataset will output words such as *typhoon*, *flooding*, and *signal* since the word *warning* is commonly used in texts to notify people of possible natural disasters.

For this study, we generated a word embedding model for each year of partitioned news article dataset from 2017 to 2019. A total of three models were generated using the Word2Vec architecture (Mikolov et al., 2013b). In addition, we used the initial seedwords list which contains 150 concepts (50 for each year) as query words to extract semantically similar terms which occur in the context of disaster.

5 Results

Table 3 shows the expanded ontology by querying four sample seedwords from the word embedding model. We only filtered the top three semantically similar resulting words from each query word that falls under a manually-annotated and qualified semantic label discussed in Section 4.2. From this, we arrived at an expanded ontology composed of 450 assertions for our disaster-related knowledge base.

From the resulting expanded ontology, it already provides us the *knowledge* that an intelligent agent can piece together or form when asked about something in the context of disasters. Take in, for example, the word *tremor*. Obtaining the top three assertions with semantic labels and expanded using word embeddings informs us that the word *tremor* is synonymous and interchangeable with the word *af-*

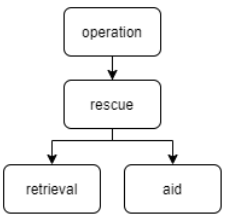


Figure 1: Hierarchy of hypernyms formed from the expanded ontology.

Seedword	Semantic Label	Assertions
police	DO	[police DO risk]
	IS	[police IS armed]
	DO	[police DO commitment]
tremor	SYN	[tremor SYN aftershock]
	dueTo	[tremor dueTo quake]
	dueTo	[tremor dueTo earthquake]
rescue	partOf	[rescue partOf operations]
	HYP	[rescue HYP retrieval]
	HYP	[rescue HYP aid]
experts	IS	[experts IS supporting]
	DO	[experts DO recommend]
	DO	[experts DO impose]

Table 3: Expansion of knowledge base using word embeddings.

tershock which scientifically means the involuntary movement of the surface due to breaking of underground rocks. Likewise, the words *quake* and *earthquake* are annotated with the semantic label *dueTo* which denotes effect since earthquakes is root cause of tremors according to scientific definition (Yose, 2013).

Another observation from the expanded ontology is the knowledge of understanding general actions to more specific ones. In the example, the action word *rescue* is a **hypernym**, which means it is a general word that can be possibly specified further, in this case, it is the hypernym of the word *aid*. Consequently, the word *operation* is a hypernym of *rescue*. Combining the words altogether, we can interpret the three assertions as a string of successive actions as shown in Figure 1 where in an *operation* involves a *rescue* and the meaning of *rescue* may vary such that it can be some form of (a) *retrieval* of missing people or (b) *aid* for the wounded or stranded people.

Lastly, the word embedding model was able to produce words that imply *responsibilities* by concepts in the form of human entities. In the two examples of human entities shown in Table 3, *experts* and *police*, most of the semantic labels are **DO** which denotes **action** and **IS** which denotes **description**. For *police*, common known actions done in the context of disasters are *risk* and *commitment* while being *armed*. For *experts*, the responsibilities are in the line of *supporting* a claim as well as *recommending* and *imposing* future actions based on scientific knowledge.

We observe the significant potential of using word embeddings for expanding knowledge bases by capturing various levels of information such as (a) understanding interchangeable context-specific words, in the case of this study on disaster-related words, and (b) understanding roles played by human entities described in this section. When used by an intelligent agent, it will have an idea of what response can be produced when queried with questions such as *"What does a police officer do?"* or *"What happens after an earthquake?"*.

5.1 Expert Validation

To properly gauge the effectivity of using word embeddings for the expansion of knowledge bases, we performed a validation process by inviting three experts in the fields of linguistics, disaster response, and meteorology to evaluate the assertions of the knowledge base. Each assertion is evaluated using a **two-scale metric**: **Agree** if the expert deems that the assertion observes a correct relationship and semantic labelling ([*doctor* SYN *medical person*]) or **Disagree** if the assertion is not properly labelled to form a correct relationship ([*flood* SYN *typhoon*]).

Table 4 shows the averaged agreeability rate of the expert validation process. Results show that the most accurate model with the highest rating is the Word2Vec model trained on disaster-related news articles collected in the year of 2019. This is fol-

Model	Aggreability Rate
Word2Vec 2019	0.64
Word2Vec 2018	0.52
Word2Vec 2017	0.49

Table 4: Expert validation for the knowledge base.

lowed by word embedding models using the 2018 and 2017 dataset. We attribute the semi-low agreeability scores due to the variation of word usage of the experts in their corresponding fields of study. For example, the assertion [*tremor* SYN *earthquake*] was evaluated by the linguist and disaster experts as **Agree** while the meteorologist contested. The expert on meteorology swears by the scientific definition of which tremors are very different with earthquakes in a way that tremors are *caused* by earthquakes and not an interchangeable term as perceived by the other two non-technical experts. This pattern is also observed with other assertions using the synonym labelling such as [*storm* SYN *typhoon*] and [*lava* SYN *magma*].

6 Discussion

In this section, we perform an even more in-depth analysis by considering the changes in semantic meaning or semantic information of the assertions of the knowledge base over time. We break the discussion into three categories we observed from the analysis.

6.1 Roles Played by Human Entities

We observe the changing roles played by three essential human entities, *governor*, *experts*, and *police* over time in the setting of a natural disaster as written in news articles. These entities are expected to be on full alert and their responsibilities are crucial towards mitigating the consequences of disasters.

The entity *governor*, or the highest commanding individual of Philippine province, performs various roles over time. As seen in Table 5, the entity is mostly connected by the semantic label **DO** to denote action with concepts such as *assure*, *explain*, *oversee*, *develop*, and *declare*. There are also a few description words connected by the label **IS** such as *political* and *mandatory* which is obvious since holding a gubernatorial position is indeed *political* and resolutions filed in a governor's office is in its essence *mandatory*.

Similarly, the entity *experts* is also expected to have various changing roles. For 2017 as seen in the Table, common action words associated are *checking* and *verifying* which provides one of the most important roles of experts in the field of disasters: validating the integrity of information being publicized. For 2018, experts play more of an information dissemination role with the concept *announce* as the connecting action word. In 2019, experts assumed a more stricter role as observed with the connected concepts such as *recommend* and *impose*.

For the entity *police*, there is consistent associated descriptor across the years regarding their responsibility: *risk*. This provides us a concrete idea of a consequence when assuming the role of a police. In 2017 and 2018, strong action words such as *control*, *damage*, and *armed* are tied with a policeman's job. In 2019, however, the entity *police* assumed a more passive role as more of an informant with the connected action words being *study* and *alerts*.

6.2 Memberships of Human Entities

We also observed changes in memberships of human entities in context of disasters over time as seen in Table 6. Memberships are denoted by the semantic labels **HYP** for **hypernyms** or general words and **partOf** for **hyponyms** more specific words. We observe memberships played of three entities: *mayor*, *student*, *teacher*.

For the entity *mayor*, memberships are more specific in 2017 and 2018. The entity is expected to partner with small groups in a municipality such as a *sitio* or a barangay cite as well as with large government agencies such as Division on the Welfare of the Urban Poor (DWUP) and Department of Environment and Natural Resources (DENR) in times of disasters. There are also general memberships such as *administration* and *government* to which the entity is an obvious member.

In the case of the entity *student*, there is a mix of specific and generalized memberships. Certain specific universities such as *UP* or the University of the Philippines and *PUP* or Polytechnic University of the Philippines were classified as institutions where a student may belong. General and obvious concepts such as *elementary*, *school*, *organization* and *Uni-*

Seed	2017	2018	2019
governor	[governor DO assure] [governor DO giving] [governor DO explain]	[governor IS political] [governor DO oversees] [governor IS mandatory]	[governor DO declare] [governor DO resolution] [governor DO develop]
experts	[experts SYN representative] [experts DO checking] [experts DO verify]	[experts DO work] [experts IS frantic] [experts DO announce]	[experts IS supporting] [experts DO recommend] [experts DO impose]
police	[police DO risk] [police DO control] [police DO damage]	[police DO risk] [police IS armed] [police DO commitment]	[police DO risk] [police DO study] [police DO alerts]

Table 5: Roles played by human entities.

Seed	2017	2018	2019
mayor	[mayor partOf hall] [mayor partOf DWUP] [mayor partOf sitio]	[mayor partOf administration] [mayor DO communication] [mayor partOf DENR]	[mayor RAND workers] [mayor partOf government] [mayor RAND employers]
student	[student partOf PUP] [student IS victim] [student IS resident]	[student partOf elementary] [student partOf school] [student IS minor]	[student partOf organization] [student partOf University] [student partOf UP]
teacher	[teacher DO education] [teacher partOf DepEd] [teacher partOf school]	[teacher DO research] [teacher DO education] [teacher partOf school]	[teacher partOf house] [teacher SYN employee] [teacher partOf school]

Table 6: Memberships of human entities.

versity contribute to the commonsense information of the knowledge base.

For the entity *teacher*, general concepts of memberships are more prominent compared to the other two entities. The concept *school* is consistent for all years. The term *DepEd* which means Department of Education, the government agency responsible in shaping the educational landscape of the Philippines, is connected to the entity. DepEd oversees elementary and intermediate level schools which may mean that the term *teacher* is commonly tied with educators from these levels. Interestingly, the concept *house* is also tied with the entity *teacher*. Although it may already be obvious that a teacher assumes a different role inside the house maybe as a mother or a breadwinner.

6.3 Common Word Association of Disaster-Related Terms

For the last category, we observe change in association of disaster-related words over time as seen in Table 7. We highlight the importance of this analysis to understand how a knowledge base using information extracted from word embeddings interchange co-occurring and similar words in the context of disasters. For this category, we analyze the four frequently occurring natural disasters in the Philippines which are **earthquakes**, **eruption**, **landslide**, and **typhoon**.

The first disaster-related word is *earthquake*. From the formed assertions for all years, *earthquakes* are synonymous with the term *quakes* which denote a shortened version of the word. In addition, the term *magnitude* is also a consistent term connected to the seedword *earthquake* which tells us that earthquakes have their own corresponding *magnitudes* quantified by some number.

In the case of *eruption*, most assertions formed are used with the semantic label **IS** to denote **description** or **property**. Across the years, most of the descriptive words associated with the term are negative such as *amplifying*, *hazardous*, *destruction*, *explo-*

Seed	2017	2018	2019
earthquake	[earthquake IS magnitude] [earthquake SYN quake] [earthquake RAND drill]	[earthquake IS magnitude] [earthquake SYN quake] [earthquake IS intensity]	[earthquake IS magnitude] [earthquake SYN quake] [earthquake RAND signal]
eruption	[eruption IS amplifying] [eruption IS fast] [eruption IS confirmed]	[eruption IS happening] [eruption IS explosive] [eruption CAUSE destruction]	[eruption IS hazardous] [eruption IS magmatic] [eruption IS happening]
landslide	[landslide dueTo flood] [landslide HYP mudslide] [landslide RAND area]	[landslide dueTo rain] [landslide IS torrential] [landslide HYP mudslide]	[landslide RAND mountain] [landslide dueTo rains] [landslide IS widespread]
typhoon	[typhoon IS expected] [typhoon partOf calamity] [typhoon SYN onslaught]	[typhoon SYN storm] [typhoon IS super] [typhoon IS powerful]	[typhoon dueTo Amihan] [typhoon SYN hurricane] [typhoon SYN cyclone]

Table 7: Disaster-related terms and its associated words.

sive, and *magmatic*. These word may denote a sense of urgency compared to the other natural disasters when it happens. The term *happening* has occurred both in 2018 and 2019 due to two of the most active volcanoes in the Philippines, Mount Mayon and Mount Taal, had shown activity[4] by erupting successively and spewing ash. The eruption caused mass postponement of public activities and over 48,000 evacuated locals.

For the concept *landslide*, the common associated word is *mudslide*. Although by scientific definition, mudslides are as specific type landslides (also called *debris flow*). Thus, the *landslide* term conforms the the semantic label **HYP** for hypernyms. The cause of landslides can be tied to *floods*, which in turn, caused by *rains* as joined by the semantic label **dueTo** denoting a **consequence** or an **effect**. In addition, landslides are also often described using the words *widespread* and *torrential* which gives us a quantifiable idea of the magnitude of landslides that occur in the Philippines.

Lastly, we have the word *typhoon*. This disaster-related concept assumes many interchangeable and synonymous terms such as *storm*, *hurricane*, *cyclone*, and *onslaught*. We note the frequent association and interchangeability of these words due to geographic locations (Khadka, 2018) but may mean the same thing—they are all **tropical storms**. Likewise, description words tied to the concept *typhoon* are *expected*, *super*, and *powerful* which also provides us an idea of the magnitude of typhoons occurring in the country similar to *landslides*.

[4]Volcano Bulletin: phivolcs.dost.gov.ph/index.php/volcano-hazard/volcano-bulletins3

7 Conclusion

As a disaster-prone country, the Philippines should increase its efforts in mitigating the effects of natural calamities with the help of technology. One way to do this is to consider the potential of intelligent agents such as chatbots as tools for disaster preparedness and information dissemination. In this study, we established a simple context-specific knowledge base by doing three important processes: (a) extracting disaster-related concepts from a collected news article dataset, (b) building a network of binary assertions from a curated list of semantic labels, and (c) expanding the ontology by querying an initial seedwords list from word embeddings generated from the original news dataset. Results show that using word embeddings captured various levels of information that may be useful for intelligent agents to produce responses such as information on roles of human entities, generalization and specification of terms, and common word association when asked in the topic of disasters. Future directions of the study include collection of even more dataset that covers not only news articles but also other media for finer-grained assertions. In addition, the study will also benefit from efforts in testing the capability of the knowledge base in practical applications.

References

Mohammed Attia, Suraj Maharjan, Younes Samih, Laura Kallmeyer, and Thamar Solorio. 2016. Cogalex-v shared task: Ghhh-detecting semantic relations via word embeddings. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 86–91.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091.

Matthew Phillip Go and Nicco Nocon. 2017. 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*, pages 81–88.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133.

Navin Singh Khadka. 2018. Hurricanes, typhoons and cyclones: What's the difference?, Sep.

Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Dionne Tiffany Ong, Christine Rachel De Jesus, Luisa Katherine Gilig, Junlyn Bryan Alburo, and Ethel Ong. 2018. Building a commonsense knowledge base for a collaborative storytelling agent. In *Pacific Rim Knowledge Acquisition Workshop*, pages 1–15. Springer.

Ethel ChuaJoy Ong. 2010. A commonsense knowledge base for generating children's stories. In *2010 AAAI Fall Symposium Series*.

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.

Joel Pocostales. 2016. Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1298–1302.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. 2016. The CogALex-v shared task on the corpus-based identification of semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 69–79, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Rajdeep Sarkar, John Philip McCrae, and Paul Buitelaar. 2018. A supervised approach to taxonomy extraction using word embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Mei-Hui Wang, Chang-Shing Lee, Kuang-Liang Hsieh, Chin-Yuan Hsu, Giovanni Acampora, and Chong-Ching Chang. 2010. Ontology-based multi-agents for intelligent healthcare applications. *Journal of Ambient Intelligence and Humanized Computing*, 1(2):111–131.

Joash Yose. 2013. Earthquake or tremor: What's the difference?, Jul.

Effective Approach to Develop a Sentiment Annotator For Legal Domain in a Low Resource Setting

Gathika Ratnayaka[*,1], Nisansa de Silva[*], Amal Shehan Perera[*], and Ramesh Pathirana[+]

[*]Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka
[+]Faculty of Law, University of Colombo, Sri Lanka
[1]gathika.14@cse.mrt.ac.lk

Abstract

Analyzing the sentiments of legal opinions available in Legal Opinion Texts can facilitate several use cases such as legal judgement prediction, contradictory statements identification and party-based sentiment analysis. However, the task of developing a legal domain specific sentiment annotator is challenging due to resource constraints such as lack of domain specific labelled data and domain expertise. In this study, we propose novel techniques that can be used to develop a sentiment annotator for the legal domain while minimizing the need for manual annotations of data.

1 Introduction

Legal Opinion Texts that elaborate on the incidents, arguments, legal opinions, and judgements associated with previous court cases are an integral part of case law. As the information that can be acquired from these documents has the potential to be directly applied in similar legal cases, legal officials use of them as information sources to support their arguments and opinions when handling a new legal scenario. Therefore, developing methodologies and tools that can be used to automatically extract valuable information from legal opinion texts while deriving useful insights from the extracted data are of significant importance when it comes to assisting legal officials via automated systems.

Sentiment analysis can be considered as one such information extraction technique that has a significant potential to facilitate various information extraction tasks. When a legal case is considered, it is built around two major parties that are opposing to each other. The party that brings forward the lawsuit is usually called the plaintiff and the other party is known as the defendant. At the beginning of a legal opinion text, a summary of the case is given, describing the incidents associated with the case and also explaining how each party is related with those incidents. Legal opinions or the opinions of judges about the associated events and laws related to the court case can be considered as the most important type of information available in a legal opinion text. Such opinions may have a positive, neutral, or negative impact on a particular party. In addition to the opinions that are directly related to the conduct of the parties, legal opinion texts also provide interpretations related to previous judgements and also on statutes that are relevant to the legal case. Such opinions may elaborate on the justifications, purposes, drawbacks and loopholes that are associated with a particular statute or a precedent. Moreover, the descriptions also contain information related to the proceeding of court cases such as adjournment of the case and lack of evidence which can be considered as factors that can directly have an impact on the outcomes. When all of the above mentioned factors are considered, sentiment analysis on legal opinion texts can be considered as a task that can facilitate a wide range of use cases. Despite its potential and usefulness, the attempts to perform sentiment analysis in legal domain are limited. This study aims to address this issue by developing a sentiment annotator that can identify sentiments in a given sentence/phrase extracted from legal opinion texts related to the United States Supreme Court.

Information that can be derived from such a sentiment annotator can then be adapted to facilitate more downstream tasks such as identifying advantageous and disadvantageous arguments for a particular party, contradictory opinion detection (Ratnayaka et al., 2019), and predicting outcomes of legal cases (Liu and Chen, 2018) .

In order to develop a reliable sentiment annotator using supervised learning, it is required to have a large amount of labelled data to train the underlying classification model. However, creating such sophisticated datasets with manually annotated data (by domain experts) for a specialised domain like legal opinion texts is not practical due to extensive resource and time requirements (Gamage et al., 2018; Sharma et al., 2018). In a low resource setting, transfer learning can be used as a potential technique to overcome the requirement of creating a sophisticated data set. Adapting these models directly into the legal domain will create drawbacks, especially due to the negative transfer; which is a phenomenon that occurs due to dissimilarities between two domains. Domain specific usage of words, domain specific sentiment polarities and meanings of words can be considered as one major reason that causes negative transfer when adapting datasets/models from one domain to another domain (Sharma et al., 2018). In this study, we demonstrate methodologies that can be used to overcome drawbacks due to negative transfer, when adapting a dataset from a source domain to the legal domain.

2 Related Work

It can be observed that the early attempts (Thelwall et al., 2010) of developing automatic sentiment analysis mechanisms make use of sentiment lexicons such as AFINN (Nielsen, 2011), ANEW (Bradley and Lang, 1999), and Sentiwordnet (Baccianella et al., 2010). The sentiment polarity and the strength of a particular word change from one lexicon to another depending on the domain that is being considered when developing the lexicon (Nielsen, 2011) due to the domain-specific behaviors of words. However, in recent works related to sentiment analysis that are based on machine learning and deep learning techniques, the learning algorithms are allowed to learn the sentiments associated with words and how their compositions affect the overall sentiment of a particular text. The Recursive Neural Tensor Network (RNTN) model proposed by Socher et al. (Socher et al., 2013) can be considered as an important step towards this direction and it has shown promising results for sentiment classification in movie reviews. However, the performances of such approaches that are based on recursive neural network architectures have been surpassed more recently by the approaches that make use of pretrained language models (eg: BERT(Devlin et al., 2018)), and such approaches have now become the state of the art for sentiment classification (Munikar et al., 2019). From this point onwards, the RNTN model proposed in (Socher et al., 2013) will be denoted as $RNTN_m$. Though the applications of sentiment analysis in the legal domain are limited, there is an emerging interest within the law-tech community to explore how sentiment analysis can be used to facilitate the legal processes(Conrad and Schilder, 2007; Liu and Chen, 2018). The study by Gamage et al. (Gamage et al., 2018) on performing sentiment analysis in US legal opinion texts can be considered as the closest to our work. However, the direct applicability of their approach into our study is prevented due to some limitations. In (Gamage et al., 2018), the sentiment annotator was developed to perform a binary-classification task (negative sentiment and non-negative sentiment). Moreover, one of the key steps in (Gamage et al., 2018) is to identify words that have different sentiments in the legal domain when compared with their sentiments in the movie domain. The identification of such words with domain-specific sentiments had been performed manually by human annotators. However, manually going through a set of words with a significant size is not ideal for a low resource setting in which the intention is to create an optimum outcome from a limited amount of human annotations. Though (Sharma et al., 2018) proposes an automatic approach based on word embeddings to minimize negative transfer by identifying transferable words that can be used for cross domain sentiment classification, the proposed approach aims only at binary sentiment classification that considers only the positive and negative sentiment classes.

3 Methodology

3.1 Identifying words that can cause negative transfer

In order to minimize the resource requirements, our intention is to utilize a labeled high resource source domain to facilitate sentiment analysis in a low resource target domain. The Stanford Sentiment Treebank (SST-5) (Socher et al., 2013) which consists of Rotten Tomato movie reviews labelled according to their sentiments was taken as the source dataset and a corpus of legal opinion texts was selected to extract legal phrases that will be used as the target dataset. As the first step, 3 categories were identified to which the words available in the source dataset can be assigned. The first category is the *Domain Generic words*, the words that behave in a similar manner across the movie review domain and the legal domain. The second category is the *Domain Specific words*, the type of words that behaves differently in the two domains and has the potential to cause negative transfer. Within this category, the most frequently used sense/meaning of a word in one domain may differ from that of the other domain. Additionally, such a word may have different sentiment polarities across the two domains. However, there is another important type of words that can be identified as *Under Represented Words*. The set of *Under Represented Words* consists of words that are frequently occurring in the target domain (legal domain), but are not available or have occurred with a very less frequency in the source dataset.

Due to the resource limitations, it is not feasible to identify *domain specific words*, *domain generic words*, and *under represented words* manually by going through each word in the legal opinion text corpus. Therefore, the following steps were followed to minimize the requirements for manual annotation. As the first step, stop words in the legal opinion text corpus were removed utilizing the Van stop list (Van Rijsbergen, 1979). Next *word frequency*, which is the frequency of occurrence of a particular word within the corpus was calculated for each word. Then, the set of words was arranged in a descending order based on the word frequency to create the sorted word set W. From W, first k-words (most frequent k-words) were chosen as the considered set of words S. Here $k = \min_j\{j \in \mathbb{Z}^+ | \sum_{i=1}^{j}(w_i) \geq 0.95 \cdot \sum_{i=1}^{n}(w_i)\}$, where w_i is the i^{th} element of W and n is the total number of elements in W.

Algorithm 1

Function $assignSentiment_o(w, sentiment)$
  if sentiment == N **then** $D_{on} \cup \{w\}, O_i - \{w\}$
  else if sentiment == P **then** $D_{op} \cup \{w\}, O_i - \{w\}$
  end if
EndFunction
Function $assignSentiment_n(w, sentiment)$
  if sentiment == N **then** $D_{nn} \cup \{w\}$
  else if sentiment == P **then** $D_{np} \cup \{w\}, N_i - \{w\}$
  else if sentiment == O **then** $D_{no} \cup \{w\}, N_i - \{w\}$
  end if
EndFunction
Function $assignSentiment_p(w, sentiment)$
  if sentiment == N **then** $D_{pn} \cup \{w\}, P_i - \{w\}$
  else if sentiment == P **then** $D_{pp} \cup \{w\}$
  else if sentiment == O **then** $D_{po} \cup \{w\}, P_i - \{w\}$
  end if
EndFunction
$P_i = P_m, N_i = N_m, O_i = O_m, D_{on} = \{\}, D_{op} = \{\}$
$D_{nn}, D_{np}, D_{no}, D_{pp}, D_{pn}, D_{po} = \{\}$
n=0,p=0
While $1 + |D_{on}| > n$ or $1 + |D_{op}| > p$ **do**
  n=$1 + |D_{on}|$, p =$1 + |D_{op}|$
  for word w in O_i **do**
    l = $mostSimilar_l(w)$
    if underRepresented(w) and affinAssignable(w) **then**
      $assignSentiment_o(w, afinn(w))$
    else if domainSpecific(w) and affinAssignable(w) **then**
      $assignSentiment_o(w, afinn(w))$
    else if domainGeneric(*l*) and $l \in N_m \cup D_{on}$ **then**
      if $notAntonym(w, l)$ **then** $assignSentiment_o(w, N)$
    else if domainGeneric(*l*) and $l \in P_m \cup D_{op}$ **then**
      if $notAntonym(w, l)$ **then** $assignSentiment_o(w, P)$
    end if
  end for
end

Next, the Stanford Sentiment Annotator ($RNTN_m$) was used to annotate the sentiment of each word in the considered word set S. After the annotation process, the words were distributed into three sets P_M, N_M, O_M based on the annotated sentiment. The set P_M is made up of words that were annotated as Very Positive or Positive and the set N_M is made up of words that were annotated as Very Negative or Negative. The words that were annotated as having a Neutral sentiment were included into O_M. The sets P_M, N_M, O_M consists of 336, 253, and 4992 words respectively. Identifying words in O_M that have different sentiments across the two domains by manually going through each word is resource extensive as it contains nearly 5000 words. To overcome this challenge and to minimize

the required number of manual annotations, we developed a heuristic approach to identify words in the neutral word set (O_M) that can have different (deviated) sentiments. It should also be noted that in our algorithmic approach, words with deviated sentiments are identified while automatically assigning each word with a legal sentiment (Algorithm 1 and Algorithm 2).

Algorithm 2

n=0,p=0
While $1 + |D_{nn}| > n$ or $1 + |D_{np}| > p$ **do**
  n=$1 + |D_{nn}|$, p =$1 + |D_{np}|$
  Q = $N_i \cup D_{on} \cup D_{nn}$, R = $P_m \cup D_{op} \cup D_{np}$
  for word w in N_i **do**
    l = $mostSimilar_l(w)$
    if domainGeneric(w) **then** $assignSentiment_n(w, N)$
    else if domainSpecific(w) and affin(w)==N **then**
      $assignSentiment_n(w, N)$
    else if domainSpecific(w) and notAntonym(w,l) **then**
      if $l \in Q$ **then** $assignSentiment_n(w, N)$
      else if domainGeneric(l) and $l \in R$ **then**
        $assignSentiment_n(w, P)$
    end if
  end for
end
for word w in N_i **do** $assignSentiment_n(w, O)$
n=0,p=0
While $1 + |D_{pp}| > p$ or $1 + |D_{pn}| > n$ **do**
  p=$1 + |D_{pp}|$, n =$1 + |D_{pn}|$
  Q = $N_m \cup D_{on} \cup D_{pn}$, R = $P_i \cup D_{op} \cup D_{pp}$
  for word w in P_i **do**
    l = $mostSimilar_l(w)$
    if domainGeneric(w) **then** $assignSentiment_p(w, P)$
    else if domainSpecific(w) and affin(w)==P **then**
      $assignSentiment_p(w, P)$
    else if domainSpecific(w) and notAntonym(w,l) **then**
      if $l \in R$ **then** $assignSentiment_p(w, P)$
      else if domainGeneric(l) and $l \in Q$ **then**
        $assignSentiment_p(w, N)$
    end if
  end for
end
for word w in P_i **do** $assignSentiment_p(w, O)$

$P_l = D_{op} \cup Dnp \cup Dpp$, $N_l = D_{on} \cup Dnn \cup Dpn$

Though it is feasible to manually annotate all the words in P_M and N_M, we have developed our algorithmic approach to identify words that can have deviated sentiments in P_M and N_M as well (Algorithm 2) because having a heuristic approach to identify such deviated words can be used to minimize the number of annotations required in case a significant number of words will be identified from O_M as having deviated sentiments exceeding the annotation budget. Moreover, such an automatic approach has the potential to be utilized as a mechanism to generate domain specific sentiment lexicons.

Within our approach to distinguish domain specific words from domain generic words, two key information that can be derived from word embedding models are considered; 1. Cosine similarity between vector representations of two words u, v as $Cosine_{domain}(u, v)$ and the most similar word for a particular word w as $mostSimilar_{domain}(w)$. Domain specific word embeddings have been utilized within our approach to identify domain specific words from domain generic words. The Word2Vec model publicly available at SigmaLaw dataset (Sugathadasa et al., 2017) that has been trained using a United States legal opinion text corpus was selected as the legal domain specific word embedding model. The SST-5 dataset does not contain an adequate amount of text data to be used as a corpus to create an effective word embedding model. Therefore, we selected the IMDB movie review corpus (Maas et al., 2011) to train the movie review domain specific Word2Vec embedding model. From this point onwards, $Cosine_{legal}$ and $Cosine_{movie-reviews}$ will be denoted by $Cosine_l$ and $Cosine_m$ respectively. Similarly, $mostSimilar_{legal}(w)$ will be denoted by $l(w)$ while using $m(w)$ to denote $mostSimilar_{movie-reviews}(w)$.

First, for a given word w, we obtain $l(w)$ and $m(w)$. As Word2Vec (Mikolov et al., 2013) embeddings are based on distributional similarity, it can be assumed that the most similar word output by a domain specific embedding model to a particular word is related to the domain specific sense of that considered word. For example, *convicted* is obtained as $l(charged)$. It can be observed that the word convicted is associated with the sense of accusation, which is the most frequent sense of *charge* in the legal domain. However, when it comes to $m(charged)$, *sympathizing* is obtained as the output. *Sympathizing* is associated with the sense of *filled with excitement or emotion*, which is the most frequent sense of *charged* in the movie reviews. After obtaining the most similar words for a given word w, we define a value $domainSimilarity(w)$ such that $domainSimilarity(w) = Cosine_l(l(w), m(w))$. As we are considering the legal embedding model when getting the cosine similarity values, a higher *domainSimilarity(w)* value will suggest that legal sense and movie sense of the word w have a similar meaning in the legal domain while a lower *domainSimi-*

larity(w) will suggest that the meanings of the two senses are less similar to each other. For example, the value obtained for *domainSimilarity(Charged)* was 0.06 while it was 0.53 for *domainSimilarity(Convicted)* (*convicted* has a similar sense across the two domains).

The next step is to identify a threshold based on domainSimilarity(w) to heuristically distinguish whether a word w is domain generic or not. To that regard, we made use of already available Verb Similarity dataset [1] developed for the legal domain. The dataset consists of 959 verb pairs manually annotated based on whether the two verbs in a pair have a similar meaning or not. First, a threshold t based on cosine similarity was defined. For a given two verbs v_i, v_j, if $Cosine_l(v_i, v_j) \geq t$, the two verbs are considered as having a similar meaning. From the experiments, it was observed that precision is less than 0.5 when the threshold value is equal to 0.1. Therefore, 0.2 is selected as the threshold value to identify domain generic words based on the *domainSimilarity(w)* score. In other words, if *domainSimilarity(w)* is greater than or equal to 0.2, the word w will be considered as domain generic and the attribute *domainGeneric(w)* will be set to true. Otherwise, the attribute *domainSpecific(w)* will be set to true. Though we have used the aforementioned approach to determine the threshold, it is a heuristic and domain specific value that can be decided based on different experimental techniques (when applying this methodology to another domain).

Even if a word behaves in a similar manner across the two domain, it still can be assigned with a wrong sentiment (neutral sentiment) due to under representation. However, it is important to identify words with sentiment polarities (positive or negative) as the descriptions with positive or negative sentiments tend to contain more specific information that will be useful in legal analysis. As a measure of identifying sentiment polarities of under represented words, we made use of AFINN (Nielsen, 2011) sentiment lexicon (denoted as set A from this point onwards), which consists of 3352 words annotated based on their sentiment polarity (positive, neutral, negative) and sentiment strength considering the domain of twitter discussions. If a frequency of a word w is less than 3 in the source dataset, *underRepresented(w)* is set to true. Assignment of AFINN sentiment for an under represented word or a domain specific word w can create a positive impact if the most frequently used sense of w in twitter discusion domain is aligned towards it's sense in the legal domain than the sense of that word (w) in the movie review domain. In order to heuristically determine this factor, we have defined an attribute name *afinnSimilarity* such that $afinnSimilarity(w) = Cosine_t(w, l(w)) - Cosine_t(w, m(w))$, where w is a given word and $Cosine_t$ is the cosine similarity obtained using a publicly available Word2Vec model (Godin, 2019) trained using tweets. If $Cosine_t(w, l(w)) > Cosine_t(w, m(w))$, it can be assumed that the sense of word w in twitter discussions is more closer to its sense in the legal domain than that of the movie-reviews. Thus, if *afinnSimilarity(w)* > 0 and $w \in A$, the attribute *afinnAssignable(w)* is set to true.

Both Algorithm 1 and Algorithm 2 are two parts of one major algorithmic approach (Algorithm 1 executes first). Therefore, the functions and attributes defined in Algorithm 1 are applied globally for both Algorithm 1 and Algorithm 2 and the states of the attributes after executing Algorithm 1 will be transferred to the Algorithm 2. In the algorithms, P, N, O denotes positive, negative, and neutral sentiments respectively. *afinn(w)* is the AFINN sentiment categorization of a given word w. When observing the algorithm, it can be observed that sentiment of $l(w)$) is also considered when determining the correct sentiments of a word. For a word in O_m, the sentiment of $l(w)$ will be assigned if $l(w)$ is *domain generic* (Algorithm 1). This step was followed as another way to identify words with sentiment polarities (positive or negative). The sentiments of domain generic words in P_m or N_m will not be changed under any condition. For a domain specific word w in P_m or N_m, if $l(w)$ has a opposite sentiment polarity to that of w, the sentiment of $l(w)$ will be assigned to w only if $l(w)$ is domain generic. All the *domain specific* words in P_m or N_m that do not satisfy any of the conditions that are required to assign a positive or negative polarity (Algorithm 2), will be assigned with a neutral sentiment. This step is taken because such *domain specific* words have a relatively higher probability to have opposite sentiment polarities in

[1] https://osf.io/bce9f/

the legal domain, thus capable of transferring wrong information to the classification models (Sharma et al., 2018). Assigning neutral sentiment will reduce the impact of negative transfer that can be caused by such words (neutral sentiment is better than having the opposite sentiment polarity). Furthermore, it should be noted that an antonym of a particular word w can be given as $l(w)$ by the embedding model due to semantic drift. To tackle this challenge, WordNet (Fellbaum, 2012) was used to check whether a given word w and $l(w)$ are antonyms. If they are not antonyms, notAntonyms() attribute is set true. After running the Algorithm 1 and 2 by taking P_m, O_m, N_m as the inputs, the word sets D_{on}, D_{op} were obtained that consist of words the overall algorithm picked from O_m as having negative and positive sentiments respectively. D_{on}, D_{op} together with P_m, N_m were given to a legal expert in order to annotate the words in these sets based on their sentiments. $|D_{on}|$ = 220 and $|D_{op}|$=116, thus reducing the required amount of annotations to 925 (925= $|W|$, where $W = D_{op} \cup D_{on} \cup P_m \cup N_m$). After the annotation process, three word sets N_a, O_a, P_a were obtained that contains words that are annotated as having positive, neutral and negative sentiments respectively. Then word sets D_n, D_o, D_p were created such that $D_n = \{w \in W | w \in N_a \& w \notin N_m\}$, $D_p = \{w \in W | w \in P_a \& w \notin P_m\}$, $D_o = \{w \in W | w \in O_a \& w \notin O_m\}$. P_l contains the set of words identified by the overall algorithm as having positive sentiment and N_l contains the words identified as having negative sentiment (without human intervention).

3.2 Fine Tuning the RNTN Model

As an approach to develop a sentiment classifier for legal opinion texts, $RNTN_m$ (Stanford Sentiment Annotator) (Socher et al., 2013) was fine tuned following a similar methodology as proposed by (Gamage et al., 2018). In the proposed methodology (Gamage et al., 2018), there is no need to further train the $RNTN_m$ model or to modify the neural tensor layer of the model. Instead, the approach is purely based on replacing the word vectors. In this approach, if a word v in a word sequence S have a deviated sentiment s_d in the legal domain when compared with its sentiment s_m as output by the $RNTN_m$, the vector corresponding to v will be replaced by the vector of word u, where u is a word from a list of predefined words that has the sentiment s_d as output by $RNTN_m$. When choosing u from the list of predefined words, PoS tag of w in word sequence S is considered in order to preserve the syntactic properties of the language. For example, if we consider the phrase *Sam is charged for a crime*, as *charged* is a word that have a deviated sentiment, the vector corresponding to *charged* will be substituted by the vector of *hated* (*hated* is the word that matches the PoS of *charged* from the predefined word list corresponding to the negative class) (Gamage et al., 2018). When extending the approach proposed in (Gamage et al., 2018) for three class sentiment classification, a predefined word list for positive class was developed by mapping a set of selected words that have positive sentiment in $RNTN_m$ to each PoS tag. The mapping can be represented as a dictionary R, where R = {JJ:beautiful, JJR:better, JJS:best, NN:masterpiece, NNS:masterpieces, RB:beautifully, RBR:beautifully, RBS:beautifully, VB:reward, VBZ:appreciates, VBP:reward, VBD:won, VBN:won, VBG:pleasing}. For the negative class and the neutral class, the PoS-word mappings provided by (Gamage et al., 2018) for negative and non-negative classes were used respectively. Furthermore, instead of annotating each word in the selected vocabulary to identify words with deviated sentiments, we used word sets D_n, D_o, D_p that were derived using the approaches described in Section 3.1. In Section 4, the fine tuned RNTN model developed in this study is denoted as $RNTN_l$.

3.3 Adapting the BERT based approaches

An approach based on $BERT_{large}$ embeddings (Munikar et al., 2019) has achieved the state of the art results for sentiment classification of sentences in SST-5 dataset. In order to adapt the same approach for our task, following steps were followed. First, sentences with their sentiment labels were extracted from SST-5 training set. The SST-5 training set consists of 8544 sentences labelled for 5 class sentiment classification. As our focus is on 3 class classification, the sentiment labels in the SST training set were converted for 3 class sentiment classification by mapping very positive, positive labels as positive and very negative,negative labels as negative.

Next, following a similar methodology as described in (Munikar et al., 2019), *canonicalization*, *tokenization* and *special token addition* were performed as the preprocessing steps. Then, the classification model was designed following the same model architecture described in (Munikar et al., 2019), that consists of a dropout regularization and a softmax classification layer on top of the pretrained BERT layer. Similarly to (Munikar et al., 2019), $BERT_{large}$ uncased was used as the pretrained model and during the training phase, dropout of probability factor 0.1 was applied as a measure of preventing overfitting. Cross Entropy Loss was used as the cost funtion and stochastic gradient descent was used as the optimizer (batch size was 8). Then, the model was trained using the SST-5 training sentences. As information related to number of training epoch could not be found in (Munikar et al., 2019), we experimented with 2 and 3 epcohs and calculated the accuracies with a test set of 500 legal phrases (Section 4). When trained for 2 epochs, the accuracy was 57% and for 3 epcohs it was reduced to 52%, possibly due to the overfitting with the source data. Therefore, 2 was choosen as the number of training epochs. This model will be denoted as $BERT_m$ in next sections.

In order to finetune the BERT based approach to the legal sentiment classification, the following steps were followed. First we selected sentences in the SST training data, that consists of words that were identified as having deviated sentiments (words in $D_o \cup D_p \cup D_n$). If the sentiment label of the sentence S that has a deviated sentiment word w is different from the sentiment label assigned to w by the legal expert, then S will be removed from the original SST training dataset as a measure of reducing negative transfer. For example, if there is a sentence S with word *charged* and if the sentiment of S is positive or neutral (sentiment of charged is negative in legal domain), then that sentence S will be removed from the training set. After removing such sentences, the training set was reduced to 6318 instances and this new training set will be denoted by D from this point forward. Next, for each word w in D_n or D_p, we randomly selected 2 sentences that contains w from the legal opinion text corpus. Then, the sentiments of the selected sentences were manually annotated by a legal expert. As $|D_n| = 206$ and $|D_p| = 82$, only 576 new annotations were needed ($|D_o| = 230$, but words in D_o were not considered for this approach as they are having a neutral legal sentiment). Then, these 576 sentences from legal opinion texts were combined together with sentences in D, thus creating a new training set L that consists of 6894 instances. The above mentioned steps were followed to remove the negative transfer from the source dataset and also to fine tune the dataset to the legal domain. Then, L was used to train a BERT based model using the same architecture, hyper parameters and number of training epochs that were used to train $BERT_m$. The model obtained after this training process is denoted as $BERT_l$.

4 Experiments and Results

4.1 Identification of words with deviated sentiments

In order to evaluate the effectiveness of the proposed algorithmic approach when it comes to identifying legal sentiment of a word, we have compared the positive word list (P_l) and negative word list(N_l) identified by the algorithm with P_m and N_m respectively as shown in Table 1. The way in which P_l and N_l were obtained is described in Algorithm 2. It can be observed that the precision of identifying words with negative sentiments is 80% in the algorithmic approach and it is a 19% improvement when compared with the $RNTN_m$ (Socher et al., 2013). Furthermore, the number of correctly identified negative words have increase to 317 from 154. Though the precision of identifying words with positive sentiment is only 62%, there is an improvement of 21% when compared with the $RNTN_m$. Precision of identifying words with positive sentiment is relatively low due to the fact that most of the words that have a positive sentiment in generic language usage have a neutral sentiment in the legal domain. Sophisticated analysis in relation to the neutral class could not be performed due to the large amount of words available in O_m. When considering these results, it can be seen that the proposed algorithm has shown promising results when it comes to determining the legal domain specific sentiment of a word. Additionally, it implies that the proposed algorithmic approach is successful in identifying words that have different sentiments across the two domains. This approach can also be extended to other domains easily as domain specific word embedding models can be trained using an unlabelled corpus. Furthermore, the proposed algorithmic approach also has the potential to be used in automatic generation of domain specific sentiment lexicons.

Table 1: Evaluating the word lists generated from Algorithm 1 and Algorithm 2

Metric / Model	Number of Words				Percentages			
	N_m	N_l	P_m	P_l	N_m	N_l	P_m	P_l
Negative	**154**	**317**	17	20	**61%**	**80%**	5%	7%
Neutral	96	73	180	89	38%	19%	54%	41%
Positive	3	4	**139**	**181**	1%	1%	**41%**	**62%**
Total	253	394	336	290	100%	100%	100%	100%

Table 2: Precision(P), Recall (R) and F-Measure (F) obtained from the considered models

Metric / Model	Negative			Neutral			Positive			Accuracy
	P	R	F	P	R	F	P	R	F	
$RNTN_m$	0.51	0.68	0.58	0.44	0.52	0.48	0.48	0.10	0.16	0.48
$RNTN_l$ (Improved)	0.55	0.70	0.62	0.54	0.51	0.52	0.73	0.44	0.55	0.57
$BERT_m$	0.68	0.73	0.70	0.47	0.68	0.56	0.57	0.13	0.21	0.57
$BERT_l$ (Improved)	0.72	0.79	0.75	0.58	0.55	0.57	0.70	0.62	0.66	0.67

4.2 Sentiment Classification

In order to evaluate the performances of the considered models when it comes to legal sentiment classification, it is needed to prepare a test set that consists of sentences from legal opinion texts annotated according to their sentiment. As the first step of preparing the test set, 500 sentences were randomly picked from the legal opinion text corpus such that there is no overlap between the test set and the sentences used to train $BERT_l$. Then sentiment of each sentence was annotated by a legal expert. According to the human annotations, the number of data instances belong to negative, neutral and positive classes in the test set were 211, 168, and 121 respectively. The results obtained for each model for the test set is shown in Table 2. The effectiveness of the fine tuning approaches proposed in this study is evident as the RNTN finetuning has achieved accuracy increase of 9% while fine tuning the dataset for BERT training has achieved an accuracy increase of 10% when compared with the performances of the respective source models. It can be observed that the $BERT_m$ has the same accuracy as the $RNTN_l$. However, the performance of $RNTN_l$ model is relatively consistent across all 3 classes while the recall, f-measure of $BERT_m$ in relation to the positive class is significantly low. It should be noted that $BERT_l$ model that was trained after fine tuning the dataset for legal domain outperforms all other models. Furthermore, the state of the art accuracy value for 5 class sentiment classification of sentences in SST-5 dataset is 55.5%(Munikar et al., 2019). An accuracy of 67% for 3 class classification in the legal domain can be considered as satisfactory when we consider the added language complexities in legal opinion texts, though the number of classes has been reduced to 3. Most importantly, the accuracy enhancement of 10% compared with $BERT_m$ was achieved by including only 576 new sentences from legal opinion texts that were annotated by a legal expert. Therefore, it can be concluded that the transfer learning approach mentioned in Section 3.3 is an effective way to develop a domain specific sentiment annotator with a considerable accuracy while utilizing a minimum amount of annotations.

5 Conclusion

Developing a sentiment annotator to analyze the sentiments of legal opinions can be considered as the primary contribution of this study. In order to achieve this primary objective in a low resource setting, we have proposed effective approaches based on transfer learning while utilizing domain specific word representations to overcome negative transfer. As a part of the overall methodology, we have also proposed an algorithmic approach that has the capability of identifying the words with deviated sentiments across the source and target domains, while assigning the target domain specific sentiment to the considered words. The data sets prepared within this study for testing and training purposes has been made publicly available [2]. Moreover, the methodologies formulated in this study are designed in a way such that they can be easily adaptable for any other domain.

Acknowledgments

This research was funded by SRC/LT/2018/08 grant of University of Moratuwa.

[2] https://osf.io/zwhm8/

References

[Baccianella et al.2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

[Bradley and Lang1999] Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology.

[Conrad and Schilder2007] Jack G Conrad and Frank Schilder. 2007. Opinion mining in legal blogs. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 231–236.

[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*.

[Fellbaum2012] Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

[Gamage et al.2018] Viraj Gamage, Menuka Warushavithana, Nisansa de Silva, Amal Shehan Perera, Gathika Ratnayaka, and Thejan Rupasinghe. 2018. Fast approach to build an automatic sentiment annotator for legal domain using transfer learning. *arXiv preprint arXiv:1810.01912*.

[Godin2019] Fréderic Godin. 2019. *Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing*. Ph.D. thesis, Ghent University, Belgium.

[Liu and Chen2018] Yi-Hung Liu and Yen-Liang Chen. 2018. A two-phase sentiment analysis approach for judgement prediction. *Journal of Information Science*, 44(5):594–607.

[Maas et al.2011] Andrew 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.

[Mikolov et al.2013] 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.

[Munikar et al.2019] 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.

[Nielsen2011] Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[Ratnayaka et al.2019] Gathika Ratnayaka, Thejan Rupasinghe, Nisansa de Silva, Viraj Salaka Gamage, Menuka Warushavithana, and Amal Shehan Perera. 2019. Shift-of-perspective identification within legal cases. *arXiv preprint arXiv:1906.02430*.

[Sharma et al.2018] Raksha Sharma, Pushpak Bhattacharyya, Sandipan Dandapat, and Himanshu Sharad Bhatt. 2018. Identifying transferable information across domains for cross-domain sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 968–978.

[Socher et al.2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y 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.

[Sugathadasa et al.2017] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 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.

[Thelwall et al.2010] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558.

[Van Rijsbergen1979] C Van Rijsbergen. 1979. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, pages 1–14.

Deriving confirmation and justification — an expectative, compositional analysis of Japanese *yo-ne*

Lukas Rieser
Tokyo University of Agriculture and Technology / Tokyo, Japan
rieserl@go.tuat.ac.jp

Abstract

The Japanese epistemic particles *yo* and *ne* have received much attention in formal pragmatics, but it remains contentious whether their combination *yo-ne* can be analyzed compositionally or is an independent lexical item. The novel, expectation-based approach I propose captures the contributions of both particles in declaratives and interrogatives on discourse-oriented as well as soliloquous uses, compositionally accounts for *yo-ne* on its well-documented confirmation as well as the emerging justification use, and predicts its badenss in polar, but not *wh*-interrogatives.

1 Intro: is *yo*(-)*ne* compositional?

While there is plenty of research on *yo* and *ne* in isolation, their combination has received somewhat less attention and there is no consensus on whether it is best analyzed as an independent expression *yone* or as compositionally derived *yo-ne*.

1.1 Issues with (un)controversiality approaches

Intuitive paraphrases typically characterize *yo* as (strongly) assertive, intending to force addressee acceptance of the prejacent; *ne* as confirming, signaling that prejacent acceptance by the addressee is likely. Implementing such intuitions in formal analysis, henceforth the "(un)controversiality approach", tends to predict contradictory meanings and thus complementary distribution of *yo* and *ne*, precluding a compositional analysis of *yo-ne*.[1]

An issue with the (un)controversiality approach unrelated to compositionality is the particles' contribution in falling interrogatives (FIs), which function as exclamations or expressions of doubt felicitous in soliloquy, *i.e.* not necessarily requiring an addressee, and where the particles have markedly different contributions than in declaratives — *yo* in FIs introduces a mirative nuance rather than forcing the prejacent on the addressee; *ne* strengthens the conveyed notion of doubt rather than signaling mutual acceptance.

1.2 Extant analyses on compositionality

Oshima (2014) takes *yone* to be a lexical item which covers different types of confirmation together with *ne*[2], but cannot be derived from *yo* and *ne*. Najima (2014) additionally connects *yone* to *yo*, but rejects a compositional analysis in favor of describing shared features, concluding *yone* is a less determinate form of *yo* and *ne*, sharing some features with each, as well as with the interrogative marker *ka*.

Takubo and Kinsui (1997), on the other hand, propose that *yo* and *ne* are applied sequentially, thus providing a compositional account of *yo-ne* and predicting the infelicity of a sequence **ne-yo*. Their analysis is built on a mental-space rather than a belief-oriented framework, thus also being applicable to soliloquous uses, *cf.* Hasegawa (2010a). My analysis of *yo* and *ne* also applies to soliloquy as it does not involve obligatory reference to addressee belief, and has the additional advantage of compositionally accounting for combination with *ka*.

[1] *Cf.* McCready (2009) for discussion of issues around extant analyses of *yo* and *ne* (excluding their use in interrogatives).

[2] See also Miyazaki (2002) for a taxonomy of various uses of *yo(-)ne* and *ne* used in confirming utterances.

1.3 Uses unaccounted for

Other than the occurrence of *yo* and *ne* (and the badness of *yo-ne*) in polar FIs, which remains outside the scope of some analyses, the use of *yo-ne* in *wh*-interrogatives has to my knowledge not been discussed in the application of previous analyses.

As for *yo-ne* in declaratives, my analysis not only covers the well-documented "confirmation" use accounted, but also an emerging use of *yo-ne* that I label "justification", which, as far as I am aware of, extant analyses have not taken into account. On this use, *yo-ne* is in complementary distrubution with *yo*, rather than *ne*, pointing towards functional overlap between *yo-ne* and *yo* as well as *ne*.

2 The expectative approach

I propose an analysis on which *yo* and *ne* encode the status of the prejacent as an expectation and connect it to premises in the conversational background. This accounts for both confirmation and justification and for the (in)felicity of *yo-ne* in interrogatives.

2.1 The expectative context

On the expectative view of discourse, the context, defined as the set of all propositions other than the prejacent relevant for utterance interpretation, is split into **premises** and **expectations**, differentiated by whether or not a proposition is epistemically settled. Epistemic particles like *yo* and *ne* mark the prejacent as a premise or expectation and inform about its relation to other members of the context set.

2.2 Contextual premises

Shared premises include propositions agreed upon in the discourse (conversational common ground[3]), what is conventionally considered to be a premise (world knowledge), along with external anchors like extralinguistic evidence and antecedent utterances.

Participant-specific premises sets, which are crucial to capture differing premises and for analyzing soliloquy, additionally include private beliefs. I write Π^x for the set of premises specific to participant x as in (1), where $B_x\pi$ indicates that π is epistemically settled to x (*i.e.* x believes π to be true).

[3] *cf.* Stalnaker (2002), among others — this is only part of the premise set, which also includes speaker-specific commitments as well as external anchors such as contextual evidence.

$$\Pi^x = \{\pi \mid B_x(\pi)\} \tag{1}$$

Within Π, epistemic and evidential premises can be distinguished. For instance, when (extralinguistic, but also hearsay) evidence constitutes a premise to an agent, only the existence of such evidence, but not necessarily the proposition it supports is epistemically settled. While I differentiate different uses of *yo* and *ne* by this distinction, I do not formally implement it as they are not sensitive to it.

2.3 Contextual expectations

Expectations can be thought of as what is normally the case, but does not necessarily hold in all cases. Assuming this is equivalent to the so-called weak epistemic nessecity reading of English *ought*, I write $\textsc{ought}(\xi)$ for "ξ normally holds", where OUGHT represents a normalcy or anticipative modal[4], and define the set Ξ^x of $x's$ expectations ξ as in (2).

$$\Xi^x = \{\xi \mid B_x\textsc{ought}(\xi)\} \tag{2}$$

This is to say that ξ is an expectation of x if $\textsc{ought}\xi$ is a premise epistemically settled to x.

Expectations negotiated by epistemic particles are typically based on premises, a relation I model as restriction of OUGHT's modal base with a premise, *i.e.* as a normalcy conditional. I label this the "expectative relation", written as $\rightsquigarrow$ in (3), where the participant-specific set of premise-based expectations is defined.

$$\Xi^x_\Pi = \{\xi \mid \exists\pi \in \Pi^x : \pi \rightsquigarrow \xi\} \tag{3}$$

It should be noted that this relation is defeasible, so that expectations can be in contrast with epistemically settled premises. If, for instance, in an expectative atypical w.r.t. p where $\textsc{ought}(p)$ is a premise, p remains expected even if $\neg p$ is settled.

2.4 Negotiating expectations

With the expectative context in place, the crucial question is how its contents are determined and how they change during the discourse. I assume the default goal of a discourse is to maximize the set of

[4] Against analyzing OUGHT as a weaker epistemic modality, von Fintel and Iatridou (2008) and Yalcin (2016) propose normality analyses; in Rieser (2020) I label this "anticipative" modality, proposing Japanese *hazu* is of this flavor.

premises, *i.e.* to settle as many propositions as possible, while resolving epistemic inconsistencies by context update. Epistemic particles function to negotiate this process and navigate issues like beliefs differing between participants, or the emergence of evidence that counters or confirms extant beliefs.

In descriptive metalanguage, I use the phrase "p is (not) expected (to x)" to indicate that p is (not) part of the set of x's expectations at a given point in the discourse, and "p is (not) settled (for x)" to indicate that p is (not) part of x's premise set.

3 Particle meanings

I propose that both *yo* and *ne* mark the prejacent's status as an expectation and impose conditions felicitous utterance contexts, henceforth (context) presuppositions, while only *yo* updates the context, making the prejacent expected. Presuppositions and updates from *yo* and *ne*, relative to prejacent p, are shown in the table below.

	presupposition	update
ne(p)	$p \in \Xi^x_\Pi$	–
yo(p)	$p \notin \Pi^x$	$p \cup \Xi^x_\Pi$

Note that the presupposition of *ne* is the same as the result of update with *yo*, crucial for deriving the meaning of *yo-ne*. Whereas *ne* requires p to be expected to an underspecified participant x, *yo* requires it not to be, and seeks to make it expected.

The various uses of the particles are differentiated by which participant x is resolved to and the of premise p is (made) expected from. Differences between interrogatives and declaratives are derived by combining each utterance type's original felicity conditions with those introduced by the particles.

3.1 In declaratives

Recall that *ne* in declaratives asserting a prejacent is often described as confirming, that of *yo* as strongly assertive. Without going into detail on their various declarative uses, the following example of a prejacent clearly not expected by the addressee illustrates the difference between *yo* and *ne* reflected in the proposed analysis. Consider a situation where the speaker is observing the addressee walking dangerously close to a fall but being oblivious of the danger and utters (4) to warn them.

(4) Abunai {yo / #ne}!
dangerous SFP SFP
"Careful!"

It should be noted that *yo* is preferred in (4) over no particle at all. This is expected, considering the presupposition $p \notin \Pi^A$ with x resolved to A is a prefect fit for prompting the addressee to add extralinguistic evidence as a premise π from which p is expected, resulting in $p \in \Xi^\Pi_A$. Figure 1 illustrates how this leads to revision of a possible extant expectation $\neg p$ and prepares addressee settlement of p.

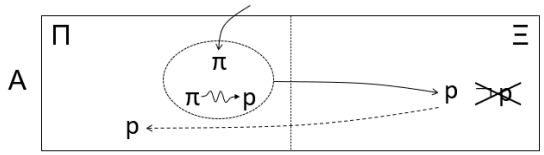


Figure 1: addressee update from *yo*-declarative

The premise π anchoring prejacent p as an expectation can be extralinguistic evidence as in the scenario sketched for (4) above, but also the speaker's assertion of the prejacent as such, constituting hearsay evidence[5], as in example (5) below.

Discourse-oriented *ne*, on the other hand, presupposes p to be addressee-expected ($p \in \Xi^A_\Pi$), *i.e.* the speaker anticipates that the prejacent is either believed by the addressee, or already expected so that it can be readily settled, as Figure 2 illustrates.

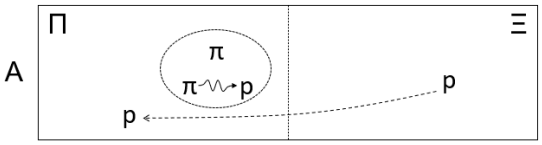


Figure 2: addressee update from *ne*-declarative

The acceptability of *yo* and *ne* in (4) can be reversed by manipulating the scenario. Assuming, for instance, the participants are watching a movie where a character is precipitously close to a fall, (4) with *ne* would be felicitous, with *yo* clearly degraded.

Without an external evidential anchor as in (4), *yo* and *ne* bring out rather different readings, as in (5).

(5) Ii {yo / ne}!
good SFP SFP
{"Sure!"/"Nice!"}

[5]Gunlogson (2008) implements a similar idea of participant commitments as source-specific premises within the context.

The evaluative predicate *ii* 'good' has a wide range of uses in Japanese. As indicated in the paraphrases, (5) with *yo* can *e.g.* be used to give permission to the addressee ("Sure!"), which is not possible with *ne*. With *ne*, on the other hand, (5) can be used to comment on the desirability of some state of affairs ("Nice!"), for which *yo* is unsuitable.

In the former case ("Sure!") it is the speaker's (performative) assertion of positive evaluative *ii* rather than external evidence that serves as a premise to make the prejacent addressee-expected. In the latter case ("Nice!") the prejacent is already presumed addressee-expected based on their knowledge of the evaluated state of affairs.

3.2 Discourse orientation vs. soliloquy

In addition to such addressee-oriented readings, *ne*, but not *yo*, in declaratives has a productive "soliloquous" reading in the sense of the agent variables being resolved to the speaker, rather than the addressee.[6] In contrast to this, *yo* in interrogative is by default soliloquous, while *ne* has both a discourse-oriented and a soliloquous use, *cf.* section 3.3.

Back to declaratives, Hasegawa (2010a) points out the following example from Takubo and Kinsui (1997), where *ne* implies "computation or confirmation on the part of the speaker", as an instance of *ne* in soliloquy. The utterance situation is one in which the speaker is checking the time on their watch, the utterance being made while reading the dial.

(6) Eeto, shichi-ji desu ne.
well 7-o'clock COP SFP
"It's seven o'clock."

While in the example, the speaker is also conveying the time to an addressee, the utterance is soliloquous in the sense of narrating the speaker's internal belief-formation process based on an external premise, which I analyze as resolving *ne*'s participant variable to the speaker, rather than the addressee. I will use the terms "discourse-oriented" and "soliloquous" in this sense.[7]

This function of *ne* will be crucial for explaining the variation in uses of compositionally derived *yo-ne*. Rather than "computation or confirmation" in particular, I take soliloquous *ne* to more generally anchor assertion of the prejacent in some extant premise, such as the external evidence provided by the watch in this example. Figure 3 illustrates the variable in the presupposition of *ne* resolved to the speaker, yielding $\varphi \in \Xi^S_\Pi$.

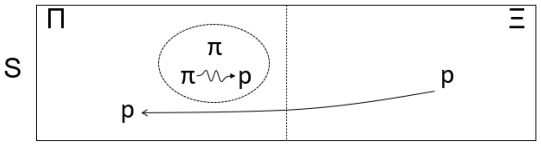


Figure 3: soliloquous *ne* in declaratives

The discourse-orientation of *yo* is even stronger than that of *ne*. When (7), a COMP-exclamative, is uttered in soliloquy, adding *yo* is infelicitous.

(7) A, soko-ni atta n da (#yo)!
oh here-LOC was COMP COP
"Oh, there it was!"

Mirativity, *i.e.* marking of the prejacent as unexpected, can also be conveyed by *yo* in FIs, as discussed below, but does not license *yo* in exclamatives like (7), or other declaratives. When the utterance is not purely soliloquous, *yo* can be added under the assumption that the addressee was also wondering where the item in question had been, thus making the prejacent unexpected. in sum, only when there is (also) an addressee expectation to be revised, *yo* can be added to declaratives.

3.3 In interrogatives

The perspective from final falling interrogatives (FIs) is crucial to fully capture the contribution of *yo* and *ne*, and, building on this, the compositional derivation of *yo-ne*. FIs, a particularly productive class of utterances in Japanese, are often characterized as rhetorical questions expressing their prejacent is not epistemically settled and can be used in soliloquy. Considering this, the contribution of *yo* and *ne* in FIs should be observable within the speaker context.

Yo in FIs indicates that the prejacent has become expected based on evidence that has just become available in the utterance situation, whereas FIs with *ne* convey the speaker's sustained doubt over whether the prejacent holds in spite of there being grounds to expect this. Consider (8) in a context

[6]The soliloquous use of *ii-ne* is, incidentially, the Japanese translation of the "like" button on Facebook.

[7]For more detailed discussion on the definition of soliloquy, see Hasegawa (2010b)

where the speaker notices some people out of a party appear to be gathering their belongings.

(8) Kaeru ka {yo! / ne...}
good SFP SFP
{"What, they're leaving!?"/
"Are they really leaving..."}

The salient reading of (8) with *yo* is translated as an exclamation, that with *ne* as an expression of doubt. Note that this is in stark contrast with the contribution of the particles in declaratives.

The schemata in Figure 4 illustrate how the present proposal accounts for *yo* and *ne* in FIs. Both utterances narrate (potential) change in the speaker context, the interrogative speech act coming with the presupposition $p \notin \Pi^S$.

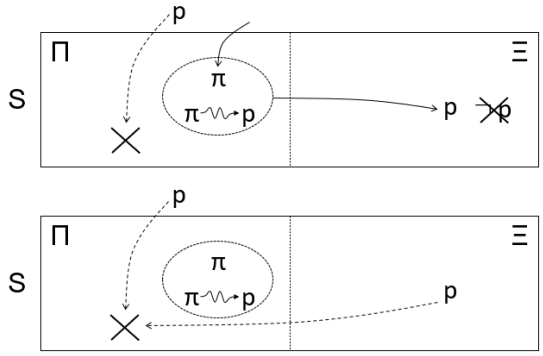


Figure 4: update of speaker context set as narrated by *ka-yo* (top) and *ka-ne* (bottom)

Yo-FIs mark possible revision of a speaker expectation $\neg p$ ($\neg p \in \Xi^S$) is revised based on the premise of newly available evidence for p ($\exists \pi \in \Pi^S : \pi \rightarrow p$). While this does not necessarily coincide with epistemic settlement p ($p \cup \Pi^S$), this is on the table and potentially imminent, given sufficient strength of the newly available evidence.[8]

Purely soliloquous *ne*-FIs indicate that even though p is, or has become, expected to the speaker ($p \in \Xi^S$), they forgo settling for it. The reasons for this can be varied, for instance a previous settlement for $\neg p$ ($\neg p \in \Pi^S$), or doubt over the sufficient strength of π as a premise.

Additionally, *ne*-FIs have a discourse-oriented use indicating p is not settled to the speaker as required by the bare FI ($p \notin \Pi^S$), but the expect the addressee to have grounds for expecting it as conveyed by *ne* ($\exists \pi \in \Pi^A : \pi \rightarrow p$). The effect is a bias towards a positive answer, or a confirmation use with stronger speaker doubt than a *ne*-assertion.

Interim summary The current proposal not only accounts for *yo* and *ne* in declaratives but also in interrogatives, where they make contributions that cannot be predicted based on the extant generalizations of *yo* being "strongly assertive", *ne* "confirming" in nature. This is achieved by assuming particles meaning that do not involve obligatory addressee-reference, in contrast to (un)controversiality approaches, and deriving utterance meaning by modification of each utterance type's basic felicity conditions.

4 Deriving the meaning of *yo-ne*

I follow Takubo and Kinsui (1997) in proposing that *yo-ne* can be derived as sequential application of the two particles — *yo* makes the prejacent expected, which *ne* presupposes. In order to account for the full range of uses, my analysis allows distinct premises and/or participant resolution for each particle. This is crucial to account for a subset of confirming uses of *yo-ne* as well as the emergent justification use, which is excluded from the scope of analyses that take *yone* to essentially be a confirmation marker.

4.1 *Yo-ne* in confirmations

Yo(-)*ne* is frequently described as a confirmation marker similar in function to *ne*, *cf.* Miyazaki (2002). Before moving on its emerging justification use which relates it to *yo* rather than *ne*, this section accounts for the distriubtion of *yo-ne* and *ne* in confirmations with reference to Oshima (2014)'s classification of *yone*-utterances summarized below.

	ne	*yone*
a. Confirmation of A's utterance or checking for A's understanding	✓	#
b. Elsewhere (not a.): Prejacent is preparatory condition for S's subsequent utterance, or S is "questioner"	✓	✓
c. Elsewhere (neither a. nor b.):	#	✓

[8] When the evidence is strong enough for the speaker to tend to settle p, this is conventionally marked with the evidential particle *no*, *cf.* Rieser (2017b),Taniguchi (2016).

Consider first (9) in a context where the speaker recalls wondering about the prejacent, and the addressee's preparing to leave reminds them of it.

(9) Ashita asa hayai yo ne.
tomorrow morning early SFP SFP
"Early start tomorrow, right?"

(9) is an instance of Oshima's category c, favoring the use of *yo-ne*, and comes in two varieties Oshima labels "call for confirmation" (seeking addressee's confirmation), and "shared information" (marking p as mutually expected or settled).[9] I suggest that on both uses, *yo* is discourse oriented, and, in absence of extralinguistic evidence, the speaker's assertion of p is added as π to the addressee premise set.

Call for confirmation On the reading illustrated in Figure 5, *ne* is soliloquous[10], indicating that the speaker has grounds to expect p ($p \in \Xi^S$).

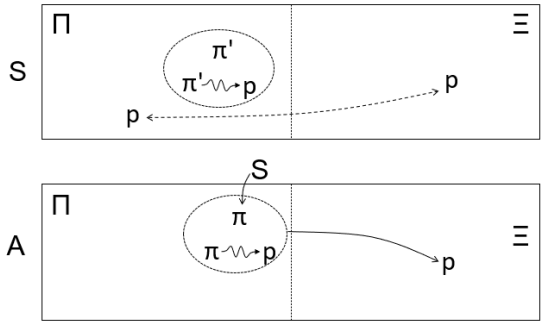


Figure 5: call for confirmation use of *yo-ne*

The degree of speaker certainty w.r.t. p as a shared expectation is relatively low, *ne* indicating the speaker has access to a premise $\pi' : \pi' \rightsquigarrow p$, and promts the addressee to add the speaker's assertion of p to their evidential premise set as π.

I propose the lack of reference to extant addressee premises prompts them to disclose their epistemic state w.r.t. p. While this interpretation may not be immediately obvious from the particle meanings alone, the marked intonation pattern of the call for confirmation use (also observed by Oshima) makes it plausible that this interpretation arises in contrast with the shared information use.

[9]The two varieties are differentiated by intonation, a compositional analysis of which is beyond the scope of this paper.

[10]This is not to say the entire utterance is not discourse-oriented.

Shared information On the reading Figure 6 illustrates, both *yo* and *ne* are discourse-oriented, indicating the speaker assumes there is another addressee- or shared premise π' also supporting p.

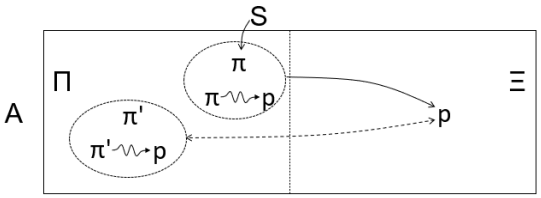


Figure 6: shared information use of *yo-ne*

In contrast to *ne*-declaratives of Oshima's category a, the speaker is not necessarily anticipating p to be addressee-settled, but can also merely confirm sufficient (shared) grounds to expect p.

Ne* vs. *yo-ne In sum, adding *yo* to *ne* on confirming use is dispreferred whenever it is implausible that the premise is not already settled for the addressee as required by *yo* ($p \notin \Pi^A$). $p \in \Pi^A$ holds in Oshima's class a, where the addressee has either asserted p or is committed to accept what was asserted by the speaker, merely checking for understanding, making *yo* illicit.

Finally, Oshima's category b is characterized by turn-holding on part of the speaker, *i.e.* subsequent acceptance of p by the addressee is anticipated. I propose that in this (addressee-oriented) case, addition of *yo* indicates that the speaker deems is not sure whether the addressee does not have (sufficient) grounds to accept p, but maintains that their own assertion of it constitutes such a premise.

4.2 *Yo-ne* in justifications

The use of *yo-ne* in what I label "justifications" is observable in younger speakers and intuitively serves to soften rejections. They can thus be considered a cases of *ne* being added to a *yo*-assertion. Consider the following variation of the confirmation example in a situation where the speaker is indirectly refusing an offer to join a second round of drinks.

Reacting to: "Are you joining us for another round?"
(10) Ashita asa hayai n da yo ne.
tomorrow morning early COMP COP PRT PRT
"I kinda have an early start tomorrow..."

The discourse-connective COMP-COP construction is used to indicate a causal relation with the (im-

plicit) refusal (a strategy available as "It's that…" in English). In contrast to confirmations, the addressee is previously unaware of p, and a *yo*-assertion is an alternative. Some speakers report this would feel harsh, so they choose adding *ne* to soften the rejection's blow. Note that a *ne*-assertion would be infelicitous assuming the addressee does not expect p to hold (hence inviting the speaker to drink more).

I propose analyzing this use of *yo-ne* as purely discourse-oriented, where *yo* prompts expectation revision, and *ne* subsequently ensure this revision has been successful by marking p as addressee-expected. This is schematically shown in Figure 7, where the previous utterance's prejacent (roughly "S joins for more drinks.") is written as ξ.

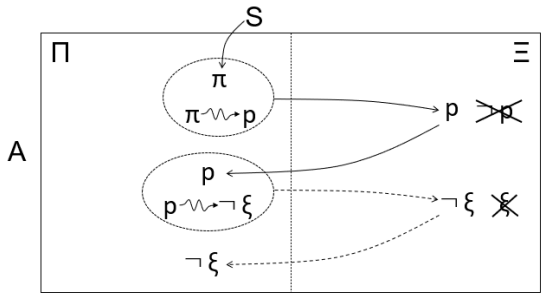


Figure 7: justifiction use of *yo-ne*

The utterance connects to an additional expectation ξ (the speaker joining a second round), as indicated by the COMP-COP construction. This expectation is indirectly negated by assertion of p, which the addressee is prompted to accept, *i.e.* settle, based on the speaker's utterance. This, in turn, gives rise to an expectation $\neg\xi$, thereby rejecting the proposal in a roundabout way.

In this example, *ne* functions to ensure that the justifying grounds for rejection have been accepted. I propose that *yo* in confirmations of Oshima's class b has a parallel, seemingly superfluous use, where it functions to ensure that a shared premise indicated by *ne* is actually present. I assume that face-saving strategies like adding *yo* to turn-holding utterances or *ne* to justifications also occur elsewhere.

In sum, the current proposal's flexibility in terms of speech-act type and discourse-orientation allows for derivation of *yo-ne* and accounting for its various uses in declaratives. Next, it can also account for its (in)felicity in (*wh*)-interrogatives.

4.3 *Yo-ne* in (*wh*)-interrogatives

While polar *yo-ne* interrogatives are degraded to the point of being labeled ungrammatical (Najima, 2014), *yo-ne* is perfectly fine in a subclass of *wh*-interrogatives which have so far not received due attention in the literature. In (11)[11], the speaker is certain that the object in question points to a treasure, but is uncertain what location it indicates.[12]

(11) Doko-o simeshiteiru no ka, yo ne.
where-ACC indicate COMP INT PRT PRT
lit.:"Where is it it really indicates…"

Figure 8 illustrates my account of *yo-ne*'s acceptability in some *wh*-interrogatives. The prejacent of a *wh*-interrogative is a proposition of the form $p(x)$, x being the *wh*-expression's referent. While an existential statement $\exists x : p(x)$ is settled, settling $p(x)$ is not possible due to ignorance w.r.t. x. On this background, discourse-oriented *yo* in *wh*-interrogatives prompts addition of a premise supporting $p(x)$ to the addressee context (as confirmed by *ne*) rather than conveying mirativity as in polar FIs.

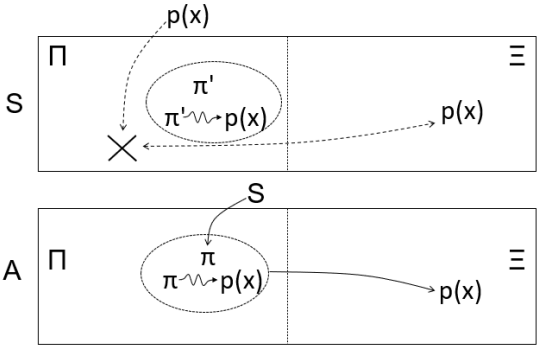


Figure 8: *yo-ne* in *wh*-interrogatives

As for the combination of *yo* and *ne* in polar FIs, I assume it is unacceptable due to complementary conveyed meanings — in light of newly available evidence, *yo*-FIs conventionally indicate belief revision, *ne*-FIs suspension of judgment. *Yo-ne* would thus represent considering, but then suspending belief revision — a process to complex for soliloquy narrating the speaker's immediate reaction.

[11] Found in the 1989 novel *Koto Pazuru* by Alice Arisugawa.

[12] As indicated by the comma, there is an obligatory pause between *ka* and *yo-ne*, presumably due to discourse-oriented *yo* being highly uncharacteristic in interrogatives.

5 Formalization

This section sketches formalization of the proposal within a CCP-model of utterance meaning,[13] taking utterances to be sets of admissible context pairs.

Premise and expectation sets (repeated)

(12) $\Pi^x = \{\pi \mid B_x(\pi)\}$

(13) $\Xi^x = \{\xi \mid B_x\text{OUGHT}(\xi)\}$

(14) $\Xi^x_\Pi = \{\xi \mid \exists\pi \in \Pi^x : \pi \rightsquigarrow \xi\}$

Expectative context set

(15) $c_x = \Pi^x \cup \Xi^x$

Utterances as CCPs, modification by particles

(16) $[\![U(p)]\!] = \{\langle c, c'\rangle \mid F^U\}$

(17) $[\![\text{PRT}[U(p)]]\!] = \{\langle c, c'\rangle \mid F^U \cup F^{\text{PRT}}\}$

Where c and c' represent input and output contexts (including all participants), respectively, F^U the characteristic felicity conditions for utterance type U, which may include agent- and partition specific restrictions on c as well as c'.

Felicity conditions of INT and DEC[14]

(18) $F^{\text{INT}}(p) = \{p \notin \Pi^{c_S}\}$

(19) $F^{\text{DEC}}(p) = \{p \in \Pi^{c_S}\}$

Particle meanings of *yo* and *ne*

(20) $F^{yo} = \{p \notin \Pi^{c_x}, p \in \Xi^{c'_x}_\Pi\}$

(21) $F^{ne} = \{p \in \Xi^{c_x}_\Pi\}$

***Yo* and *ne* in assertions**

(22) $[\![yo[\text{DEC}(p)]]\!] = \{\langle c, c'\rangle \mid p \in \Pi^{c_S} \wedge p \notin \Pi^{c_x} \wedge p \in \Xi^{c'_x}_\Pi\}$

(23) $[\![ne[\text{DEC}(p)]]\!] = \{\langle c, c'\rangle \mid p \in \Pi^{c_S} \wedge p \in \Xi^{c_x}_\Pi\}$

***Yo* and *ne* in interrogatives**

(24) $[\![yo[\text{INT}(p)]]\!] = \{\langle c, c'\rangle \mid p \notin \Pi^{c_S} \wedge p \notin \Pi^{c_x} \wedge p \in \Xi^{c'_x}_\Pi\}$

(25) $[\![ne[\text{INT}(p)]]\!] = \{\langle c, c'\rangle \mid p \notin \Pi^{c_S} \wedge p \in \Xi^{c_x}_\Pi\}$

(Note that the discourse-orientation and soliloquous nature of *yo* in declaratives and interrogatives respectively follows from incompatibility of a speaker-oriented presupposition $p \notin \Pi^c_S$ with the declarative condition and congruence of the same presupposition with the interrogative condition. *Ne* is more flexible in terms of orientation.)

Composing *yo-ne* assertions (sequential update)

(26) $[\![ne[yo[\text{DEC}(p)]]]\!] = \{\langle c, c'\rangle \mid p \in \Pi^{c_S} \wedge p \notin \Pi^{c_x} \wedge p \in \Xi^{c'_x}_\Pi \wedge p \in \Xi^{c'_y}_\Pi\}$

Note that *ne* constrains the context updated by *yo*, which in case of identity $x = y$ satisfies *ne*'s presupposition, but targets another partition in case of split participant resolution.

Predicting (il)licit uses The present proposal predicts limits of particle felicity *e.g.* as follows:

- *Yo* is illicit where: $\neg\exists x : p \notin \Pi^{c_x}$
- *Ne* is illicit where: $\neg\exists x : p \in \Xi^{c_x}_\Pi$

With discourse-orientation (x is resolved to A):

- *Yo* is illicit where: $p \in \Pi^{c_A}$
- *Ne* is illicit where: $p \notin \Xi^{c_A}_\Pi$
- *Yo-ne* is illicit where: $\neg\exists\pi : \pi \rightsquigarrow p \wedge \pi \notin \Pi^{c_A}$

This predicts that *yo-ne* is quite widely acceptable (while other particles may be chosen in its stead as they are more informative w.r.t. the prejacent's epistemic status), except for cases where the speaker has no new grounds for settling the prejacent to provide to the addressee. This corresponds to Oshima's condition a, where the addressee has already or intend to commit to p, *yo-ne* thus being dispreferred.

6 Summary

I have proposed an analysis of *yo* and *ne* that is flexible enough to capture their contributions in both interrogatives and declaratives, on both discourse-oriented and soliloquous uses, but specific enough to predict the limits of their acceptability, as well as their combination *yo-ne*, in various speech-act types. The lack of obligatory reference to shared or addressee belief makes a compositional analysis of *yo-ne* possible, where the possibility of split participant resolution for each particle, or particle and host utterance, as well as flexible premise resolution, allow to account for a wide range of uses.

[13] *Cf.* Rieser (2017b), Rieser (2017a), Davis (2011) for CCP-analyses of Japanese particles, building on Gunlogson (2003) and Heim (1983), among others.

[14] More precisely, the DEC(p) speaker should have grounds to expect p and not have settled for $\neg p$ as a premise, reflecting Gricean Quality, *cf.* Grice (1975). For the purposes of illustrating the role of particles, the simplified condition suffices.

References

Christopher Davis. 2011. *Constraining Interpretation: Sentence Final Particles in Japanese*. Ph.D. thesis, University of Massachusets - Amherst.

Herbert P Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York.

Christine Gunlogson. 2003. *True to form: Rising and falling declaratives as questions in English*. Ph.D. thesis, UCSC.

Christine Gunlogson. 2008. A question of commitment. *Belgian Journal of Linguistics*, 22(1):101–136.

Yoko Hasegawa. 2010a. The sentence-final particles *ne* and *yo* in soliloquial Japanese. *Pragmatics*, 20(1):71–89.

Yoko Hasegawa. 2010b. *Soliloquy in Japanese and English*. John Benjamins Publishing Company.

Irene Heim. 1983. On the projection problem for presuppositions. *Proceedings of the West Coast Conference of Formal Linguistics (WCCFL)*, 2:249–260.

E. McCready. 2009. Particles: Dynamics vs. utility. *Japanese/Korean Linguistics*, 16(6).

Kazuhito Miyazaki. 2002. Kakunin'yookyuu [= confirmation requests]. In Kazuhito Miyazaki, Taro Adachi, Harumi Noda, and Shino Takanashi, editors, *Modariti [= Modality]*, pages 203–228. Kuroshio Shuppan, Tokyo.

Yoshinao Najima. 2014. 'yone' to 'yo'.'ne' — genshikei-ka fukugoukei-ka [= 'yone' and 'yo.ne' — primitive or composite form]. Ms. Presented at 8th Symposium of Japanese Education, Nagoya University.

David Yoshikazu Oshima. 2014. On the functional differences between the discourse particles ne and yone in Japanese. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 442–451.

Lukas Rieser. 2017a. *Belief States and Evidence in Speech Acts: The Japanese Sentence Final Particle* no. Ph.D. thesis, Kyoto University.

Lukas Rieser. 2017b. Doubt, incredulity, and particles in Japanese falling interrogatives. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 25–33.

Lukas Rieser. 2020. Anticipative and ethic modalities: Japanese *hazu* and *beki*. *JSAI-isAI 2019 Workshops: LNCS*, 12331:309–324.

Robert C. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25:701–721.

Yukinori Takubo and Satoshi Kinsui. 1997. Discourse management in terms of mental spaces. *Journal of Pragmatics*, 28(6):741–758.

Ai Taniguchi. 2016. Setnece-final *ka-yo* in Japanese: A compositional account. *Proceedings of FAJL 8: Formal Approaches to Japanese Linguistics*, pages 165–176. (MITWPL 79).

Kai von Fintel and Sabine Iatridou. 2008. How to say *ought* in foreign: The composition of weak necessity modals. In *Time and modality*, pages 115–141. Springer.

Seth Yalcin. 2016. Modalities of normality. In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*, pages 230–255. Oxford University Press.

Combining Thai EDUs: Principle and Implementation

Chanatip Saetia **Supawat Taerungruang** **Tawunrat Chalothorn**
Kasikorn Labs
Kasikorn Business Technology Group (KBTG)
Nontaburi, Thailand
chanatip.sae@kbtg.tech, supawat.t@kbtg.tech, tawunrat.c@kbtg.tech

Abstract

Due to the lack of explicit end-of-sentence marker in Thai, Elementary Discourse Units (EDUs) are usually preferred over sentences as the basic linguistic units for processing Thai language. However, some segmented EDUs lack of structural or semantic information. To obtain a well-form unit, which represents a complete idea, this paper proposes combining EDUs with rhetorical relations selected depending on our proposed syntactic and semantic criteria. The combined EDUs can be then used without considering other parts of the text. Moreover, we also annotated data with the criteria. After that, we trained a deep learning model inspired by coreference resolution and dependency parsing models. As a result, our model achieves the F1 score of 82.72%.

1 Introduction

Generally, sentences have served as the basic linguistic units required by many tasks in NLP (e.g., text summarization and question answering) for processing long bodies of text (Mihalcea, 2004; Raiman & Miller, 2017; Van Lierde & Chow, 2019). Basic linguistic units must contain complete propositional content that represents a single piece of idea. However, in Thai language, the sentence cannot clearly specify the boundary (Intasaw & Aroonmanakun, 2013, p. 491; Lertpiya et al., 2018) since there is no explicit end-of-sentence marker like a period in English.

For this reason, prior works (Ketui et al., 2015; Singkul et al., 2019; Sinthupoun & Sornil, 2010; Sukvaree et al., 2004) have suggested to use the Elementary Discourse Unit (EDU), which is the basic unit in discourse based on Rhetorical Structure Theory (RST) (Mann & Thompson, 1988; Taboada & Mann, 2006), as a processing unit for Thai language. However, a single EDU, on its own, may not contain complete information to be understandable. Generally, EDUs are segmented based solely on syntactic criteria (Carlson & Marcu, 2001; Intasaw & Aroonmanakun, 2013; Ketui et al., 2012) and used alongside RST relation (which contains semantic and structural information) to represent complete discourse structure.

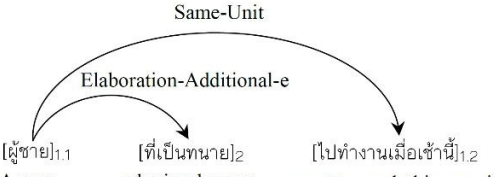


Figure 1. EDU that is separated due to embedded clause

To highlight that the EDU segmentation based on syntactic criteria creates individual EDUs with incomplete in structure and meaning, an example is shown in Figure 1. The figure illustrates RST d-tree (Morey et al., 2018) of EDU, which is separated by an embedded clause. A matrix clause [*ผู้ชายไปทำงานเมื่อเช้านี้*] 'A man went to work this morning', which should be one unit of EDU, is separated into two units, [*ผู้ชาย*] 'A man' and [*ไปทำงานเมื่อเช้านี้*] 'went to work this morning', because there is embedded clause

[*ที่เป็นทนาย*] 'who is a lawyer' modifying a noun [*ผู้ชาย*] 'A man' in matrix clause.

As such, combining incomplete EDUs is necessary for obtaining a complete idea that is to represent through a well-formed EDU.

This paper states the syntactic and semantic **criteria** for considering rhetorical relations used to combine EDUs into a well-formed EDU, which represents a complete idea. After that, rhetorical **relations** that correspond to those criteria are proposed.

We conducted experiments by building a dataset based on our proposed criteria and relations, which is then trained on a deep learning model. Our first experiment investigates methods to score the combination of EDU pairs. Since there is no prior work on this exact task, two methods from dependency parser task (Dozat & Manning, 2016) and coreference resolution task (Lee et al., 2017) were adapted on this task. In the second experiment, various EDU representations constructed from contextual word vectors are compared against one another. Lastly, we discuss how the result of our method covers the proposed rhetorical relations stated above.

The rest of this paper is structured as follows. In Section 2, the background knowledge related to this work is reviewed. The type of rhetorical relation to combine EDUs is explained in Section 3. While, in Section 4, we describe the architecture of our deep learning model. The dataset, implementation detail, and evaluation metrics are mentioned in Section 5. The results are shown in Section 6 and discussed in Section 7. Finally, Section 8 concludes the paper.

2 Theoretical Background

2.1 Elementary discourse unit

For processing to obtain single pieces of information from text, the text needs to be segmented into units that are related to each other. For this reason, Rhetorical structure theory (RST), a text organization theory, was proposed by Mann and Thompson (1988). RST is widely applied in text and discourse processing. The essence of this theory is to analyze relationships between subunits in the discourse, which is based on the intention of the messenger and the content of text. A tree diagram is then used to represent the relationships between subunits.

RST's theoretical concept was developed to be more practical by Carlson et al. (2001). In the discourse structure, there is the smallest unit that can convey complete content and meaning, called Elementary Discourse Unit (EDU). EDU can determine the nuclearity status of each unit and can also specify the type of rhetorical relation between them. After that, the criteria used to segment EDUs in different languages and sets of rhetorical relations were proposed in large numbers.

In this paper, the criteria for segmenting Thai EDUs that proposed by Intasaw and Aroonmanakun (2013) was applied.

2.2 The prior principle to combine discourse units.

Schauer (2000) states that not only clauses but sometimes prepositional phrases can also be considered as discourse units. To support this statement, three principles for including phrases as discourse units are proposed as follows.

First, consideration of the complement-status and adjunct-status of a clause or phrase is the principle that if any discourse unit has a status as a complement of another discourse unit, they will be combined to express the complete meaning. On the other hand, if any discourse unit has a status as an adjunct of another discourse unit, there is no need to combine them into a larger unit, since the main discourse unit is already meaningful.

Second, consideration of semantic specification of lexeme is to consider whether each word in discourse unit needs to rely on other linguistic units for expressing its lexical meaning. For example, the unit that indicates the agent, the patient, the instrument, the location, or the time frame. If the words in the discourse unit need to rely on another discourse unit in order to express their lexical meaning, those units will be combined.

Finally, consideration of discourse relation between units is the principle that the combined

discourse units always show the discourse relation between each other. These relations can be identified by the function of conjunctions at the beginning of the unit.

In this paper, the principles proposed by Schauer (2000) are applied as criteria to combine EDUs by considering the rhetorical relations between EDUs that are related to syntactic and semantic characteristics of text. The details on applying those criteria are presented in the next section.

3 Types of Rhetorical Relation to Combine

By applying the concepts from Schauer (2000), EDUs occurred in discourse structure can be considered as a clause or a phrase with the status of complement or adjunct. Besides that, four types of rhetorical relations only used to link between the matrix clause and unit with complement-status. The units with complement-status are necessary to express the complete meaning. If such units are omitted, the expressed meaning is inadequate. The criteria for combing EDUs are applied by using four types of rhetorical relations that are linked between units with complement-status. The details of each type are explained in the following section.

3.1 Attribution

The attribution relation is used to link between a clause containing a reporting verb or attribution verb, e.g. *พูด* 'speak', *รายงาน* 'report', *บอก* 'tell', *คิด* 'think', *สั่ง* 'order', and a clause containing the content of the reported message (Carlson & Marcu, 2001, p. 46).

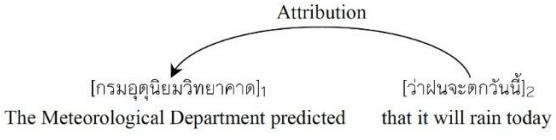


Figure 2. EDUs that linked by attribution relation

As shown in Figure 2, EDU_2 contains the content of the reporting verb *คาด* 'predict' in EDU_1. In terms of syntactic and semantic characteristics, EDU_2 is a complement clause of the main verb in EDU_1. If there is any part missing, the incomplete meaning will be conveyed. For this reason, both EDUs must be combined.

3.2 Attribution-negative

The attribution-negative has properties similar to the attribution relation. The difference is that the attribution-negative is used for marking a negative attribution.

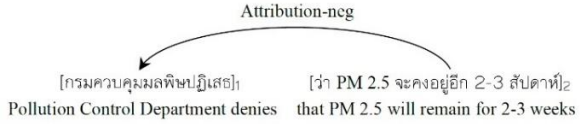


Figure 3. EDUs that linked by attribution-negative relation

Like the attribution relation, in Figure 3, EDU_2 contains the content which cannot be omitted of an attributive verb *ปฏิเสธ* 'deny' in EDU_1. But, in this case, the main verb appeared in EDU_1 is semantically negative. Therefore, the attribution-negative is applied.

3.3 Elaboration-object-attribute

Elaboration-object-attribute is a relation involving a clause, usually a postmodifier of a noun phrase in the matrix clause, that is required to give meaning to an animate or inanimate object (Carlson & Marcu, 2001, p. 54). In this case, omitting modifier clauses may result in incomplete meaning.

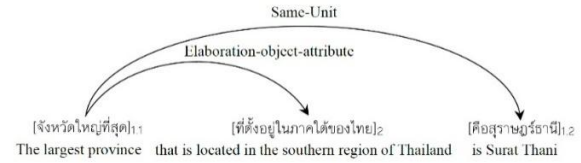


Figure 4. EDUs that linked by elaboration-object-attribute

Figure 4 shows that $EDU_{1.1}$ need EDU_2 as a modifier for a noun phrase *จังหวัดใหญ่ที่สุด* 'the largest province'. If EDU_2 is omitted, the combination of $EDU_{1.1}$ and $EDU_{1.2}$ will mean 'The largest province is Surat Thani', which is the meaning that causes misunderstandings. Therefore, to become a meaningful unit, EDU_2 is a complement that must always be combined to $EDU_{1.1}$.

3.4 Same-unit

Same-unit is pseudo-relation used as a device for linking two discontinuous text fragments that are

really a single EDU, but which are broken up by an embedded unit (Carlson & Marcu, 2001, p. 66).

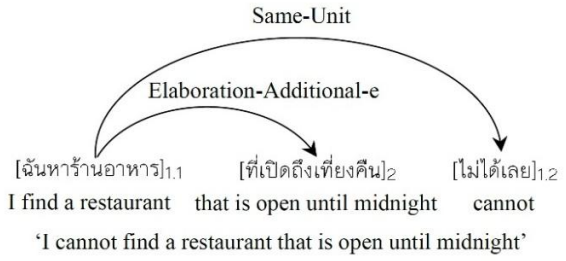


Figure 5. Single EDU that is separated due to embedded clause

Considering Figure 5, a matrix clause *ฉันหาร้านอาหารไม่ได้เลย* 'I cannot find a restaurant', which is a single EDU, is broken up into 2 units by an embedded clause with adjunct-status *ที่เปิดถึงเที่ยงคืน* 'that is open until midnight' modifying a noun *ร้านอาหาร* 'a restaurant' in a matrix clause. In this case, $EDU_{1.1}$ [*ฉันหาร้านอาหาร*] 'I find a restaurant' and $EDU_{1.2}$ [*ไม่ได้เลย*] 'cannot' have to be combined to represent the meaning as a single EDU.

The types of rhetorical relations presented in this section are used as theoretical backgrounds for implementing combining EDUs model. Details are discussed in the next section.

4 Model architecture

The architecture of our deep learning model for combining EDUs is presented in this section. The architecture, as shown in Figure 6, is separated into four parts: Contextual word representation, EDU representation, EDU combination scoring, and the training and inference process. The detail of each part is elaborated in the following subsections.

4.1 Contextual word representation

This module is responsible for converting a sequence of words $\vec{w} = [w_1, w_2, \dots, w_N]$ and the corresponding part-of-speech tags (POS) $\vec{p} = [p_1, p_2, \dots, p_N]$ into a sequence of contextual word vectors $H = [\overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_N}]$ where N is the word sequence length, as illustrated in Figure 6. First, each word and its POS in the sequence is converted into embedding vectors $E = [\overrightarrow{e_1}, \overrightarrow{e_2}, \dots, \overrightarrow{e_N}]$ where $\overrightarrow{e_t}$ is the concatenation of word embeddings and POS embeddings. After that, the sequence of concatenated embedding vectors is fed into Bidirectional Long-short memory network (Bi-LSTM) to create contextual word vectors H.

4.2 EDU representation

This module aggregates the contextual word

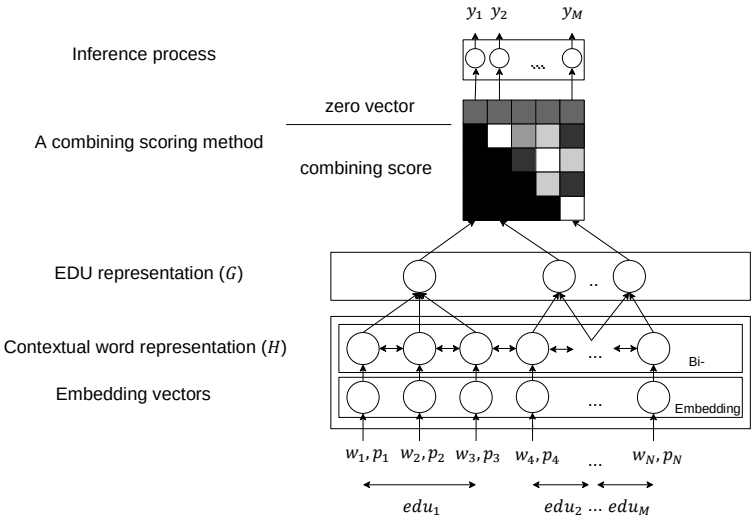


Figure 6. Model Architecture

vectors H into EDU representation $G = [\overrightarrow{g_1}, \overrightarrow{g_2}, \dots, \overrightarrow{g_M}]$ where M is the number of EDUs

and $\overrightarrow{g_t}$ represents one EDU (edu_t). For this module, we adopt two methods to create the EDU representation. These methods were proposed by Lee et al. (2017) to create span representation for performing coreference resolution. Therefore, in this module, each EDU vector $\overrightarrow{g_t}$ is concatenated from an end-point vector $\vec{g}_t^{(ep)}$ and a self-attention vector $\vec{g}_t^{(atten)}$,which are described in the following subsections.

End-point representation presents each edu_t by concatenating with three parts, as shown in Eq. 1. The beginning and end contextual word vectors ($\vec{h}_{begin(t)}, \vec{h}_{end(t)}$) are concatenated with the length of the EDU in words ($\phi(t)$). Therefore, this representation captures the keyword at the beginning and the end alongside its length.

$$\vec{g}_t^{(ep)} = [\vec{h}_{begin(t)}, \vec{h}_{end(t)}, \phi(t)] \#1$$

Self-attention representation is the weighted summation of contextual word vectors in edu_t, as shown in Eq. 2. Each word is assigned a weight which is computed from $\overrightarrow{\mathrm{h_k}}$. First, each contextual word vector ($\overrightarrow{\mathrm{h_k}}$) is fed to a linear function ($\mathrm{f_{gate}}$). Second, the output vector is applied with Softmax function to calculate the weight for each word.

$$\vec{g}_t^{(atten)} = \sum_{\mathrm{k=begin(t)}}^{\mathrm{end(t)}} \mathrm{softmax}\left(\mathrm{f_{gate}}\left(\overrightarrow{\mathrm{h_k}}\right)\right) \cdot \overrightarrow{\mathrm{h_k}} \#2$$

4.3 EDU combination scoring

This module is responsible for calculating the scores $s_t \in \mathbb{R}^{\{t-1\}}$ for edu_t where $s_{t,j}$ is a score between edu_t and edu_j. The two proposed methods are inspired by dependency parser and coreference resolution tasks.

Dependency-parsing-based method is based on biaffine dependency parsing, proposed by Dozat and Manning (2016). First, each EDU representation $\overrightarrow{g_t}$ is embedded into the child vector $\vec{h}_t^{(ch)}$ and the parent vector $\vec{h}_t^{(pa)}$ by linear functions (f_{ch} and f_{pa}), as shown in Eq. 3 and Eq. 4.

$$\vec{h}_t^{(ch)} = f_{ch}(\overrightarrow{g_t}) \#3$$

$$\vec{h}_t^{(pa)} = f_{pa}(\overrightarrow{g_t}) \#4$$

After that, both sequences of vectors (H_{ch} and H_{pc}) are applied to bilinear matrix attention to compute the scores $\overrightarrow{s_t}$, which are computed as shown in Eq. 5.

$$\overrightarrow{s_t} = \left(H^{(ch)} \oplus \vec{1}\right) \times U \times \vec{h}_t^{(pa)} \#5$$

Coreference-based method is based on end-to-end coreference resolution, Lee et al. (2017), as shown in Eq. 6. The score of each pair of EDU (edu_t, edu_i) is composed of three parts: two individual scores ($s_t^{(in)}, s_i^{(in)}$) of EDUs and the antecedent score calculated from both EDUs ($s_{t,i}^{(a)}$).

$$s_{t,i} = s_t^{(in)} + s_i^{(in)} + s_{t,i}^{(a)} \#6$$

Both types of scores are calculated based on linear functions (f_{in}, f_a), which are describes below:

$$s_t^{(in)} = w_{in} \cdot f_{in}(\overrightarrow{g_t}) \#7$$

$$s_{t,i}^{(a)} = w_a \cdot f_a\left(\overrightarrow{g_t}, \overrightarrow{g_i}, \overrightarrow{g_t} \circ \overrightarrow{g_i}, \zeta(t,i)\right) \#8$$

where $\cdot$ denotes the dot product, $\circ$ denotes element-wise multiplication, w_{in} and w_a are weights for calculating an individual score and an antecedent score respectively, and $\zeta(t,i)$ is the distance between edu_t and edu_i.

4.4 Inference and training process

This process is fed with combination scores of each EDU. In this case, zero is added to the list of scores to represent an individual EDU. Then, Softmax function is applied to the scores to find the probability distribution of combining edu_t.

The probability indicates if the EDU should be combined and which EDU is combined with. The answer is the position of highest probability y_t, as shown in Eq. 9. If y_t is the index of added zero, the considered EDU is individual. In the other hand, if y_t is other position, the considered EDU is combined with the EDU at y_t.

$$y_t = \mathrm{argmax}\left(\mathrm{softmax}([0, \overrightarrow{s_t}])\right) \#9$$

In the training process, the marginal log-likelihood of all correct combined EDU position ($\hat{y}_t$), which is indicated from the label. The calculation of the loss L is shown in Eq. 10.

$$L = log \prod_{t=1}^{M} P(\hat{y}_t) \ \#10$$

5 Experimental setup

5.1 Dataset

In this work, the dataset is collected from social media sources, including conversations and posts. After that, the text is segmented into EDUs by our in-house model and then combined EDUs from mentioned criteria by linguists. The dataset contains 161,515 arbitrary texts, which can be segmented into 847,186 EDUs. There are 59,564 pairs of EDUs that are combined.

In the annotation process, we also specify which relation in the criteria is used to combine. The number of each relation in the data is shown in Table 1.

type	# of relations
attribution	23,949
attribution-n	3,948
elaboration-object-attribute	23,971
same-unit	7,696

Table 1. The frequency of each relation that is used to combine EDUs

In pre-processing, in-house models that are trained by social media data are used (Lertpiya et al., 2018). The text is tokenized into a sequence of words. After that, each word is tagged with part-of-speech.

In the training and inference process, the dataset is randomly split with a ratio 9:1 for training set and testing set, respectively. After that, we split 10% of training set for validation set. Meanwhile, the rest of training set is truly used for training the model.

5.2 Implementation details

The hyperparameters of the trained model and the optimizer are described in this section. The word and POS embedding sizes are 300 and 100. The word embedding is pre-trained on social media data with the Skip-gram technique (Mikolov et al., 2013). Four layers of Bi-LSTM are stacked. The hidden size of each Bi-LSTM is 32, and there are dropout layers whose rate is 0.1 between the layers. In end-point representation, the size of length embedding $\phi(t)$ is 20.

In the dependency-based scoring method, the hidden sizes of fully connected layers f_{ch} and f_{pa} are both 64. Meanwhile, in the coreference-based scoring method, the embedding size of the distance $\zeta(t,i)$ is 20. The hidden size of fully connected layers f_{in} and f_a are 30 and 150, respectively.

The optimizer is ADAM (Zhang, 2018), whose initial learning rate is 0.001. The learning rate is reduced by 0.5 when the F1 score has stopped improving for five epochs. The batch size is 16. The model is trained 40 epochs and selects the model, which gains the highest F1 score on validation set.

5.3 Evaluation

The metric for this paper should have three characteristics. First, the metric reflects the intersection between combined EDUs label and prediction to measure the performance of the model. Second, since a combined EDUs is not always limited to be constructed from only two EDUs, so the metric needs to consider combining more than two EDUs. Finally, the metric ignores the order or direction of relations that are used to combine because combining EDUs does not need to be interested in the relations.

According to, a coreference resolution task also considers those characteristics for evaluation. Therefore, the metrics from coreference resolution are chosen for evaluation for this paper. However, there are many methods proposed for calculating F1 on coreference resolution. Each method has different advantages and drawbacks. Thus, the average of F1 that used in this paper are calculated from three methods: MUC (Vilain et al., 1995), B^3 (Bagga & Baldwin, 1998), and $CEAF_{\varphi_4}$ (Luo, 2005) as same as Lee et al. (2017).

6 Results

The results of two experiments are discussed in this section. The first experiment compares the different EDU combination scoring method. Meanwhile, the second experiment shows the performance of each EDU representation constructed from contextual word vectors.

6.1 EDU combination scoring

The different of EDU combination scoring methods are compared in the experiment. In this case, EDU representation module is composed of only end-point module. Table 2 shows that scoring with the coreference-based method outperforms the dependency-based method in terms of F1 score. The result occurs because the coreference-based method includes the distance between considered EDUs in a score calculation. Meanwhile, the dependency-based method exploits only the representation of the considered EDUs.

representation	precision (%)	recall (%)	F1 (%)
endpoint	82.81	82.41	82.61
self-attention	83.10	81.82	82.46
both	83.04	82.41	82.72

Table 2. The comparison of each combining scoring method

Figure 7 shows the frequency of the distance between combined EDUs to prove the aforementioned statement. The result indicates that the distance of combined EDUs is usually short.

In other words, if EDUs are far from each other, those EDUs should not be combined. Therefore, the distance is an important feature for the model on combining EDUs. As a result, the coreference-based method, which includes the distance feature, can perform better than the dependency-based method.

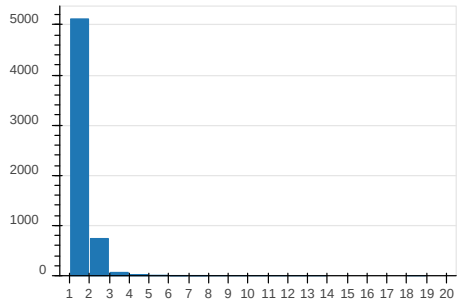


Figure 7. Distance between combined EDU

6.2 EDU representation

In this section, the difference between an end-point representation and self-attention representation is shown. Table 3 shows the results that, by using end-point representation, the F1 is slightly higher than using self-attention representation.

scoring method	precision (%)	recall (%)	F1 (%)
coreference-based	82.81	82.41	82.61
dependency-based	77.28	83.08	80.06

Table 3. The comparison of each representation

The reason is that most of combined EDUs contain a complementizer (‘ว่า’, ‘ที่’) as a marker, which is usually the first word of preceded EDU in the combined EDUs. The statement can be proved in Figure 8. This figure shows that the two most frequent words at the beginning of the preceded EDU in combined EDU are the mentioned complementizers. Therefore, the end-point representation can be trained easier because the representation focuses on the beginning and end words of the EDU.

Instead of using each representation separately, both representations are trained in the model. The model achieves the highest score at 82.72% in terms of F1.

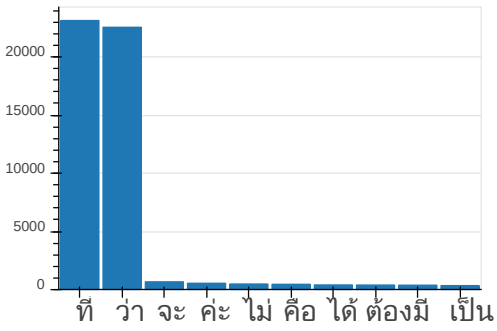


Figure 8. The frequency of the first word on the preceded EDU in the combined EDUs

7 Discussion

According to the results of the above, the discussion of the model is presented in this section. The details are focused on the process that the best model works on each relation for combining EDUs. Table 4 shows the recall of each relation. In this case, because the model does not concern about which class is predicted, so the precision of each relation is not evaluated.

Relation type	Recall (%)
attribution	82.17
attribution-n	95.02
elaboration-object-attribute	71.35
same-unit	53.01

Table 4. Recall score on each relation type that is used to combine EDUs

The model achieves 82.17% and 95.02% on 'attribution' and 'attribution-n'. According to, there is a marker like a complementizer 'ว่า' to indicate that they should be combined.

Meanwhile, 'elaboration-object attribute' is usually indicated with a complementizer 'ที่' as a marker. However, this complementizer is also included as a marker in other RST relations. Therefore, the model achieves only 71.35% recalls on this relation due to the various usage of the marker.

Besides that, the model cannot perform well on 'same-unit' relation, of which recall is only 53.01%. The reason is that there is no marker to classify this relation. Moreover, this relation is rarely found on our dataset. Therefore, there is insufficient information for the model to learn this relation without any marker.

8 Conclusion

In this paper, we propose the criteria of both the syntactic and semantic way for selecting rhetorical relations, which are used to combine EDUs. Moreover, the dataset was annotated by using those criteria before being used to train a deep learning model.

Two experiments were conducted to find the best configuration of the model. In the EDU combination scoring, the coreference-based method achieved a better F1 score. We suspect this because a distance between EDUs is correlated to its likeliness to combine. Meanwhile, using both proposed EDU representations lead to the best F1 score for combining EDUs. The best model achieves 82.72% in terms of F1 score. In an ablation study, the model works well on 'attribution' and 'attribution-n' relation due to the keywords at the beginning of EDU. However, 'same-unit' relation is hard for the model to predict as a relation for combining EDUs because there is no keyword to guide the model and lack of annotated data.

In this paper, we have primarily focused on the combining EDUs task. However, further experiments should be performed to evaluate how the use of combined EDUs affects downstream tasks (e.g., intention classification and sentiment analysis).

References

Bagga, A., & Baldwin, B. (1998). Entity-Based Cross-Document Coreferencing Using the Vector Space Model. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, 79–85.

Carlson, L., & Marcu, D. (2001). Discourse tagging reference manual. University of Southern California Information Sciences Institute.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 16, 1–10.

Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734.

Intasaw, N., & Aroonmanakun, W. (2013). Basic principles for segmenting Thai EDUs. Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), 491–498.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2012). A rule-based method for thai elementary discourse unit segmentation (ted-seg). 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems, 195–202.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2015). An EDU-Based Approach for Thai Multi-Document Summarization and Its Application. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 14(1), Article 4.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045.

Lertpiya, A., Chaiwachirasak, T., Maharattanamalai, N., Lapjaturapit, T., Chalothorn, T., Tirasaroj, N.,

& Chuangsuwanich, E. (2018). A preliminary study on fundamental thai nlp tasks for user-generated web content. 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP), 1–8.

Luo, X. (2005). On coreference resolution performance metrics. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 25–32.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3), 243–281.

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proceedings of the ACL Interactive Poster and Demonstration Sessions, 170–173.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.

Morey, M., Muller, P., & Asher, N. (2018). A dependency perspective on rst discourse parsing and evaluation. Computational Linguistics, 44(2), 197–235.

Raiman, J., & Miller, J. (2017). Globally normalized reader. arXiv preprint arXiv:1709.02828.

Schauer, H. (2000). From elementary discourse units to complex ones. 1st SIGdial Workshop on Discourse and Dialogue, 46–55.

Singkul, S., Khampingyot, B., Maharattamalai, N., Taerungruang, S., & Chalothorn, T. (2020). Parsing Thai Social Data: A New Challenge for Thai NLP. 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP), 1–7.

Sinthupoun, S., & Sornil, O. (2010). Thai rhetorical structure analysis. International Journal of Computer Science and Information Security (IJCSIS), 7(1), 95-105.

Sukvaree, T., Charoensuk, J., Wattanamethanont, M., & Kultrakul, A. (2004). RST based Text Summarization with Ontology Driven in Agriculture Domain. Department of Computer Engineering, Kasetsart University, Bangkok, Thailand.

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: looking back and moving ahead. Discourse Studies, 8(3), 423-459.

Van Lierde, H., & Chow, T. W. (2019). Query-oriented text summarization based on hypergraph transversals. Information Processing & Management, 56(4), 1317-1338.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995, 45–52.

Zhang, Z. (2018). Improved adam optimizer for deep neural networks. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 1–2.

Evaluation of Pretrained BERT Model by Using Sentence Clustering

Naoki Shibayama Rui Cao Jing Bai Wen Ma Hiroyuki Shinnou
Ibaraki University, Department of Computer and Information Sciences
4-12-1 Nakanarusawa, Hitachi, Ibaraki JAPAN 316-8511
{19nm714t, 18nd305g, 19nd301r, 19nd302h, hiroyuki.shinnou.0828}
@vc.ibaraki.ac.jp

Abstract

For evaluation of pre-trained models like bidirectional encoder representations from transformers (BERT), task-based approaches are frequently adopted and there is a possibility that meta parameters for fine-tuning influence results of the evaluations. However, task-based approaches for languages, except English, have a problem- there is no common dataset for their evaluation. Hence, evaluating pre-trained models for these languages with task-based approaches is challenging. In this work, we evaluate Japanese pre-trained BERT models with CLS token. We input labeled sentences to models, get CLS token embeddings, and calculate scores from in-class and out-of-class dispersions, which can be calculated from embeddings and labels of sentences. Experiment results show that a model released by Laboro.AI Inc. is the best Japanese pre-trained BERT model. Meanwhile, the results of evaluation with sentence clustering are different from those of evaluations that are based on fill mask task.

1 Introduction

BERT (Devlin et al., 2019) is a high-performance pre-training model. It helped in the improvement of the performance of natural language processing tasks. Generally, task-based approaches were adopted for evaluating pre-training models like BERT. In English language, a dataset for task-based evaluation, such as the general language understanding evaluation (GLUE) (Wang et al., 2018), can be used, and it is easy to compare models. However, when a pre-trained model is fine-tuned for task-based evaluation, meta parameters for fine-tuning may influence scores of the model. Hence, task-based evaluation with fine-tuning has a possibility of biased evaluation. Also, there is no common task-based dataset for languages except English, so it is challenging to compare pre-trained models for other languages.

In this work, we evaluate Japanese pre-trained BERT models using CLS token embeddings in outputs of target models. CLS token embedding can be regarded as an input sentence embedding, and models can be rated with evaluating embeddings itself. However, how to evaluate sentence embeddings is also challenging. Here, we use clustering to evaluate sentence embeddings. Also, we prepare sets of sentences sorted by genre and use BERT models to get embeddings of each sentence. Then, we cluster those embeddings and evaluate models with clustering score.

2 Related Works

Generally, a task-based approach for evaluation is adopted to compare and evaluate pre-trained models like BERT. Although this simple method requires data for evaluation, it consists of the following 3 steps:

1. Solve a task with pre-trained model A and get its accuracy.
2. Solve this same task with pre-trained model B and get its accuracy.
3. Compare the accuracies and evaluate models A

and B.

The GLUE can be used for English, but there is no common dataset for other languages, so we have to prepare the dataset for evaluation ourselves.

There is a work that compared and evaluated some Japanese pre-trained BERT models. In this work, we evaluated three BERT models using document classification tasks with the Amazon dataset (Shibayama et al., 2019). However, BERT is a model for sentences, and there is no established method of document classification with BERT. Therefore, whether document classification is the right task to evaluate or not is questionable. We use a sentence as input of BERT and evaluate models using CLS token embeddings, which can be considered as sentence embeddings from outputs of BERT.

The approaches for evaluation of embeddings are task-based, but in the case of word embeddings from outputs of some method like word2vec (Mikolov et al., 2013), there is a viewpoint that embeddings represent the meaning of words. Also, there is a research that evaluated embeddings with correlation of similarities between words calculated from the similarity of embeddings and by hand (Sakaizawa and Komachi, 2016).

3 Evaluation of BERT

In Section 2, we mentioned that a task-based approach is frequently adopted to evaluate embeddings. Also, we mentioned that there is a viewpoint that embeddings represent the meaning of words. When this viewpoint is applied to clustering, we can say that a cluster can be represented by a group of embeddings in it. In what follows, we use this to evaluate pre-trained BERT models with sentence clustering.

3.1 Method of the Evaluation

Embeddings that were outputted from BERT model m, were evaluated by the following 5 steps. Labels for sentences of model m's input were required to do this evaluation.

1. Get CLS token's embedding from the output of each sentence of model m, and use the embedding as the sentence vector.
2. Check which class contains the sentence vector, and calculate $g_i^{(m)}$: centroid of each class of model m.
3. Calculate A_m: in-class dispersion of each class from the following expression[1].

$$A_m = \sum_{i=1}^{N} \sigma_i^2 \tag{1}$$

where $\sigma_i^2 = \sum_{j \in C_i} ||g_i^{(m)} - x_{i,j}||^2$, C_i is class i and N is number of classes.

4. Calculate $g^{(m)}$: average centroids of all classes and calculate B_m: out-of-class dispersion from the following expression.[2]

$$B_m = \sum_{i=1}^{N} ||g^{(m)} - g_i^{(m)}||^2 (N = \text{Number of classes}) \tag{2}$$

5. Calculate a degree of separation: $M_m = \frac{A_m}{B_m}$, and use M_m as a score of model m. This score becomes smaller when clustering with model m is performed properly.

Figure 1 summarizes the flow of the evaluation.

3.2 Re-evaluation by Using Fill Mask Task

We re-evaluated models with a fill mask task in order to verify the results of sentence clustering evaluation. The steps for the re-evaluation are as following:

1. Prepare a dataset- we prepared a dataset that contains sentences and which word to be masked in matching sentence as labels.
2. Predict masked word with model- we calculated percentages that mask token was the word in matching label which was defined in a dataset from outputs of models.
3. Average and comparison- we compared averages of percentages that were calculated in step 2.

[1] We consider the second power of deviation as the dispersion in this work in order to calculate easily. So, true in-class dispersion can be calculated from σ_i^2/N.

[2] Also, we consider the second power of deviation of centroids of all classes as dispersion like A_m.

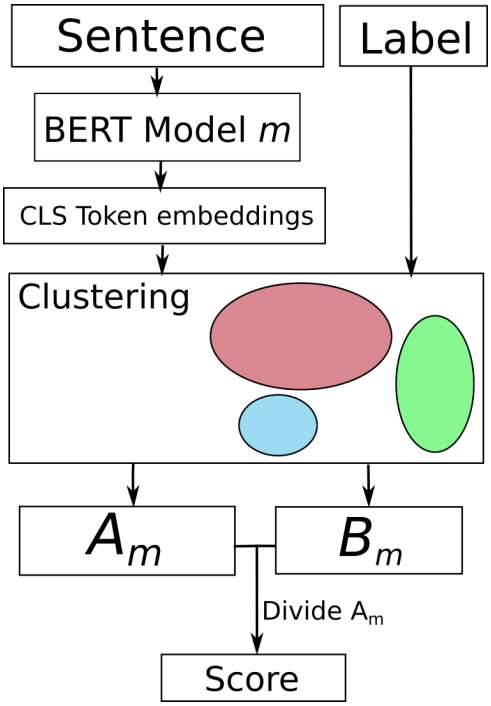


Figure 1: The flow of evaluation with sentence clustering

Detailed information on the abovementioned comparison is in the next subsection.

3.3 Experimental Setups

Firstly, we explain common setups for evaluation with sentence clustering and fill mask task. We compared six models: a model which was released by Kurohasi & Murawaki Lab at Kyoto University (hereafter, Kyoto Univ. Ver.)[3], Mr. Morinaga (hereafter, MeCab Ver.)[4], Mr. Yohei Kikuta (hereafter, SP Ver.)[5], Inui & Suzuki Lab at Tohoku University (hereafter, Tohoku Univ. Ver.)[6], National Institute of Information and Communications and Technology (NICT) (hereafter, NICT Ver.)[7], and Laboro.AI Inc.(hereafter, Laboro Ver.)[8]. We did not fine-tune models for comparisons.

Table 1 summarizes the word tokenizer and pre-training corpus of pre-trained models. Model size of all models are base format of official BERT(Devlin et al., 2019): 12-layer, 768-hidden, and 12-heads. So sentence vectors we got in the evaluation with sentence clustering have 768 dimentions. Juman++ and MeCab are software for morphological analysis. Juman++ uses Recurrent Neural Network Language Model and MeCab uses bi-gram Markov model for analysing. SentencePiece is unsupervised text tokenizer and detokenizer, so model publishers which use SentencePiece as tokenizer release pre-trained SentencePiece model with their BERT model.

Table 1: Information of pre-trained BERT models

Models	Tokenizer (characteristic)	Pre-training corpus
Kyoto Univ. Ver.	Juman++	Wikipedia
MeCab Ver.	MeCab + NEologd (No subword tokenize)	Articles of business news
SP Ver.	SentencePiece (do_lower_case = True)	Wikipedia
Tohoku Univ. Ver.	MeCab + NEologd	Wikipedia
NICT Ver.	MeCab + Jumandic	Wikipedia
Laboro Ver.	SentencePiece	Texts on the Internet (12GB)

In the evaluation with sentence clustering, we used Livedoor news corpus[9]. This dataset contains nine categories of articles and we used one hundred articles per category. We extracted titles from selected articles and regarded categories as classes. Then, we calculated scores with the method in Section 3.1 and compared these scores.

In the evaluation with fill mask task, we made a fill mask dataset from Japanese domain of Webis-

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル We used normal version.

[4] https://qiita.com/mkt3/items/3c1278339ff1bcc0187f

[5] https://github.com/yoheikikuta/bert-japanese

[6] https://github.com/cl-tohoku/bert-japanese This model can be used easily as "cl-tohoku/bert-base-japanese" from HuggingFace's transformers (Wolf et al., 2019), and we used it.

[7] https://alaginrc.nict.go.jp/nict-bert/index.html We used the byte pair encoding (BPE) version.

[8] https://laboro.ai/column/laboro-bert/

[9] http://www.rondhuit.com/download.html#ldcc

CLS-10 (Prettenhofer and Stein, 2010) and used it. The following two steps show how to make a fill mask dataset from Webis.

1. Pick twenty nouns that have the highest frequencies of occurence from the test data of each domain: books, DVDs, and music.

2. Pick five sentences that contain matching selected words from test data of the matching domain randomly to each selected word.

3. Use nouns which were selected in step 1 as labels for matching sentences which were selected in step 2.

We replaced "selected nouns" which appeared for the first time in matching sentence with mask token. The following shows selected Japanese nouns for each domain.

books: 本, 人, 著者, 内容, 自分, 作品, 本書, 感じ, 文章, 主人公, 小説, 部分, 最後, 言葉, 読者, 作者, 人間, 物語, 他, 世界
DVDs: 映画, 作品, 人, シーン, 映像, 原作, 自分, ストーリー, 内容, ファン, 感じ, 主人公, 最後, アニメ, ドラマ, 物語, 人間, 世界, 子供, 部分
music: 曲, アルバム, 作品, 人, 音楽, 感じ, ファン, 音, バンド, 自分, 歌詞, 声, ギター, 歌, ＣＤ, 楽曲, サウンド, ライブ, シングル, 前作

We calculated percentages that mask token is the word in matching label with prepared dataset and each model. Then, we averaged percentages and compared these. The following shows notices of this comparison.

- We used transformers (Wolf et al., 2019) to solve the fill mask task.

- We replaced "ＣＤ" (Fullwidth form of CD) with CD (Halfwidth form of CD).

- We lowercased sentences when we used SP Ver. model. We did not lowercase sentences when we used SP Ver. model for the first time, and the model tokenized CD as token "C" and token "D". We checked vocabulary file of the model and found the word "cd", not "CD". So we recognized we needed to activate lowercasing option.

4 Results

In this section, we show the resluts of the evaluations. First, we show the result of the evaluation with sentence clustering, and then the result of evaluation with fill mask task.

4.1 Result of Evaluation with Sentence Clustering

Table 2 summarizes A_m, B_m, and scores of evaluation with sentence clustering. Bigger B_m is better, and smaller A_m and score are better.

Table 2: Values and scores of evaluation with sentence clustering

Models	A_m	B_m	Score
Kyoto Univ. Ver.	240131.79	337.83	710.81
MeCab Ver.	97536.21	154.37	631.06
SP Ver.	67744.36	104.05	651.06
Tohoku Univ. Ver.	49991.31	65.64	761.58
NICT Ver.	106698.11	151.27	705.37
Laboro Ver.	153378.22	273.83	**560.13**

The following shows the results of comparing models by score, A_m, and B_m, and figure 2 is a bar graph of the results.

Score: Laboro Ver. $<$ MeCab Ver. $<$ SP Ver. $<$ NICT Ver. $<$ Kyoto Univ. Ver. $<$ Tohoku Univ.Ver.
A_m: Tohoku Univ. Ver. $<$ SP Ver. $<$ MeCab Ver. $<$ NICT Ver. $<$ Laboro Ver. $<$ Kyoto Univ. Ver.
B_m : Kyoto Univ. Ver. $>$ Laboro Ver. $>$ MeCab Ver. $>$ NICT Ver. $>$ SP Ver. $>$ Tohoku Univ.Ver.

Table 3 shows scores of models except NICT Ver., and Laboro Ver. in previous work (Shibayama et al., 2020) and this work. According to this table, Scores of models except MeCab Ver. changes about 0.8–30 from results in previous evaluation (Shibayama et al., 2020). These changes did not influence the results of comparisons. Score of MeCab Ver. model became 100 or more higher than the previous result, but this change also did not influence results.

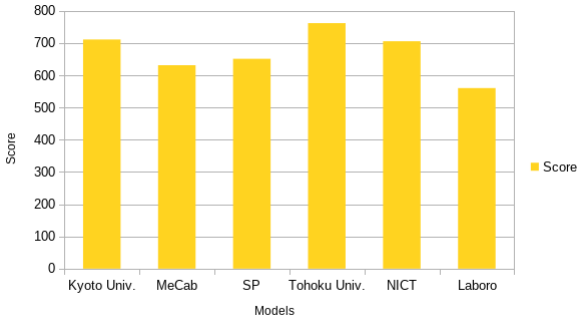


Figure 2: The results of comparing models by score

Table 3: Scores of previous work(Shibayama et al., 2020) and this work

Models	previous	this
Kyoto Univ. Ver.	710.88	710.81
MeCab Ver.	458.19	631.06
SP Ver.	668.92	651.06
Tohoku Univ. Ver.	792.34	761.58

4.2 Result of Re-evaluation with Fill Mask Task

Table 4 shows average of percentages that mask token is the word in matching label of all domains and three domains: books, DVDs, and music. Figure 3 shows a bar graph of column "All" in table 4.

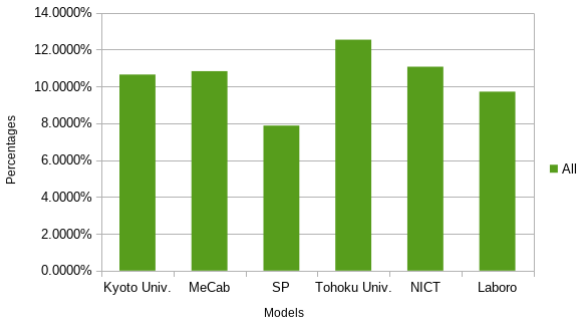


Figure 3: Averages of percentages of all domains

The following shows the result of comparing models by percentages of all domains, and this is different from the results in Section 4.1.

Table 4: Average of percentages that mask token is true masked word

Models	books	DVDs	music	All
Kyoto Univ. Ver.	11.53%	11.18%	9.24%	10.65%
MeCab Ver.	11.24%	13.62%	7.62%	10.83%
SP Ver.	7.36%	9.86%	6.41%	7.88%
Tohoku Univ. Ver.	14.04%	12.76%	10.81%	**12.54%**
NICT Ver.	11.90%	12.63%	8.68%	11.07%
Laboro Ver.	8.86%	10.44%	9.85%	9.72%

Tohoku Univ. Ver. > NICT Ver. > MeCab Ver. > Kyoto Univ. Ver. > Laboro Ver. > SP Ver.

5 Discussion

In this section, we describe the results in Section 4, and why there is a difference between the results in Section 4.1 and Section 4.2.

We changed the tokenizer settings for MeCab Ver. model from previous evaluation not to use subword tokenize[10]. We think this influenced the score of MeCab Ver. model, which caused a difference from a previous work (Shibayama et al., 2020).

As mentioned earlier, we considered A_m and B_m as in-class and out-of-class dispersion, respectively, in order to calculate easily (see, footnotes of Section 3.1). Therefore, comparing A_m means evaluating whether embeddings in the same class are close, and B_m means evaluating differences of embeddings that are not in the same class. We can dedude the general tendencies of each model from the results in Section 4.1. The best model is Laboro Ver., which has the second-highest B_m and about 100000 smaller A_m than Kyoto Univ. Ver. model. MeCab Ver. model that has the best score in previous eval-

[10]According to an article of MeCab Ver. model, we have to change scripts that use only MeCab as a tokenizer.

uation (Shibayama et al., 2020) is the second-best model. SP Ver. model is the third, which A_m of model is it of Tohoku Univ.Ver. model or more and it of MeCab Ver. model less. Tohoku Univ. Ver. model has the worst score, which has smallest A_m and B_m. This means the dispersion of all embeddings is smaller than the other models.

However, the results in Section 4.1 are different from the results in Section 4.2. Thus, we could not conclude that the results of methods of evaluation with sentence clustering and fill mask task have the same tendency. We used the title of articles in evaluation with sentence clustering, but we used a sentence in product reviews (see synopsis of Webis-CLS-10 (Prettenhofer and Stein, 2010)) with fill mask task. This difference may have caused the differences between the results in Section 4.1 and Section 4.2.

6 Conclusion

We evaluated Japanese pre-trained BERT models using sentences that were labeled, and outputs of BERT that inputted those sentences. Then, we obtained the following result.

> Laboro Ver. $<$ MeCab Ver. $<$ SP Ver. $<$ NICT Ver. $<$ Kyoto Univ. Ver. $<$ Tohoku Univ. Ver.

Also, we masked a specific noun in each sentence, calculated percentage that mask token is the word in matching label, and re-evaluated with averages of that percentage. However, we obtained the following result, and this is different from result of sentence clustering.

> Tohoku Univ. Ver. $>$ NICT Ver. $>$ MeCab Ver. $>$ Kyoto Univ. Ver $>$ Laboro Ver. $>$ SP Ver.

If we decrease the difference of type or domain of documents that are used in both experiments, there is a chance that the comparison results will be different from what we obtained in this work.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP19K12093.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-2019*, pages 4171–4186.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS-2013*, pages 3111–3119.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pages 1118–1127. Association for Computational Linguistics, July.

Yuuya Sakaizawa and Mamoru Komachi. 2016. Building similarity dataset of japanese verbs and adjectives (in Japanese). *The Twenty-second Annual Meeting of the Association for Natural Language Processing*, pages 258–261.

Naoki Shibayama, Rui Cao, Jing Bai, Wen Ma, and Hiroyuki Shinnou. 2019. A comparison of japanese pretrained bert models (in Japanese). *IEICE Techn. Rep.*, 119(212):89–92.

Naoki Shibayama, Rui Cao, Jing Bai, Wen Ma, and Hiroyuki Shinnou. 2020. Evaluation of pretrained BERT model by using sentence clustering (in Japanese). In *The Twenty-sixth Annual Meeting of the Association for Natural Language Processing*, pages 1233–1236.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Basic Hyperparameters of Models We Evaluated

In this section, we show basic hyperparameters of pre-trained BERT models we evaluated. However, some parameters were not written in both config file of model and model publisher's web site. Table 5 shows basic pre-training information of pre-trained models. "No Info" cell is a parameter that we could not found the correct value. Some publishers pre-trained the model with two step pre-training, and we

show those as Ph1 parameter and Ph2 parameter if there is differences.

Table 5: Basic pre-training information of BERT models

Models	Model Size	Whole Word Mask-ing	Vocabulary Size	max_ seq_ length
Kyoto Univ. Ver.	Base	No	32,000	128
MeCab Ver.	Base	No	32,000	No Info
SP Ver.	Base	No	32,000	No Info
Tohoku Uinv. Ver.	Base	No	32,000	512
NICT Ver.	Base	No	32,000	Ph1-128 Ph2-512
Laboro Ver.	Base	No	32,000	Ph1-128 Ph2-512

Music and speech are distinct in lexical tone normalization processing

Ran Tao
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
rantao@polyu.edu.hk

Gang Peng
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
gpeng@polyu.edu.hk

Abstract

This paper presents the results of a Cantonese lexical tone normalization experiment. Two non-linguistic (i.e., music and nonspeech) context conditions were used in addition to a speech context condition to elicit listeners' normalization of level tones following contexts with fundamental frequency alterations. Participants showed clear tone normalization in the speech context condition. Participants' normalization performances in the non-linguistic conditions were not significant but comparable to each other. The findings indicated that the non-linguistic music context is not sufficient to support lexical tone normalization.

1 Introduction

In daily life, people speak with huge inter- and intra-talker variance (Peng et al., 2012). The task of normalizing variable vocal streams to identical linguistic codes seems challenging. When it comes to the tone languages such as Cantonese, the situation is even more complicated (Wong and Diehl, 2003). Cantonese has three level tones conveying distinct meanings when superimposing to identical syllables, leaving the pitch height the only clue for listeners' successful judgments. The level tones' pitch heights frequently overlap with each other among speakers and within a speaker, such as in different emotional states. However, the task of identifying words in a vocal stream, i.e., lexical tone normalization, can be accomplished by typical Cantonese without conscious awareness (Wong and Diehl, 2003; Francis et al., 2006; Zhang et al., 2016; Zhang et al., 2017).

Research on lexical tone normalization has revealed the significant role of immediate context which provides a reference to normalize the target pitch height. Researchers debated on whether such processes depend on a general-purpose auditory mechanism or a speech-specific mechanism. There are contradictory findings in the literature, providing evidence to support both sides. Huang and Holt (2009; 2011) found both nonspeech and speech context is sufficient to facilitate the listener's tone normalization, with a restricted effect in the nonspeech context. However, other researchers failed to replicate the facilitation effect of nonspeech context, and mainly support the speech specific mechanism (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2016; Zhang et al., 2017).

While the nonspeech context can replicate the pitch pattern in speech context, they differ in aspects other than whether or not containing a linguistic meaning. The nonspeech context is novel to listeners because such stimuli are rare in daily lives. It is plausible that listeners cannot take advantage of the immediate context of nonspeech as they are not familiar with such auditory patterns.

Music, as non-linguistic stimuli, frequently appears in daily lives. More interestingly, previous research suggested that music could benefit linguistic abilities. Nan et al., (2018) found that piano training can enhance the neural processing of pitch and thus improves speech perception in Mandarin-speaking children. Other researchers have found that music training can benefit the lexical tone perception of

non-tone language speakers (Wayland et al., 2010), suggesting that lexical tone and music may share a similar processing mechanism. However, a closer study on Cantonese speakers failed to find such a facilitating effect (Mok and Zuo, 2012). Thus, it is intriguing whether music can serve as an efficient context for lexical tone normalization. An evident tone normalization following the music context may support the general-purpose auditory mechanism. If the music context does not evoke tone normalization, the result may favor the speech-specific mechanism.

In addition to the investigation into the musical context, it is also interesting to compare the function of different non-linguistic contexts. One limitation in previous studies is that researchers tended to include only one contrastive condition to the speech context condition (Zhang et al., 2013; Zhang et al., 2017), which may limit the interpretation and generalization of the findings.

In this study, we inspect the context's role in native Cantonese speakers' lexical tone normalization by including two non-linguistic conditions to compare with the speech context condition, e.g., a nonspeech context condition and a music context condition. Our primary question is whether music context can suffice the normalization of Cantonese level tones. We are also interested in the profile of the contextual effect between the two non-linguistic contexts.

2 Methodology

We adopted similar stimuli and experiment design as our teams' previous studies. The stimuli preparation and experiment procedure are reported briefly as follows, but see (Zhang et al., 2013; Zhang et al., 2017) for detailed descriptions.

Participants

28 native Cantonese speakers (15 female, mean age = 21.9 yrs, SD = 2.95) were recruited in the current study, all without hearing impairment. Two female participants were left-handed. None of the participants worked as professional musicians. All participants signed written consent before the experiment. Experiment protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

Stimuli

Stimuli of this study consisted of contexts and targets in the four context conditions, e.g., no context, speech context, and two non-linguistic contexts. speech context and all targets were produced by four native Cantonese speakers, who were a female speaker with a high pitch range, a female speaker with a low pitch range, a male speaker with a high pitch range, and a male speaker with a low pitch range (coded as FH, FL, MH, and ML respectively in the following text). Speech context was a four-syllable meaningful sentence, i.e., 呢個字係 (/li55 ko33 tsi22 hɐi22/, "This word is meaning"). After recording the natural production of the sentence from the four talkers, the F0 trajectories of the sentences were then lowered and raised three semitones. In sum, three kinds of speech contexts were formed: an F0 lowered context, an F0 unshifted context, and an F0 raised context. All targets from three context conditions were the natural production of the Chinese character 意 (e.g., /ji33/ mid-level tone, "meaning").

The nonspeech contexts were produced by applying the F0 trajectory and intensity profile from speech contexts to triangle waves. The music contexts were piano notes that had the closest pitch height to each of the syllables in the speech context. All the targets were adjusted to 55dB in intensity and 450 ms in duration. All speech contexts were adjusted to 55 dB in intensity. The non-linguistic contexts were adjusted to 65dB in intensity to match the hearing loudness of speech contexts. The duration of non-linguistic contexts were the same as their corresponding speech contexts.

There were also fillers in the tasks. In the speech context condition, the filling context was a four-syllable sentence, i.e., 我以家讀 (/ŋɔ23 ji21 ka55 tuk2/, "Now I will read"). All the targets were Chinese characters 醫 (e.g., /ji55/ high level tone, "a doctor") or 意. The nonspeech contexts and music contexts of the fillers were produced with the same procedure above.

Experiment Procedure

Participants attended two practice blocks and four experiment blocks. The four experiment blocks con-

sisted of four context conditions respectively, e.g., the no context condition, the speech context condition, the nonspeech context condition, and the music context condition. In the no context condition, the participants heard the targets without preceding contexts. The no context condition is coded as isolated in the following text.

The task was a word identification task that asked participants to make a judgment on the target syllable after listening to the preceding context attentively. In each experiment trial, participants first heard a context, and after a jittering silence (range: 300 - 500 ms), a target syllable was presented. In the isolated condition, Participants heard the target without a context. Participants then made a judgment on the target syllable, whether it is 醫, 意, or 二 (e.g., /ji22/low-level tone, "two") by pressing corresponding keys on the keyboard when they saw a cue on the screen. The cue was presented 800 ms after the onset of the target. Such a manipulation of response time widow was because the experiment was part of a large project in which participants were tested with EEG recording. This restriction minimized the artifacts of EEG signal due to muscle movement. In this kind of setting, reaction times were not a meaningful index of participants' psycholinguistic properties and thus not analyzed in this study. We focused on the judgments of the targets from the participants, which was also the standard procedure in previous research.

The isolated condition consisted of 16 repetitions of each target. The three context conditions each consisted of nine repetitions of three F0 shifts of four talkers, making 27 trials for each talker in each context condition. The four experiment blocks were counterbalanced to prevent order effects.

Analysis

Following previous research (Wong and Diehl, 2003; Zhang et al., 2012; Zhang et al., 2017), perceptual height (PH) and identification rate (IR) were analyzed to investigate participants' lexical tone normalization performance. For the PH analysis, a response of high-level tone was coded as 6, middle-level tone as 3, and low-level tone as 1. The mean Perceptual Height close to 6 indicated that participants generally perceived the targets as high-level tones. In a lowered F0 condition, this could serve as an evidence of evoking participants' tone normalization. Perceptual height close to 1 indicated that participants generally perceived the targets as low-level tone. In a raised F0 condition, this could serve as an evidence of evoking participants' tone normalization. The identification rate was the percentage of expected responses in each condition. The expected responses were the judgments that participants should make when successfully evoked tone normalization, e.g., low-level tone response in the raised F0 condition, middle-level tone response in the unshifted F0 condition, and high-level tone response in the lowered F0 condition.

We conducted one-way repeated measures ANOVAs on PH and IR, with Context as the main factor. Then, we conducted three-way repeated measures ANOVAs on PH and IR. The isolated condition was excluded from this analysis because it did not match the design matrix of other context conditions, e.g., there was no context and thus no F0 Shift manipulations. Three main factors were Context (music, nonspeech, speech), F0 shift (lowered, unshifted, raised), and talker (FH, FL, MH, ML). Greenhouse–Geisser correction was applied when the data violated the Sphericity hypothesis. Tukey method for comparing families of multiple estimates were applied for necessary post-hoc analysis.

3 Results

Perceptual Height

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(2.78, 75.18) = 4.91, p = 0.004, ges = 0.054$), indicating that the PH is influenced by the targets' preceding contexts (see Figure 1). Post-hoc analysis revealed that the speech context condition (3.23) had higher PH than nonspeech (2.81) and music (2.80) conditions ($ps < 0.01$), but not isolated (2.94) condition ($p = 0.117$). All other comparisons were not significant. Next, we report the results of the three-way repeated measures ANOVA on PH.

There was a significant main effect of Context ($F(1.83, 49.44) = 7.20, p = 0.002, ges = 0.031$), replicating that participants perceived the same set of targets as different lexical tones across three context conditions. Post-hoc analysis showed that the

PH is highest in the speech context (3.23, $ps < 0.01$ when compared with the two non-linguistic contexts) and showed no difference between music context (2.80) and nonspeech context (2.81). The main effect of F0 Shift was also significant ($F(1.13, 30.50) = 162.77, p < 0.001, ges = 0.173$), indicating participants perceived targets as different lexical tones with contexts' F0 manipulated. As expected, the PH was highest in the lowered F0 conditions (3.62) and lowest in the raised F0 conditions (2.34). The PH in the unshifted F0 condition was 2.87 and the comparisons among the three PHs were all significant (all $ps < 0.001$) which indicated participants' tone normalization evoked in general. The main effect of talker was also significant ($F(2.04, 55.14) = 8.80, p < 0.001, ges = 0.100$). Participants perceived targets produced by FH with the highest (3.41) picth height, significantly higher than that of FL (2.35), MH (3.00), and ML (3.01). There was no difference between FL, MH, and ML.

The interaction between Context and Shift was significant ($F(1.33, 35.92) = 142.88, p < 0.001, ges = 0.286$). The large suggests the participants perception of the same set of targets was strongly influenced by the preceding context with different F0 Shift manipulation (see Figure 1). Post-hoc analysis revealed that only speech context elicited successful lexical tone normalization: the PH was significantly different among the three F0 Shift conditions (lowered: 5.21, unshifted: 3.02, raised: 1.46, all $ps < 0.001$). The PHs were not different among the three F0 Shift conditions of the two non-linguistic contexts (PH range: 2.77 - 2.85, all $ps > 0.7$).

The interaction between Context and talker was also significant ($F(2.96, 80.03) = 11.59, p < 0.001, ges = 0.046$), indicating participants' perception of tones was modulated by the talkers in the three Contexts. However, such modulation was not significant in speech contexts. Participants perceived the four talkers as the same PH (all $ps > 0.7$). In the two non-linguistic contexts, the targets produced by FL (music: 1.95, nonspeech: 1.92) were always perceived lowest (all $ps < 0.01$). No other comparisons were significant except that the targets produced by FH (3.46) were perceived higher than that of ML (2.77) in the music context ($p < 0.05$).

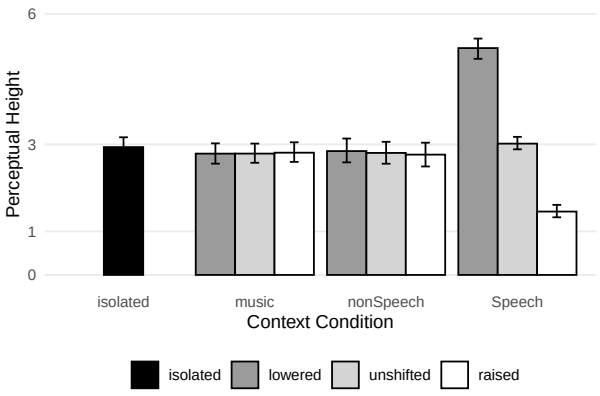


Figure 1: Average perceptual heights (PH) of the word identification task. Black bar respresents the PH in isolated condition. Gray bars represent the PH in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

Identification Rate

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(1.75, 47.37) = 95.47, p < 0.001, ges = 0.705$, see Figure 2). The large effect size indicated that participants' IR is strongly influenced by the targets' preceding contexts. Post-hoc analysis revealed that the speech context condition (78.4%) had higher IR than nonspeech (34.1%), music (33.3%) and isolated (45.2%) conditions (all $ps < 0.01$). The isolated condition had higher IR than the two non-linguistic context conditions (all $ps < 0.01$). There was no IR difference between the two non-linguistic conditions. Next, we report the results of the three-way repeated measures ANOVA on IR.

There was a significant main effect of Context ($F(1.10, 29.75) = 135.93, p < 0.001, ges = 0.359$). This replicated that participants' identification rate is strongly influenced by targets' attaching context. The post-hoc analysis also revealed the same result as in the one-way repeat measures ANOVA. The main effect of Shift ($F(1.52, 41.06) = 8.83, p = 0.002, ges = 0.058$) was also significant, indicating participants' performance was different under the three F0 Shift conditions. Specifically, the lowered F0 condition (38.8%) yielded lower IR than unshifted (54.0%) and raised (53.1%) F0 conditions. There was no

IR difference between the unshifted and raised F0 conditions. However, the main effect of talker was not significant ($F(2.84, 76.79) = 1.37, p = 0.259$). Although the PHs was different as percieved from targets produced by the four talkers, this perceptual difference did not influence participants' judgment: their performance was overall the same on the four talkers.

There were two significant two-way interactions. The interaction between Context and Shift was significant ($F(3.08, 83.09) = 8.55, p < 0.001, ges = 0.035$), indicating participants' performance was modulated by F0 shifts in different contexts. Post-hoc analysis showed that there was no difference in IRs among F0 Shifts of speech context (range: $75.7 - 81.7\%$, all comparisons' $ps > 0.79$). However, the patterns of IR were highly similar in music and nonspeech conditions. In both non-linguistic context conditions, lowered F0 context (music: (17.4%), nonspeech: (21.1%)) elicited lower identification rate than unshifted ((45.8%), (40.4%)) and raised ((36.8%), (40.7%)) F0 context (all $ps < 0.001$, see Figure 2).

The interaction between Shift and talker was also significant ($F(3.18, 85.93) = 6.22, p = 0.002, ges = 0.055$), indicating participants' performance was also modulated by F0 Shift in different talkers. Interestingly, for targets produced by the male talkers, the IRs showed no difference among three F0 shifts (range: $40.3 - 54.9\%$, all $ps > 0.1$). For targets produced by female talkers, The IR was lower in the lowered F0 condition compared with unshifted F0 condition (in FL: 27.7% and $50.9\%, p < 0.01$; in FH: 44.8% and $61.6\%, p < 0.05$). In addition, the IRs were higher in the raised F0 condition (64.6%) than lowered F0 condition in FL ($p < 0.001$), and higher in the raised F0 condition (40.2%) than unshifted F0 condition in FH ($p < 0.05$).

Finally, the three-way interaction among Context, Shift and talker was significant ($F(6.74, 182.05) = 6.22, p < 0.001, ges = 0.039$). To decode this interaction, we conducted two-way ANOVAs on the IRs of targets produced by each of the four talkers (see Figure 3). The Context main effect were significant across talkers (all $ps < 0.001$). Post-hoc analysis revealed that IRs of the speech contexts were always highest (all $ps < 0.001$), with no difference between

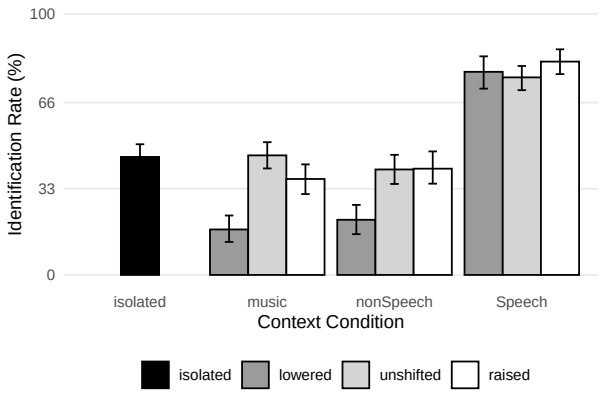


Figure 2: Identification rates (IR) of the word identification task. Black bar respresents the IR in isolated condition. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

the two non-linguistic contexts (all $ps > 0.8$). The Shift main effect was only significant in the analysis of female talkers (all $ps < 0.05$) and the lowered F0 condition always produced the lowest IRs. The interaction between Context and Shift was significant in the analysis of FH, FL, and ML (all $ps < 0.001$). The post-hoc analysis revealed that in music conditions, lowered F0 contexts had consistently low IRs (all $ps < 0.01$), while in nonspeech conditions, lowered F0 context had low IRs only in the analysis of FL ($ps < 0.001$).

4 Discussion

In this study, we included two non-linguistic context conditions in addition to a speech context condition to investigate native Cantonese's lexical tone normalization. Specifically, we were interested in participants' performance in the music context condition. The results revealed successful tone normalization only in speech contexts, suggesting that the music and language abilities were not mutually transferrable in the scenario of lexical tone normalization. Besides, the two non-linguistic conditions' performance were comparable. In general, the finding favors the speech-specific mechanism in lexcial tone normalization.

Our results on the perceptual heights and identification rates under different conditions largely

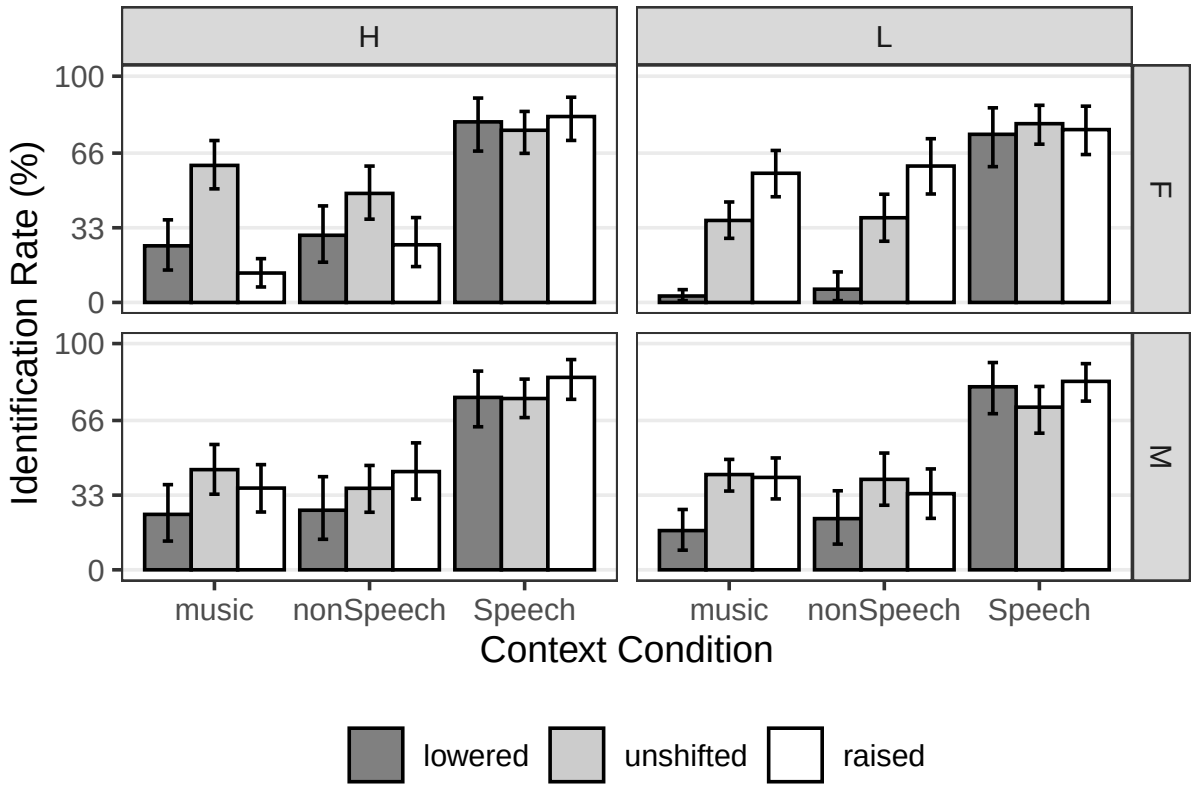


Figure 3: Identification rates (IR) of the targets produced by four talkers in the word identification task. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of three Context conditions. Error bars represent 95% confidence intervals. Rows of the panels are divided by the talkers gender: Femal and Male; columns of the panels are divided by the talkers' pitch range: High and Low.

replicated our previous studies (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2017). In accordance with our previous research, participants only showed evident lexical tone normalization effect under the speech context conditions. The perceptual height under music and nonspeech conditions were close to 3, indicating that participants perceived the middle-level tone as it was in these non-linguistic conditions.

The participants' perceived perceptual heights were influenced by the talker who produced the targets under the non-linguistic conditions. In both music and nonspeech context conditions, targets produced by FL was lowest. Targets produced by FH were perceived highest, however, the perceptual height differences between targets produced by FH and two male talkers were not significant. In contrast, the identification rates among the targets produced by four talkers were not different from each other. This suggests that the perceived height does not influence the ability to correctly normalize lexical tones.

Interestingly, the identification rate was significantly lower in the lowered F0 conditions, compared with unshifted and raised F0 conditions. Such a pattern was consistent across the non-linguistic contexts but not evident in the speech context condition. In both music and nonspeech context conditions, the identification rates in the lowered F0 contexts were significantly lower than unshifted and raised F0 contexts, while the identification rate under unshifted and raised F0 contexts showed no difference.

Here we offer two possible explanations. Firstly, the normalization of level tones under the lowered F0 context conditions was probably harder than that of unshifted and raised. The pitch difference

between high-level tone and middle-level tone is greater than that between low-level tone and middle-level tone (Wong and Diehl, 2003). To elicit successful tone normalization, the F0 shift has a higher requirement for high-level tones, i.e., in the lowered F0 condition. In our manipulation, the lowered and raised F0 condition both altered 3 semitones. While this manipulation was sufficient to elicit tone normalization in speech contexts, it might bring bias in the non-linguistics context, resulting in a unbalanced IR in the three F0 alternation conditions. A future study with an unbalanced alteration (larger F0 shift in lowered F0 condition) may explicitly examine this explanation.

However, such a bias was not consistent across talkers: the Shift main effect was only significant in the analysis of female talkers. Thus, a second explanation is that the low IR in the lowered F0 condition was influenced by the response tendency driven by the natural pitch range of different talkers. Indeed, the interaction between Context and Shift had variable patterns among the analysis of four talkers.

Both of the two explanations seemed hard to be compatible with the speech-specific mechanism which would predict a null effect of F0 Shift in non-linguisitic conditions. In future studies, a closer examination on the low IR in the lowered F0 Shift conditions of non-linguistic contexts will be meaningful to unravel the possible role of general auditory mechanism in lexcial tone normalization. For example, the F0 alterations can be exagerated to increase the effect of non-linguistic contexts.

5 Conclusion

We revisited the lexical tone normalization process by examining the possible function of music as an untested non-linguistic context. Participants successfully normalized target lexical tones following the speech context but not the music or nonspeech context. The behaivoral pattern in the two non-linguistic conditions were similar. Our findings mainly supported the idea that lexical tone normalization relies on the speech-specific mechanism.

Acknowledgments

This work was partially supported by the General Research Fund (No. 15607518) from Research Research Grants Council (RGC) of Hong Kong and a PolyU internal fund (PolyU 156002/17H).

References

Alexander L. Francis, Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu. 2006. Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, 119(3):1712–1726.

Jingyuan Huang and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, 125(6):3983–3994.

Jingyuan Huang and Lori L. Holt. 2011. Evidence for the central origin of lexical tone normalization (l). *The Journal of the Acoustical Society of America*, 129(3):1145–1148.

P. K. Peggy Mok and Donghui Zuo. 2012. The separation between music and speech: evidence from the perception of cantonese tones. *The Journal of the Acoustical Society of America*, 132(4):2711–2720.

Yun Nan, Li Liu, Eveline Geiser, Hua Shu, Chen Chen Gong, Qi Dong, John D. E. Gabrieli, and Robert Desimone. 2018. Piano training enhances the neural processing of pitch and improves speech perception in mandarin-speaking children. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28):E6630–E6639.

Gang Peng, Caicai Zhang, Hong-Ying Zheng, James W. Minett, and William S.-Y. Wang. 2012. The effect of intertalker variations on acoustic–perceptual mapping in cantonese and mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2):579–595.

Ratree Wayland, Elizabeth Herrera, and Edith Kaan. 2010. Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, 38(4):654–662.

Patrick C. M. Wong and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2):413–421.

Caicai Zhang, Gang Peng, and William S-Y Wang. 2012. Unequal effects of speech and nonspeech contexts on the perceptual normalization of cantonese level tones. *The Journal of the Acoustical Society of America*, 132(2):1088–1099.

Caicai Zhang, Gang Peng, and William S-Y Wang. 2013. Achieving constancy in spoken word identification: time course of talker normalization. *Brain and language*, 126(2):193–202.

Caicai Zhang, Kenneth R. Pugh, W. Einar Mencl, Peter J. Molfese, Stephen J. Frost, James S. Magnuson, Gang Peng, and William S-Y Wang. 2016. Functionally integrated neural processing of linguistic and talker information: An event-related fmri and erp study. *NeuroImage*, 124(Pt A):536–549.

Kaile Zhang, Xiao Wang, and Gang Peng. 2017. Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism. *The Journal of the Acoustical Society of America*, 141(1):38.

Music and speech are distinct in lexical tone normalization processing

Ran Tao
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
rantao@polyu.edu.hk

Gang Peng
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
gpeng@polyu.edu.hk

Abstract

This paper presents the results of a Cantonese lexical tone normalization experiment. Two non-linguistic (i.e., music and nonspeech) context conditions were used in addition to a speech context condition to elicit listeners' normalization of level tones following contexts with fundamental frequency alterations. Participants showed clear tone normalization in the speech context condition. Participants' normalization performances in the non-linguistic conditions were not significant but comparable to each other. The findings indicated that the non-linguistic music context is not sufficient to support lexical tone normalization.

1 Introduction

In daily life, people speak with huge inter- and intra-talker variance (Peng et al., 2012). The task of normalizing variable vocal streams to identical linguistic codes seems challenging. When it comes to the tone languages such as Cantonese, the situation is even more complicated (Wong and Diehl, 2003). Cantonese has three level tones conveying distinct meanings when superimposing to identical syllables, leaving the pitch height the only clue for listeners' successful judgments. The level tones' pitch heights frequently overlap with each other among speakers and within a speaker, such as in different emotional states. However, the task of identifying words in a vocal stream, i.e., lexical tone normalization, can be accomplished by typical Cantonese without conscious awareness (Wong and Diehl, 2003; Francis et al., 2006; Zhang et al., 2016; Zhang et al., 2017).

Research on lexical tone normalization has revealed the significant role of immediate context which provides a reference to normalize the target pitch height. Researchers debated on whether such processes depend on a general-purpose auditory mechanism or a speech-specific mechanism. There are contradictory findings in the literature, providing evidence to support both sides. Huang and Holt (2009; 2011) found both nonspeech and speech context is sufficient to facilitate the listener's tone normalization, with a restricted effect in the nonspeech context. However, other researchers failed to replicate the facilitation effect of nonspeech context, and mainly support the speech specific mechanism (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2016; Zhang et al., 2017).

While the nonspeech context can replicate the pitch pattern in speech context, they differ in aspects other than whether or not containing a linguistic meaning. The nonspeech context is novel to listeners because such stimuli are rare in daily lives. It is plausible that listeners cannot take advantage of the immediate context of nonspeech as they are not familiar with such auditory patterns.

Music, as non-linguistic stimuli, frequently appears in daily lives. More interestingly, previous research suggested that music could benefit linguistic abilities. Nan et al., (2018) found that piano training can enhance the neural processing of pitch and thus improves speech perception in Mandarin-speaking children. Other researchers have found that music training can benefit the lexical tone perception of

non-tone language speakers (Wayland et al., 2010), suggesting that lexical tone and music may share a similar processing mechanism. However, a closer study on Cantonese speakers failed to find such a facilitating effect (Mok and Zuo, 2012). Thus, it is intriguing whether music can serve as an efficient context for lexical tone normalization. An evident tone normalization following the music context may support the general-purpose auditory mechanism. If the music context does not evoke tone normalization, the result may favor the speech-specific mechanism.

In addition to the investigation into the musical context, it is also interesting to compare the function of different non-linguistic contexts. One limitation in previous studies is that researchers tended to include only one contrastive condition to the speech context condition (Zhang et al., 2013; Zhang et al., 2017), which may limit the interpretation and generalization of the findings.

In this study, we inspect the context's role in native Cantonese speakers' lexical tone normalization by including two non-linguistic conditions to compare with the speech context condition, e.g., a nonspeech context condition and a music context condition. Our primary question is whether music context can suffice the normalization of Cantonese level tones. We are also interested in the profile of the contextual effect between the two non-linguistic contexts.

2 Methodology

We adopted similar stimuli and experiment design as our teams' previous studies. The stimuli preparation and experiment procedure are reported briefly as follows, but see (Zhang et al., 2013; Zhang et al., 2017) for detailed descriptions.

Participants

28 native Cantonese speakers (15 female, mean age = 21.9 yrs, SD = 2.95) were recruited in the current study, all without hearing impairment. Two female participants were left-handed. None of the participants worked as professional musicians. All participants signed written consent before the experiment. Experiment protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

Stimuli

Stimuli of this study consisted of contexts and targets in the four context conditions, e.g., no context, speech context, and two non-linguistic contexts. speech context and all targets were produced by four native Cantonese speakers, who were a female speaker with a high pitch range, a female speaker with a low pitch range, a male speaker with a high pitch range, and a male speaker with a low pitch range (coded as FH, FL, MH, and ML respectively in the following text). Speech context was a four-syllable meaningful sentence, i.e., 呢個字係 (/li55 ko33 tsi22 hɐi22/, "This word is meaning"). After recording the natural production of the sentence from the four talkers, the F0 trajectories of the sentences were then lowered and raised three semitones. In sum, three kinds of speech contexts were formed: an F0 lowered context, an F0 unshifted context, and an F0 raised context. All targets from three context conditions were the natural production of the Chinese character 意 (e.g., /ji33/ mid-level tone, "meaning").

The nonspeech contexts were produced by applying the F0 trajectory and intensity profile from speech contexts to triangle waves. The music contexts were piano notes that had the closest pitch height to each of the syllables in the speech context. All the targets were adjusted to 55dB in intensity and 450 ms in duration. All speech contexts were adjusted to 55 dB in intensity. The non-linguistic contexts were adjusted to 65dB in intensity to match the hearing loudness of speech contexts. The duration of non-linguistic contexts were the same as their corresponding speech contexts.

There were also fillers in the tasks. In the speech context condition, the filling context was a four-syllable sentence, i.e., 我以家讀 (/ŋɔ23 ji21 ka55 tuk2/, "Now I will read"). All the targets were Chinese characters 醫 (e.g., /ji55/ high level tone, "a doctor") or 意. The nonspeech contexts and music contexts of the fillers were produced with the same procedure above.

Experiment Procedure

Participants attended two practice blocks and four experiment blocks. The four experiment blocks con-

sisted of four context conditions respectively, e.g., the no context condition, the speech context condition, the nonspeech context condition, and the music context condition. In the no context condition, the participants heard the targets without preceding contexts. The no context condition is coded as isolated in the following text.

The task was a word identification task that asked participants to make a judgment on the target syllable after listening to the preceding context attentively. In each experiment trial, participants first heard a context, and after a jittering silence (range: 300 - 500 ms), a target syllable was presented. In the isolated condition, Participants heard the target without a context. Participants then made a judgment on the target syllable, whether it is 醫, 意, or 二 (e.g., /ji22/low-level tone, "two") by pressing corresponding keys on the keyboard when they saw a cue on the screen. The cue was presented 800 ms after the onset of the target. Such a manipulation of response time widow was because the experiment was part of a large project in which participants were tested with EEG recording. This restriction minimized the artifacts of EEG signal due to muscle movement. In this kind of setting, reaction times were not a meaningful index of participants' psycholinguistic properties and thus not analyzed in this study. We focused on the judgments of the targets from the participants, which was also the standard procedure in previous research.

The isolated condition consisted of 16 repetitions of each target. The three context conditions each consisted of nine repetitions of three F0 shifts of four talkers, making 27 trials for each talker in each context condition. The four experiment blocks were counterbalanced to prevent order effects.

Analysis

Following previous research (Wong and Diehl, 2003; Zhang et al., 2012; Zhang et al., 2017), perceptual height (PH) and identification rate (IR) were analyzed to investigate participants' lexical tone normalization performance. For the PH analysis, a response of high-level tone was coded as 6, middle-level tone as 3, and low-level tone as 1. The mean Perceptual Height close to 6 indicated that participants generally perceived the targets as high-level tones. In a lowered F0 condition, this could serve as an evidence of evoking participants' tone normalization. Perceptual height close to 1 indicated that participants generally perceived the targets as low-level tone. In a raised F0 condition, this could serve as an evidence of evoking participants' tone normalization. The identification rate was the percentage of expected responses in each condition. The expected responses were the judgments that participants should make when successfully evoked tone normalization, e.g., low-level tone response in the raised F0 condition, middle-level tone response in the unshifted F0 condition, and high-level tone response in the lowered F0 condition.

We conducted one-way repeated measures ANOVAs on PH and IR, with Context as the main factor. Then, we conducted three-way repeated measures ANOVAs on PH and IR. The isolated condition was excluded from this analysis because it did not match the design matrix of other context conditions, e.g., there was no context and thus no F0 Shift manipulations. Three main factors were Context (music, nonspeech, speech), F0 shift (lowered, unshifted, raised), and talker (FH, FL, MH, ML). Greenhouse–Geisser correction was applied when the data violated the Sphericity hypothesis. Tukey method for comparing families of multiple estimates were applied for necessary post-hoc analysis.

3 Results

Perceptual Height

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(2.78, 75.18) = 4.91, p = 0.004, ges = 0.054$), indicating that the PH is influenced by the targets' preceding contexts (see Figure 1). Post-hoc analysis revealed that the speech context condition (3.23) had higher PH than nonspeech (2.81) and music (2.80) conditions ($ps < 0.01$), but not isolated (2.94) condition ($p = 0.117$). All other comparisons were not significant. Next, we report the results of the three-way repeated measures ANOVA on PH.

There was a significant main effect of Context ($F(1.83, 49.44) = 7.20, p = 0.002, ges = 0.031$), replicating that participants perceived the same set of targets as different lexical tones across three context conditions. Post-hoc analysis showed that the

PH is highest in the speech context (3.23, $ps < 0.01$ when compared with the two non-linguistic contexts) and showed no difference between music context (2.80) and nonspeech context (2.81). The main effect of F0 Shift was also significant ($F(1.13, 30.50) = 162.77, p < 0.001, ges = 0.173$), indicating participants perceived targets as different lexical tones with contexts' F0 manipulated. As expected, the PH was highest in the lowered F0 conditions (3.62) and lowest in the raised F0 conditions (2.34). The PH in the unshifted F0 condition was 2.87 and the comparisons among the three PHs were all significant (all $ps < 0.001$) which indicated participants' tone normalization evoked in general. The main effect of talker was also significant ($F(2.04, 55.14) = 8.80, p < 0.001, ges = 0.100$). Participants perceived targets produced by FH with the highest (3.41) picth height, significantly higher than that of FL (2.35), MH (3.00), and ML (3.01). There was no difference between FL, MH, and ML.

The interaction between Context and Shift was significant ($F(1.33, 35.92) = 142.88, p < 0.001, ges = 0.286$). The large suggests the participants perception of the same set of targets was strongly influenced by the preceding context with different F0 Shift manipulation (see Figure 1). Post-hoc analysis revealed that only speech context elicited successful lexical tone normalization: the PH was significantly different among the three F0 Shift conditions (lowered: 5.21, unshifted: 3.02, raised: 1.46, all $ps < 0.001$). The PHs were not different among the three F0 Shift conditions of the two non-linguistic contexts (PH range: 2.77 - 2.85, all $ps > 0.7$).

The interaction between Context and talker was also significant ($F(2.96, 80.03) = 11.59, p < 0.001, ges = 0.046$), indicating participants' perception of tones was modulated by the talkers in the three Contexts. However, such modulation was not significant in speech contexts. Participants perceived the four talkers as the same PH (all $ps > 0.7$). In the two non-linguistic contexts, the targets produced by FL (music: 1.95, nonspeech: 1.92) were always perceived lowest (all $ps < 0.01$). No other comparisons were significant except that the targets produced by FH (3.46) were perceived higher than that of ML (2.77) in the music context ($p < 0.05$).

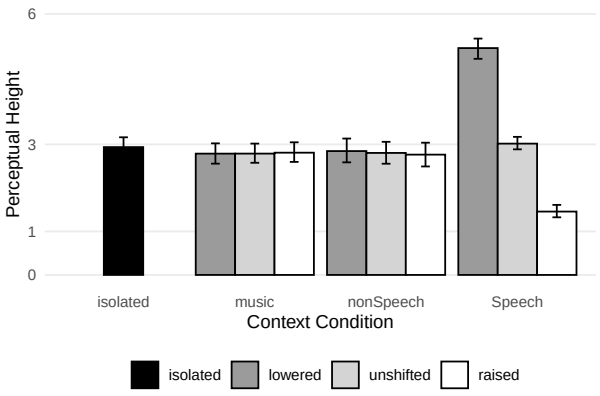


Figure 1: Average perceptual heights (PH) of the word identification task. Black bar respresents the PH in isolated condition. Gray bars represent the PH in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

Identification Rate

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(1.75, 47.37) = 95.47, p < 0.001, ges = 0.705$, see Figure 2). The large effect size indicated that participants' IR is strongly influenced by the targets' preceding contexts. Post-hoc analysis revealed that the speech context condition (78.4%) had higher IR than nonspeech (34.1%), music (33.3%) and isolated (45.2%) conditions (all $ps < 0.01$). The isolated condition had higher IR than the two non-linguistic context conditions (all $ps < 0.01$). There was no IR difference between the two non-linguistic conditions. Next, we report the results of the three-way repeated measures ANOVA on IR.

There was a significant main effect of Context ($F(1.10, 29.75) = 135.93, p < 0.001, ges = 0.359$). This replicated that participants' identification rate is strongly influenced by targets' attaching context. The post-hoc analysis also revealed the same result as in the one-way repeat measures ANOVA. The main effect of Shift ($F(1.52, 41.06) = 8.83, p = 0.002, ges = 0.058$) was also significant, indicating participants' performance was different under the three F0 Shift conditions. Specifically, the lowered F0 condition (38.8%) yielded lower IR than unshifted (54.0%) and raised (53.1%) F0 conditions. There was no

IR difference between the unshifted and raised F0 conditions. However, the main effect of talker was not significant ($F(2.84, 76.79) = 1.37, p = 0.259$). Although the PHs was different as percieved from targets produced by the four talkers, this perceptual difference did not influence participants' judgment: their performance was overall the same on the four talkers.

There were two significant two-way interactions. The interaction between Context and Shift was significant ($F(3.08, 83.09) = 8.55, p < 0.001, ges = 0.035$), indicating participants' performance was modulated by F0 shifts in different contexts. Post-hoc analysis showed that there was no difference in IRs among F0 Shifts of speech context (range: $75.7 - 81.7\%$, all comparisons' $ps > 0.79$). However, the patterns of IR were highly similar in music and nonspeech conditions. In both non-linguistic context conditions, lowered F0 context (music: (17.4%), nonspeech: (21.1%)) elicited lower identification rate than unshifted ((45.8%), (40.4%)) and raised ((36.8%), (40.7%)) F0 context (all $ps < 0.001$, see Figure 2).

The interaction between Shift and talker was also significant ($F(3.18, 85.93) = 6.22, p = 0.002, ges = 0.055$), indicating participants' performance was also modulated by F0 Shift in different talkers. Interestingly, for targets produced by the male talkers, the IRs showed no difference among three F0 shifts (range: $40.3 - 54.9\%$, all $ps > 0.1$). For targets produced by female talkers, The IR was lower in the lowered F0 condition compared with unshifted F0 condition (in FL: 27.7% and $50.9\%, p < 0.01$; in FH: 44.8% and $61.6\%, p < 0.05$). In addition, the IRs were higher in the raised F0 condition (64.6%) than lowered F0 condition in FL ($p < 0.001$), and higher in the raised F0 condition (40.2%) than unshifted F0 condition in FH ($p < 0.05$).

Finally, the three-way interaction among Context, Shift and talker was significant ($F(6.74, 182.05) = 6.22, p < 0.001, ges = 0.039$). To decode this interaction, we conducted two-way ANOVAs on the IRs of targets produced by each of the four talkers (see Figure 3). The Context main effect were significant across talkers (all $ps < 0.001$). Post-hoc analysis revealed that IRs of the speech contexts were always highest (all $ps < 0.001$), with no difference between

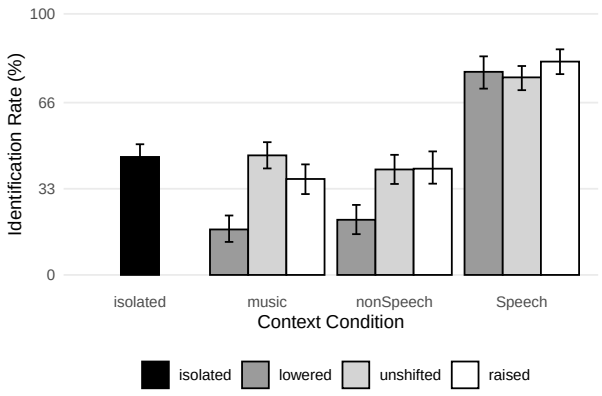


Figure 2: Identification rates (IR) of the word identification task. Black bar respresents the IR in isolated condition. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

the two non-linguistic contexts (all $ps > 0.8$). The Shift main effect was only significant in the analysis of female talkers (all $ps < 0.05$) and the lowered F0 condition always produced the lowest IRs. The interaction between Context and Shift was significant in the analysis of FH, FL, and ML (all $ps < 0.001$). The post-hoc analysis revealed that in music conditions, lowered F0 contexts had consistently low IRs (all $ps < 0.01$), while in nonspeech conditions, lowered F0 context had low IRs only in the analysis of FL ($ps < 0.001$).

4 Discussion

In this study, we included two non-linguistic context conditions in addition to a speech context condition to investigate native Cantonese's lexical tone normalization. Specifically, we were interested in participants' performance in the music context condition. The results revealed successful tone normalization only in speech contexts, suggesting that the music and language abilities were not mutually transferrable in the scenario of lexical tone normalization. Besides, the two non-linguistic conditions' performance were comparable. In general, the finding favors the speech-specific mechanism in lexcial tone normalization.

Our results on the perceptual heights and identification rates under different conditions largely

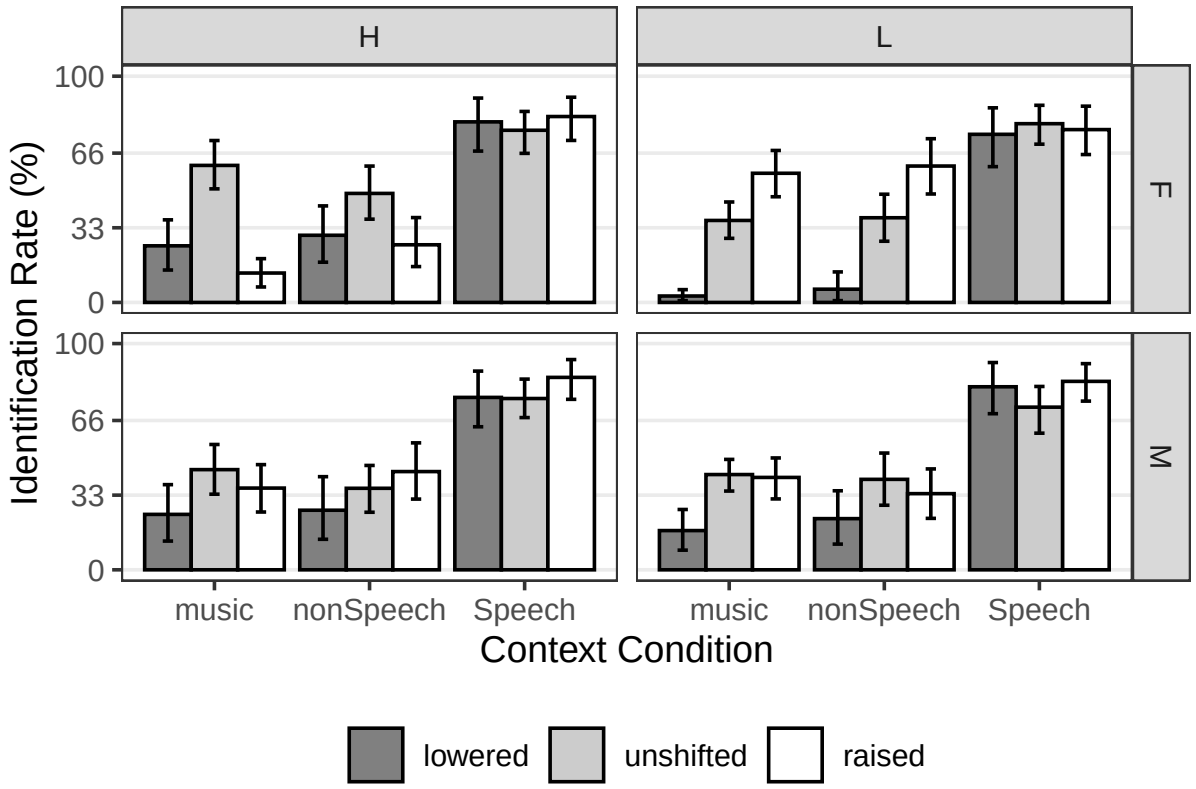


Figure 3: Identification rates (IR) of the targets produced by four talkers in the word identification task. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of three Context conditions. Error bars represent 95% confidence intervals. Rows of the panels are divided by the talkers gender: Femal and Male; columns of the panels are divided by the talkers' pitch range: High and Low.

replicated our previous studies (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2017). In accordance with our previous research, participants only showed evident lexical tone normalization effect under the speech context conditions. The perceptual height under music and nonspeech conditions were close to 3, indicating that participants perceived the middle-level tone as it was in these non-linguistic conditions.

The participants' perceived perceptual heights were influenced by the talker who produced the targets under the non-linguistic conditions. In both music and nonspeech context conditions, targets produced by FL was lowest. Targets produced by FH were perceived highest, however, the perceptual height differences between targets produced by FH and two male talkers were not significant. In contrast, the identification rates among the targets produced by four talkers were not different from each other. This suggests that the perceived height does not influence the ability to correctly normalize lexical tones.

Interestingly, the identification rate was significantly lower in the lowered F0 conditions, compared with unshifted and raised F0 conditions. Such a pattern was consistent across the non-linguistic contexts but not evident in the speech context condition. In both music and nonspeech context conditions, the identification rates in the lowered F0 contexts were significantly lower than unshifted and raised F0 contexts, while the identification rate under unshifted and raised F0 contexts showed no difference.

Here we offer two possible explanations. Firstly, the normalization of level tones under the lowered F0 context conditions was probably harder than that of unshifted and raised. The pitch difference

between high-level tone and middle-level tone is greater than that between low-level tone and middle-level tone (Wong and Diehl, 2003). To elicit successful tone normalization, the F0 shift has a higher requirement for high-level tones, i.e., in the lowered F0 condition. In our manipulation, the lowered and raised F0 condition both altered 3 semitones. While this manipulation was sufficient to elicit tone normalization in speech contexts, it might bring bias in the non-linguistics context, resulting in a unbalanced IR in the three F0 alternation conditions. A future study with an unbalanced alteration (larger F0 shift in lowered F0 condition) may explicitly examine this explanation.

However, such a bias was not consistent across talkers: the Shift main effect was only significant in the analysis of female talkers. Thus, a second explanation is that the low IR in the lowered F0 condition was influenced by the response tendency driven by the natural pitch range of different talkers. Indeed, the interaction between Context and Shift had variable patterns among the analysis of four talkers.

Both of the two explanations seemed hard to be compatible with the speech-specific mechanism which would predict a null effect of F0 Shift in non-linguisitic conditions. In future studies, a closer examination on the low IR in the lowered F0 Shift conditions of non-linguistic contexts will be meaningful to unravel the possible role of general auditory mechanism in lexcial tone normalization. For example, the F0 alterations can be exagerated to increase the effect of non-linguistic contexts.

5 Conclusion

We revisited the lexical tone normalization process by examining the possible function of music as an untested non-linguistic context. Participants successfully normalized target lexical tones following the speech context but not the music or nonspeech context. The behaivoral pattern in the two non-linguistic conditions were similar. Our findings mainly supported the idea that lexical tone normalization relies on the speech-specific mechanism.

Acknowledgments

This work was partially supported by the General Research Fund (No. 15607518) from Research Grants Council (RGC) of Hong Kong and a PolyU internal fund (PolyU 156002/17H).

References

Alexander L. Francis, Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu. 2006. Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, 119(3):1712–1726.

Jingyuan Huang and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, 125(6):3983–3994.

Jingyuan Huang and Lori L. Holt. 2011. Evidence for the central origin of lexical tone normalization (l). *The Journal of the Acoustical Society of America*, 129(3):1145–1148.

P. K. Peggy Mok and Donghui Zuo. 2012. The separation between music and speech: evidence from the perception of cantonese tones. *The Journal of the Acoustical Society of America*, 132(4):2711–2720.

Yun Nan, Li Liu, Eveline Geiser, Hua Shu, Chen Chen Gong, Qi Dong, John D. E. Gabrieli, and Robert Desimone. 2018. Piano training enhances the neural processing of pitch and improves speech perception in mandarin-speaking children. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28):E6630–E6639.

Gang Peng, Caicai Zhang, Hong-Ying Zheng, James W. Minett, and William S.-Y. Wang. 2012. The effect of intertalker variations on acoustic–perceptual mapping in cantonese and mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2):579–595.

Ratree Wayland, Elizabeth Herrera, and Edith Kaan. 2010. Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, 38(4):654–662.

Patrick C. M. Wong and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2):413–421.

Caicai Zhang, Gang Peng, and William S-Y Wang. 2012. Unequal effects of speech and nonspeech contexts on the perceptual normalization of cantonese level tones. *The Journal of the Acoustical Society of America*, 132(2):1088–1099.

Caicai Zhang, Gang Peng, and William S-Y Wang. 2013. Achieving constancy in spoken word identification: time course of talker normalization. *Brain and language*, 126(2):193–202.

Caicai Zhang, Kenneth R. Pugh, W. Einar Mencl, Peter J. Molfese, Stephen J. Frost, James S. Magnuson, Gang Peng, and William S-Y Wang. 2016. Functionally integrated neural processing of linguistic and talker information: An event-related fmri and erp study. *NeuroImage*, 124(Pt A):536–549.

Kaile Zhang, Xiao Wang, and Gang Peng. 2017. Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism. *The Journal of the Acoustical Society of America*, 141(1):38.

A corpus-based comparative study of light verbs in three Chinese speech communities

Benjamin K Tsou
City University of Hong Kong,
The Hong Kong University of Science and Technology / Hong Kong SAR
btsou99@gmail.com

Ka-Fai Yip
Yale University / New Haven,
Connecticut, United States
kafai.yip@yale.edu

Abstract

The universal category of light verbs has drawn considerable interest among researchers on Asian languages in recent years. For Chinese, recent research focused on several light verbs have included multiword expressions, some with parallels in English. This study begins with a common but little studied light verb DA 打 ("hit") in Chinese, based on 22 years of data curated in the LIVAC Pan-Chinese corpus (https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus). The verb DA has differentially evolved from a regular transitive verb involving physical strike action to take on increasingly metaphorical extension and is well on the way to become a light verb. From 2175 lexical entries in LIVAC from 1995-2016, we are able to trace some longitudinal developments of this light verb across three Chinese speech communities (Beijing, Hong Kong, Taiwan) by comparing their differences in metaphorical extension. We also provide in detail how the three communities differ in the usage of this popular light verb, and the broader significance beyond linguistics, with Beijing and Hong Kong showing a higher degree of metaphorical shift in different ways. We also explore some means by which metaphorization may be compared with grammatical change. Moreover, we demonstrate how fruitful and new findings may be obtained with a rigorously curated corpus.

1 Background

Light verbs constitute a universal and unique class of words whose basic substantive meaning has been made opaque (or bleached) so that they could have become function words with little substantive meaning. This universal linguistic class has interested linguists for sometime. In the case of English, Jespersen (1954) and Cattell (1984) have noticed them, such as *have* in *have a look*, *take* in *take a drive* etc. In the case of Chinese, as early as the Song dynasty Ouyang Xiu 歐陽修 (A.D. 1007-1072) had noted them in his book *Guitianlu* 歸田錄. In the 1980s modern linguists such as Lü Shuxiang, who used the term "dummy verbs" 形式動詞 (1999[1980]:294) and Wang Li, who referred to them as "markers for verbs" 動詞的記號 (1985:142) had also began to notice them.[1]

In 1985, the first major study of Chinese light verbs was published by Zhu Dexi. It focused on 6 major light verbs in Chinese (*jinxing* 進行 "proceed", *zuo* 作 "make", *jiayi* 加以 "add", *geiyi* 給以 "provide", *jiyu* 給予 "provide" and *yuyi* 予以 "provide"). This paper was followed by Zhou 1987, Li & Chai 1995, Mao 1997, Yan 1998, Chen 2003, Li 2003 and many others, which provided more descriptive coverage. Then came Diao (2004) who doubled the number of Chinese light verbs to cover 7 more of them (*congshi* 從事 "engage", *zuo* 做 "do", *guo* 搞 "make", *gan* 幹 "do", *nong* 弄 "make", *jia* 加 "add" and *yu* 予 "provide"). His concern was with their grammatical usage in Mainland China. With data drawn from the Sinica Corpus, Wang (2004) adopted a corpus-based approach to study differences in the usage of three light verbs (*zuo* 做 "do", *guo* 搞 "make", *nong* 弄 "make") in Taiwan. In 2014, Lin et al. and Huang et al. initiated comparison between light verb usage

[1] Light verbs are also widely found in other East and Southeast Asian languages, e.g. *suru* "do" in Japanese (Grimshaw & Mester 1988), *ha* "do" in Korean (Chae 1996), *lam* "do" in Vietnamese (Pham 1999), etc.

variations in Mainland and Taiwan Mandarin. They used data from the Annotated Chinese Gigaword Corpus which combines data from both Mainland and Taiwan in 1990-2002 to discuss the differential distribution of five light verbs (*jinxing* 進行 "proceed", *congshi* 從事 "engage", *zuo* 做 "do", *guo* 搞 "make", *jiayi* 加以 "add"). They pointed out that 進行 *jinxing* in Taiwan might take verb-object phrases as complements (e.g *jinxing toupiao* 進行投票 "proceed to voting, (lit.) cast-ticket"), but not in Mainland. The general conclusion as reported in Jiang et al. (2016) was that light verbs in Taiwan were "more transitive" (i.e. more verbal) and less grammaticalized.[2]

It is interesting that one light verb stands out for escaping the attention of most linguists. This is *da* 打 "hit", which had surprisingly interested Ouyang Xiu (1007-1072).[3] Most modern studies have focused on *da*'s diachronic development (J. Zhu 2004, Su 2009, Zhuang 2014 *i.a.*) but seldom on its light verb usage in Modern Chinese. Wang (1985) simply called it "a marker for verbs". Ren (2013) studied how *da* was interpreted specifically in different grammaticalized contexts in Grounding Theory. A detailed empirical study is thus needed.

Da 打 is particularly remarkable in undergoing robust metaphorical extensions from denoting initially a physical strike action (e.g. *dasi* 打死 "beat to death") to a complete light verb (e.g. *daya* 打壓 "suppress, (lit.) hit-press"). A systematic study on *da* could shed light on light verb developments and more generally on the process of metaphorization. Also, regional variations in metaphorization among Chinese communities may be fruitfully studied from a comparative perspective.

This paper focuses on metaphorization of *da* with reference to both latitudinal and longitudinal variations in three Chinese speech communities Beijing (BJ), Hong Kong (HK) and Taiwan (TW). From 2175 lexical entries in LIVAC drawn from 1995-2016, we trace the developments of this light verb across a 22-year long time span and also across three Chinese speech communities (BJ, HK and TW) by initiating comparison of their differential metaphorical extensions. We show in some details how the three communities differ in their usage of this popular light verb, and the broader significance beyond linguistics, with BJ and HK showing a higher degree of metaphorical shift in different ways. We also explore how metaphorical extension might come about and demonstrate the valuable use which can be made of a rigorously cultivated corpus.

The paper is organized as follows. Section 2 introduces the corpus base LIVAC and the methodology for analysis. Section 3 overviews *da*'s literal usage and metaphorical usage with regard to word structures. Section 4 investigates the latitudinal and longitudinal variations among three Chinese speech communities BJ, HK, and TW in metaphorization. Section 5 explores theoretical implications for metaphorization and corpus-based studies on variations, and concludes the paper.

2 Corpus base and methodology

This paper draws on the Pan-Chinese synchronous database LIVAC (http://www.livac.org/). Since 1995, LIVAC has processed and filtered representative media texts from Pan-Chinese communities including Beijing, Guangzhou, Hong Kong, Macau, Shanghai, Shenzhen, Singapore, Taiwan. By 2016, more than 600 million characters of news media texts have been rigorously curated (Tsou & Kwong 2015).

A total of 2175 lexical entries with *da* 打 in word initial (1730) and final (445) positions are found in LIVAC (1995-2016). These form the basis for the current study.[4] After filtering out singleton *da*, nouns and loanword entries, we have 812 remaining entries with *da* in initial position

[2] For a related discussion on "verbalness", see Shen & Zhang (2013). Also see Her et al. (2016) and Tsai (2017) for the light verb developments in Taiwan, and see Jiang (2020) for a recent comprehensive study on the semantics of light verbs.

[3] Ouyang Xiu noted in his book *Guitianlu* several examples of *da* with bleached meaning, e.g. *dachuan* 打船 "to make ships", *dache* 打車 "to make cars", *dayu* 打魚 "to fish", *dashui* 打水 "to get water", *dafan* 打飯 "to buy a meal", *dayiliang* 打衣糧 "to distribute clothes and food to soldiers", *dasan* 打傘 "to hold an umbrella", *danian* 打黏 "to stick papers", *daliang* 打量 "to measure", *dashi* 打試 "to examine vision". While most of them are still in use (e.g. *danian*, *dashi*), some have undergone semantic changes in Modern Chinese, e.g. *dache* means "to take a taxi" instead of manufacturing carts.

[4] Compare with 774 entries in Taiwan's *Guoyu Cidian* 國語辭典 (http://dict.revised.moe.edu.tw/cbdic/), among which 284 overlap with LIVAC and 238 entries in *Xiandai Hanyu Cidian* 現代漢語辭典, among which 198 overlap with LIVAC.

and 354 in final position.[5] They are analyzed into three types in terms of metaphorization in usage:

Type I (Literal): Only literal meaning, e.g. *dasi* 打死 "beat to death, (lit.) hit-die", *daquan* 打拳 "fist boxing, (lit.) hit-fist"

Type III (Metaphorical): Only metaphorical meaning, e.g. *daya* 打壓 "suppress, (lit.) hit-press", *dajia* 打假 "crack down on counterfeit, (lit.) hit-falsehood"

Type II (Incipient Metaphorization): An intermediate type whose words could be used both literally (Type I) and metaphorically (Type III), e.g. *dazao* 打造 "(lit.) fabricate (furniture)" vs. "(metaph.) forge (bright future)", *daci* 打氣 "(lit.) pump air" as in *wei luntai daqi* 為輪胎打氣 "pump air into tires" vs. "(metaph.) cheer on" as in *wei qiuyuan daqi* 為球員打氣 "cheer the players on".

The extent of metaphorization in different speech communities could be represented by the ratio between tokens in metaphorical usage and total usage. A higher ratio indicates a greater degree of metaphorization. To capture metaphorical variations across the three communities, we further analyze Type II entries.

BJ, HK and TW have in total 672 entries in use among them. However, only 151 entries are in common use among them with minimum frequency of two. In Table 1, we can see that the small ratio of common usage indicates great regional diversity where the common denominator is smaller than the peripheral variations, while the reverse would indicate a greater uniformity. This is indicative of the richness of variety in metaphorization in the Chinese language. It is noteworthy that these 151 entries take up 95% of the total usage frequency in the three communities and thus are sufficiently representative.

Distribution	Types (%)	Tokens (%)
LIVAC	812 (100)	318,095 (100)
HK, BJ or TW	672 (83)	179,182 (56)
HK, BJ and TW	151 (19)	170,245 (54)

Table 1. Distribution of *da*-X entries

Among the 151 *common* entries, 48 of these have both literal and metaphorical usage. They take up 60% (102,277) of the *common* usage (170,245).[6] To analyze and compare the relative frequencies of literal vs. metaphorical usage, we extract sample sentences containing them from two time periods: (a) 1995-2000, (b) 2011-2016 and obtain 16955 sentences from our database.[7] We then manually tag each case as either literal usage or metaphorical usage,[8] We can quantify the extent of metaphorization for *each* Type II entry by comparing these two kinds of usage.

An overall Metaphorization Index (MI) may be proposed to provide an objective measure:

$$MI = \frac{C + (B_1 + B_2 \ldots + B_n)}{A + B + C} \times 100\%$$

(where A, B, C = total tokens of Types I, II & III respectively, and B_n = tokens of metaphorical usage of a Type II entry)

Following analysis, the incipient type (II) may be redistributed as either literal (I) or metaphorical usage (III) in a bipolar division among the 16955 sample sentences.

The current approach differs from previous studies in at least five aspects: **(a) Choice of light verbs**. While the previous studies were concerned with 13 light verbs mentioned above such as *jinxing* 進行 and rarely studied *da* 打, this study focuses on 2175 lexical entries of *da*. **(b) Metaphorization**. While the previous studies emphasized grammatical differences of light verbs, e.g. compatibility with aspectual markers, argument structures and eventualities of the complements, etc., the current study, on the other hand, centers on metaphorization, a crucial process in light verb development and in language. One of the defining properties of light verbs is the reduction of substantial meaning as the shift from literal meaning to metaphorical meaning. **(c) Regional variation**. In contrast to the previous studies focused on BJ and TW, this study also includes HK to provide a fuller picture of the Pan-Chinese language situation. **(d) Longitudinal variation**. Huang et al. (2014), as a major corpus-based study, focuses on only synchronic variations. The present study samples sentences through a

[5] Three uses are excluded: (i) singleton *da* 打, (ii) noun uses such as *dahuoji* 打火機 "lighter", as well as proper names, e.g. *Daguling* 打鼓嶺 "Ta Kwu Ling (Hong Kong place name)", (iii) loanwords, e.g. *dabi* 打吡 (from English *derby*).

[6] The literal Type I takes up 12% (20179) and metaphorical Type III takes up 28% (47792).

[7] For each Type II word at each community each year, a maximum of 50 sentences were extracted (i.e. at most 600 sentences for each word).

[8] Every sentence was tagged by two annotators. Sentences with disagreement were adjudicated by a third annotator.

time window of 22 years to explore longitudinal comparison. **(e) Sampling size**. Huang et al. (2014) extracted 2000 sentences from the Chinese Gigaword Corpus in total, 200 sentences for each light verb in each region. This study extracts 16955 sentences in LIVAC, with over 170,000 tokens of 打 *da* from two periods.

3 Metaphorization of *da*

3.1 Literal vs. Metaphorical

The primary meaning of *da* 打 is 'hitting by hands or with an additional instrument', and often refers to intentional actions and involves patients (e.g. *daren* 打人 "to beat someone"). Its meaning is extended to hitting in general (e.g. *damen* 打門 "to score goal", "hit-door" in soccer) and actions involving physical contact (e.g. *dajing* 打井 "to dig a well"), as well as fighting (e.g. *dadou* 打鬥 "to fight"). All these senses involve physical actions though the function of the hand is no longer foregrounded in some cases. We consider them to convey the basic and literal meaning.

Da may also be used metaphorically as in *daqi* 打氣 "hit-air", which could mean cheering on somebody, in addition to literally "pumping air". Metaphor involves an understanding of one conceptual domain in terms of another (Lakoff & Johnson 1980). Thus, metaphorical meaning of *da* could be characterized as the extension from physical actions to other conceptual domains, such emotion and culture, i.e. for physical pumping of air to the injection of emotional support to encouragment. The same extension can also be observed in *daji* 打擊 "(lit.) hit-strike". Literally, *daji* means physical striking (e.g. *daji bangqiu* 打擊棒球 "hit a baseball"), but it could undergo metaphorical extension to mean affecting someone emotionally (e.g. *daji xinqing* 打擊心情 "hit-heart status" "affect one's feeling").

3.2 Word structure

Da may be in initial position (*da*-X) or final position (X-*da*) in compound words. A fundamental difference between them is in their degree of metaphorization. It is noteworthy that most (87%) of the X-*da* cases retain literal meaning among the 354 entries, e.g. *ouda* 毆打 "to beat up, (lit.) beat-hit", *duda* 毒打 "to beat cruelly, (lit.) poison-hit", with rare exception of metaphorical usage (only 13%) like *zhida* 主打 "promote, (lit.) main-hit", *yanda* 嚴打 "to severely crack down, (lit.) severe-hit". By comparison, few *da*-X cases do. Among the 812 *da*-X entries only 31% retain literal meaning (Type I) like *dadou* 打鬥 "to fight, (lit.) hit-fight", *dagu* 打鼓 ("to drum, (lit.) hit-drum"). 45% are fully metaphorized (Type III) such as *daya* 打壓"suppress, (lit.) hit-press", and 24% may be used both literally and metaphorically (Type II), i.e. *daci* 打氣 "(lit.) pump air" and "(metaph.) cheer on".[9] The asymmetry is summarized in Table 2:

	entries	Type I	Type II	Type III
X-*da*	354	**87%**	8%	5%
da-X	812	31%	24%	**45%**

Table 2. Distribution of 3 types by word structure

This asymmetry can also be appreciated from a minimal pair: *jida* 擊打 "(lit.) strike-hit" and *daji* 打擊 "(lit.) hit-strike". *Jida* can only mean literally physical striking, e.g. *jida luogu* 擊打鑼鼓 "strike gongs and drums", whereas *daji* may additionally have metaphorical meaning of affecting someone emotionally or cracking down on something, e.g. *daji zuian* 打擊罪案 "strike at crimes".

4 Longitudinal and regional variations in metaphorization of *da*

4.1 Incipient longitudinal variations

Developments of incipient metaphorization (Type II) verbs could be traced by comparing their metaphorical usage in the two periods 1995-2000 and 2011-2016. Some entries have increased metaphorical usage over time, indicating a shift to metaphorical type (III).[10] As an example, *datong*

[9] *Da*-X may be further classified into mainly three groups in terms of the nature of X: X could be objects (*dagu* 打鼓 "to drum", *daci* 打氣 "(lit.) pump air"), resultative or directional complements (*dasi* 打死 "beat to death", *dajin* 打進 "compete to enter") or verbs (*dadou* 打鬥 "to fight", *dazao* 打造 "(lit.) forge, fabricate"), symbolized as VO, VC and VV respectively. While VO (42%) and VV (64%) have the highest proportion of Type III, VC is distinct from them in having the largest proportion of Type II (65%).

[10] About half have undergone no significant change, e.g. *dapo* 打破 "(lit.) break (a vase) vs. (metaph.) break (a deadlock)". A

打通 "(lit.) hit-through" could mean literally "to open up passage by hitting or digging" as in (1) below, and could also mean metaphorically "to connect" as in (2). The overall metaphorical usage of *datong* has been doubled from 29% to 60% in the two periods. Notably, there are latitudinal variations which have resulted from longitudinal developments. As shown in Figure 1, BJ has the greatest increase in metaphorical usage (+58%), followed by TW (+19%), with HK having a slight decrease (-6%).

(1) *Zhongyu* ***datong*** *suidong jiuchu san-ming-gongren* 終於**打通**隧洞救出三名工人 "Finally dug through a tunnel and rescued three workers." (1995)

(2) ***Datong*** *yu shehui dazhong de lianxi* **打通**與社會大眾的聯繫 "Punched through the connection to the public in society." (2016)

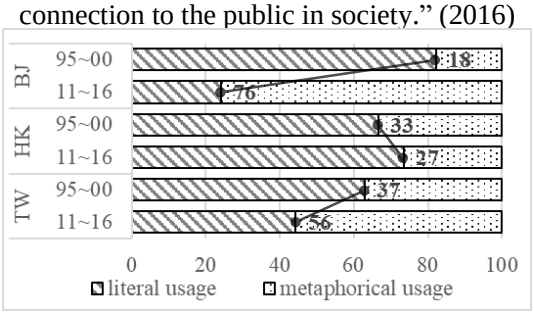


Figure 1. Change in usage of *datong* 打通

Another example of a shift to full metaphorical usage is *daxiang* 打響 "(lit.) hit-loud". Literally, it could mean "to make noise by gun firing" as in (3). Metaphorically, it could mean "to increase one's popularity" as in (4). Different from *datong* 打通, *daxiang* 打響 has almost shifted to full metaphorical usage. The overall percentage of metaphorical usage has increased from 82% to 92% in the two periods. The results of regional variations are shown in Figure 2. In BJ, the metaphorical usage of *daxiang* has increased from 72% to 85% in the two periods. In contrast, HK and TW have maintained around 95% metaphorical usage in the two time periods. This shows that while *daxiang* in BJ is still on its way to gaining metaphorical usage, HK and TW have almost completed the final stage of metaphorization.

few have gained literal usage, e.g. *daji* 打擊 "(lit.) hit (a baseball) vs. (metaph.) affect (one's feeling)", which could be attributed to common popularity of baseball in Taiwan.

(3) *Nanchang Qiyi* ***daxiang*** *di-yi-qiang de zhandou didian* 南昌起義**打響**第一槍的戰鬥地點 "The place where they fired the first shot in the Nanchang uprising" (1997)

(4) *Bing xunsu zai meishiquan-zhong* ***daxiang*** *zhimingdu* 並迅速在美食圈中**打響**知名度 "And rapidly struck up popularity in the restaurant trade." (2015)

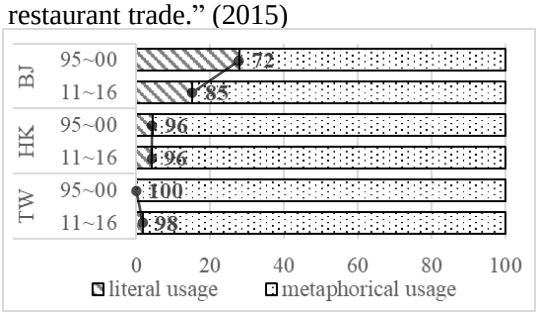


Figure 2. Change in usage of *daxiang* 打響

Dazhang 打仗 "hit-conflict",[11] on the other hand, shows the initial stage of metaphorization. It could literally mean "to fight a war" as in (5), or metaphorically mean "to compete" as in (6). *Dazhang*'s metaphorical usage has increased from 5% to 12% in the two periods. Figure 3 shows that while *dazhang* retains full literal usage in BJ (0% metaphorical usage), it has increasing metaphorical usage in HK (+15%) and TW (+14%). This could point to the initiation of metaphorization in HK and TW but not in BJ.

(5) *Budui xingjun* ***dazhang*** 部隊行軍**打仗** "The troops are transferred to fight in war." (1998)

(6) *Gai-ju bei zhi nalai gen Gangshi kaitaiju duihan* ***dazhang*** 該劇被指拿來跟港視開台劇對撼**打仗** "This show was said to be taken to compete against HKTV's show." (2014)

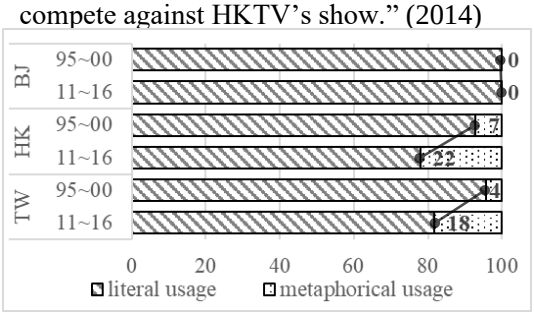


Figure 3. Change in usage of *dazhang* 打仗

[11] Note that *zhang* 仗 means "conflict" only in collocation with *da* 打 "hit". For example, *da* in *dabizhang* 打筆仗 "to fight a pen battle" cannot be replaced by *fadong* 發動 "launch", in contrast with *zhan* 戰, which means "war/battle" by itself, e.g. *fadongbizhan* 發動筆戰 "to wage a pen battle".

4.2 Incipient regional variations

The *overall* latitudinal variations among the three communities are also significant. The most frequent entry *daji* 打擊 "(lit.) hit-strike" may be used literally to refer to physical hitting in (7) and metaphorically as affecting someone emotionally as in Sent. (8). Remarkably, BJ, HK and TW differ in the *overall* percentage of metaphorical usage of *daji* in the two time periods. Almost all usage in HK are metaphorical (99%), while by comparison, there are about 88% in BJ and about 73% in TW.

(7) *Junfang yi jueding zhe kuan diduidi **daji** de duoguan huojian jiang bu-zai qianjin bushu waidao* 軍方已決定這款地對地**打擊**的多管火箭將不再前進部署外島 "The military had decided that this type of land-to-land multi-tube strike rocket would no longer be deployed to the outter islands." (2011)

(8) *Gai ming pengyou zai jijin chushi hou jingshen da shou **daji*** 該名朋友在基金出事後精神大受**打擊** "Following the incident with the Foundation, this friend was emotionally stricken." (2012)

Another prominent example is *dadiao* 打掉 "(lit.) hit-drop". It literally means "to let something drop through physical striking", as in (9). It may also be used metaphorically to mean "wipe out some (illegal) parties" as in (10). It displays great regional variations with BJ having almost full metaphorical usage (98%) while 79% for HK and 39% for TW.

(9) *que bei hui quan dadao zuolian, lian shouji ye bei **dadiao*** 卻被揮拳打到左臉，連手機也被**打掉** "Rather his left face was struck by the swinging fist, even his handphone was struck off." (2011)

(10) *gongan bumen **dadiao** fanzui tuanhuo 1230 ge* 公安部門**打掉**犯罪團夥 1230 個 "The public security bureau struck down 1230 criminal gangs." (2012)

Among the top 20 most frequent incipient words (covering 95% of all tokens of common incipient words), BJ stands out by having the largest number of words (12) with top percentage of metaphorical usage. HK, by comparison, only has 6, and TW has even fewer, only 4:[12]

BJ (12/20): *dazao* 打造 "hit-make", *dapo* 打破 "hit-break", *dakai* 打開 "hit-open", *dachu* 打出 "hit-out", *daxia* 打下 "hit-down", *datong* 打通 "hit-through", *dajia* 打架 "hit-fight", *dajiaodao* 打交道 "hit-contact-road", *dasao* 打掃 "hit-sweep", *daduan* 打斷 "hit-break", *dadiao* 打掉 "hit-drop", *dazhong* 打中 "hit-at"

HK (6/20): *daji* 打擊 "hit-strike", *daru* 打入 "hit-in", *daqi* 打氣 "hit-air", *dacheng* 打成 "hit-become", *dazhang* 打仗 "hit-conflict", *dadao* 打倒 "hit-collapse"

TW (4/20): *dajin* 打進 "hit-enter", *daqi* 打氣 "hit-air", *daxiang* 打響 "hit-loud", *dazhong* 打中 "hit-at"

These findings on regional variations could be confirmed by the metaphorical usage of the top 20 incipient words in each community. Figure 4 shows that BJ again has the highest percentage (71.8%), followed by HK (69.6%) and TW has the smallest (66.8%). These findings point to BJ having a greater tendency than either HK or TW to use common words metaphorically.

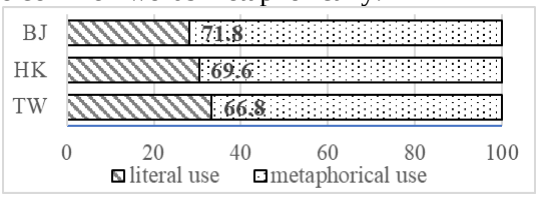


Fig 4. Metaphorical usage of top 20 Type II entries

4.3 The extent of metaphorization in three Chinese communities

The important incipient Type II verbs with variable entries have provided the basis for the study on the variations in metaphorization across the three communities. The Metaphorization Index (MI) we have proposed characterizes the extent of metaphorization. When aggregated for the top 50 common entries in each community as shown in Figure 5, BJ has the highest MI (81.6%), closely followed by HK (80.2%), with TW the lowest (73.3%). It shows that both BJ and HK have a greater tendency to use words metaphorically than TW. This supports the finding in the previous section and underscores even more striking overall differences among the three communities.

[12] *Daji* 打氣 has the same percentage in HK and TW while *dazhong* 打中 has the same percentage in BJ and TW.

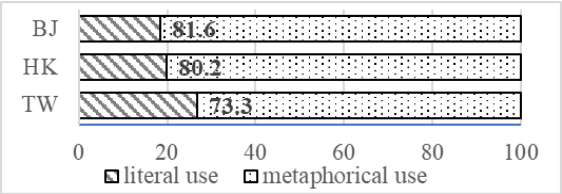


Figure 5. MI for top 50 common *da*-X verbs

4.4 Predominant local entries in three Chinese communities

In contrast to comparing the MI of the common incipient entries in the three communities, we also examine verbs found only in a single community, i.e. top most frequency (≥90%) local verbs with metaphorical usage in each community. Among the top 200 entries in each community, HK stands out with 19.5% predominant metaphorical entries, which is even larger than the sum of BJ (5.5%) and TW (10.5%):

BJ (5.5%): e.g. *dafen* 打分 "to score, (lit.) hit-points", *daxin* 打新 "staging (to buy initial round of publicly offered stocks), (lit.) hit-new", *danian* 打蔫 "being tired, (lit.) hit-wither", etc.

HK (19.5%): e.g. *dajiu* 打救 "to rescue, (lit.) hit-save", *dazhengqihao* 打正旗號 "under the name of, (lit.) hit-right-flag", *dashuding* 打書釘 "to read at bookstores without purchase, (lit.) hit-book-nail", etc.

TW (10.5%): e.g. *dalian* 打臉 "to offend or upset someone/ give a slap in the face, (lit.) hit-face", *dapi* 打屁 "to chit-chat, (lit.) hit-fart", *dashu* 打書 "to promote new books, (lit.) hit-book", etc.

It shows that HK has a higher degree of metaphorization when it comes to items that are unique to their own community. This could be a source of new metaphorical *da* 打 in the Chinese language as a basis of internal language contact.

5 Concluding remarks

5.1 Metaphorization and light verb development

Light verbs are distinct from regular verbs in terms of both *grammatical* and *semantic* properties. For instance, English *give* could only take concrete nouns in its regular ditransitive usage with the meaning of "offering something to someone". In terms of light verb usage (e.g. *give a pull*, *give a wink*), however, *give* may take eventive nouns with bleached meaning of performing actions. In Chinese, previous studies had often focused on the *grammatical* properties of light verbs (D. Zhu 1985, Diao 2004, *i.a.*). However, the process of *semantic* bleaching has been rarely studied, especially in the context of metaphorization.

Based on 2175 entries with *da* 打, the current study shows a striking asymmetry in metaphorization on the basis of word structure. *Da* in word-final position (X-*da*) tends to be used literally such as *jida (luogu)* 擊打(鑼鼓) "strike gongs and drums", whereas in word-initial position (*da*-X) it tends to be used metaphorically such as *daji (zuian)* 打擊(罪案) "strike at crimes". We observe that *grammatical* structure is correlated with *semantic* bleaching in terms of metaphorization. We note that X-*da* is often manifested as a modifier-head structure where the modifier specifies the manner of *da* "hit". Here, *ji* 擊 "strike" is a hyponym of *da* 打 "hit" and it specifies that the main action of hitting is accompanied by forceful striking, as opposed to other hitting actions such as *guoda* 摑打 "to slap" or *chuida* 捶打 "to pound". As noted, in *jida* "strike-hit" *da* is the main action modified by *ji* "strike" which denotes a subset of hitting actions. This hierarchical relationship preempts metaphorization which involves an opposite relaxation of meaning. In contrast, *da* in *daji* "hit-strike" is not modified by *ji* but forms with it a *coordinate* VV structure of near-synonyms. Since *da*'s meaning is not hierarchically constrained but forms an equal partnership in VV structure, it could undergo semantic bleaching and metaphorization by becoming associated with other kinds of objects when there is an appropriate triggering agent.[13]

The correlation of grammatical properties and metaphorization is also reflected in regional variations. Our current study arrives at a conclusion similar to that of Jiang et al. (2016)

[13] Metaphorization may be triggered by quasi-concrete objects, such as *(dazao) qicheye de hangkongmujian* (打造)汽車業的航空母艦 "(forge) an aircraft carrier of the automotive business" (Tsou, Chin & Kwong 2011). Notably, the metaphorization of *dazao* is achieved by relaxing the selectional restriction on concrete nouns to include abstract nouns through pseudo objects. This again shows an interaction of grammatical properties and metaphorization in light verb development and deserves further exploration.

which studied the grammatical variations of light verbs in BJ and TW. They noted that *jinxing* 進行 in TW may take VO complements, but not BJ. *Jinxing* in TW was thus regarded as "more transitive" and more verbal than in BJ. On the other hand, the current study has found that the MI for TW is lower than BJ and HK. This echoes the conclusion of Jiang et al. (2016) from a different perspective: while they found TW light verbs to be more conservative *grammatically*, this study has found them to be more conservative in *metaphorization* of *da*, which is a *semantic* process. The correlation of grammatical properties and metaphorization in light verb development seems to hold across Chinese speech communities.

5.2 Latitudinal and longitudinal variations based on a synchronous corpus, LIVAC

The current study is based on the LIVAC corpus which differs from other corpora through the adoption of a rigorous sampling ("Windows") approach in curating data from six Chinese speech communities since 1995 (Tsou & Kwong 2015). Thus LIVAC could contribute to the study of longitudinal variations by addressing latitudinal variations across the 22-year time span. In particular, it offers an opportunity to view some longitudinal variations in light verb developments. In the case of *da*, it is observed that incipient metaphorization has developed at different rates in the three communities from 1995 to 2016. As an illustration, while metaphorization of *daxiang* 打響 "(lit.) hit-loud" may be completed in HK and TW with steady metaphorical usage (>95% over time), it is still ongoing in BJ as indicated by an increase in metaphorical usage from 72% to 85% over the same period. This kind of comparison has to rely on a rigorously curated corpus.

The efficacy of a rigorously cultivated corpus could also be demonstrated by comparing two other top light verbs in Chinese speech communities such as *jinxing* 進行 "proceed" and *zuo* 做 "make", which have been studied in the literature. Under normal circumstance of natural societal equilibrium, it could be assumed that there would be no significant variation in usage, and their relative percentages within the two periods 1995-2000 and 2011-2016 should be about equal. However, the distribution of these two light verbs has varied drastically with different extent in BJ, HK and TW. As shown in figure 6, the relative percentage of *jinxing* has undergone decrease in the three communities and HK's drop is the sharpest (-43%), followed by TW (-32%) and Beijing the modest (-11%), accompanied by a corresponding increase of *zuo*. The change in relative percentage of *jinxing* and *zuo* is striking in terms of both the sharp rates and the regional variations. While the cause for this shift awaits careful exploration, it may be suggested that the formal register marked by *jinxing* is on the decline more prominently in HK and TW than BJ with a possible compensatory shift to the most colloquial *zuo*. A systematic study in terms of register and genre markers in LIVAC could provide fuller and better answers.

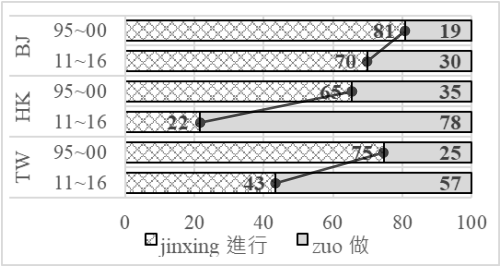


Figure 6. Relative percentage changes of *jinxing* 進行 and *zuo* 做

5.3 Towards the future

This paper examines a very common but little studied light verb *da* 打 in three Chinese speech communities: Beijing, Hong Kong and Taiwan, by monitoring a synchronous corpus. The findings offer some glimpses into the differential metaphorization of *da* in the three communities and also their longitudinal developments as well as into how and why metaphorization has come about, particularly the general interaction between grammatical structure and semantic bleaching. We expect this approach would facilitate further explorations of the underlying mechanisms of metaphorization by examining more Chinese dialects in which it is an ongoing process and other languages as well.

Acknowledgements

For data preparation and annotation, we wish to thank Wing Fu Tsoi, Janice Chong, and especially Renee Ji. We also wish to thank the two anonymous reviewers for their valuable comments. All errors remain the authors' own responsibilities.

References

Cattell, Ray. 1984. *Composite Predicates in English*. North Ryde, New South Wales: Academic Press Australia.

Chae, Hee-Rahk. 1996. Light verb constructions and structural ambiguity. *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation (PACLIC 11)*: 99-107.

Chen, Yongli. 2003 Xingshi dongci de fanwei, cilei ji tezheng [The scope, subcategories and features of dummy verbs]. *Academic Journal of Jinyang* 3: 92-94.

Diao, Yanbin. 2004. *Xiandai Hanyu Xuyi Dongci Yanjiu* [A Study on Delexical Verb in Modern Chinese]. Dalian: Liaoning Normal University Press.

Grimshaw, Jane, and Armin Mester. 1988. Light verbs and θ-marking. *Linguistic Inquiry* 19: 205-232.

Her, One-Soon, Wei-Tien Dylan Tsai, Jung-Hsing Chang, Kawai Chui, Jennifer M. Wei, and D. Victoria Rau. 2016. *Yuyan Ai-bu-ai? Yuyanxuejia de Kanfa* [Language Cancer: From the Perspective of Linguists]. Taipei: Linking.

Huang, Chu-Ren, Jingxia Lin, Menghan Jiang, and Hongzhi Xu. 2014. Corpus-based study and identification of Mandarin Chinese light verb variations. In Marcos Zampieri, Liling Tan, Nikola Ljubešic and Jörg Tiedemann, eds, *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 1-10. Dublin: Association for Computational Linguistics and Dublin City University.

Jespersen, Otto. 1954. *A Modern English Grammar on Historical Principles*. London: George Allen & Unwin & Copenhagen: Ejnar Munksgaard.

Jiang, Haiyan. 2020. Chouxiang dongzuo shijian yuyi fanchou yu Hanyu xingshi dongci [Abstract action event semantic category and Chinese dummy verbs]. Doctoral dissertation, Jilin University.

Jiang, Menghan, Dingxu Shi, and Chu-Ren Huang. 2016. Transitivity in light verb variations in Mandarin Chinese – a comparable corpus-based statistical approach. *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*: 459-468.

Lakoff, George, and Mark Johnson. 1980. *Metaphor We Live By*. Chicago and London: University of Chicago.

Li, Feng, and Yun Chai. 1995. "V+yi" lei xuhua dongci de binyu [The objects of "V+yi" dummy verbs]. *Journal of Xinjiang Education Institute* 4: 68-71.

Li, Hanlei. 2003. "Jiayi" de yuyong gongneng [The pragmatic functions of *jiayi*]. *Journal of Suzhou Education College* 1: 11-16.

Lin, Jingxia, Hongzhi Xu, Menghan Jiang, and Chu-Ren Huang. 2014. Annotation and classification of light verbs and light verb variations in Mandarin Chinese. In Takuya Nakamura, ed., *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, 75-82. Dublin: Association for Computational Linguistics and Dublin City University.

Lü, Shuxiang. 1980. *Xiandai Hanyu Babaici* [Modern Chinese 800 Words]. Beijing: Commercial Press.

Lü, Shuxiang. 1999. *Xiandai Hanyu Babaici (Zengdingben)* [Modern Chinese 800 Words (Enlarged Edition)]. Beijing: Commercial Press.

Mao, Hongyuan. 1997. Hua xingshi dongci he xingshihua dongci [On formal verbs and formalized verbs]. *Journal of Kashgar Teachers College* 4: 78-79.

Pham, Hoa. 1999. Cognate objects in Vietnamese transitive verbs. *Toronto Working Papers in Linguistics* 27: 227-246.

Ren, Fengmei. 2013. A grounding approach to the semantic meaning of the light verb *da*. In Pengyuan Liu and Qi Su, eds., *Chinese Lexical Semantics: 14th Workshop, CLSW 2013, Zhengzhou, China, May 10-12, 2013. Revised Selected Papers*, 88-96. Berlin & Heidelberg: Springer-Verlag.

Shen, Jiaxuan, and Jiangzhi Zhang. 2013. Ye tan xingshi dongci de gongneng [On the grammatical function of dummy verbs in Chinese]. *TCSOL Studies* 2: 8-17.

Su, Lei. 2009. Lun "da" de yufahua [On the grammaticized process of the word "fight"]. *Journal of Hubei University of Education* 5: 30-31.

Tsai, Wei-Tien Dylan. 2017. Jiwuhua, shiyong jiegou yu qingdongci fenxi [Transitivization, applicative construction and light verb analysis]. *Contemporary Research in Modern Chinese* 19: 1-13.

Tsou, Benjamin K., and Oi Yee Kwong. 2015. LIVAC as a monitoring corpus for tracking trends beyond linguistics. *Journal of Chinese Linguistics Monograph Series* 25: 447-471.

Tsou, Benjamin K., Andy C. Chin, and Oi Yee Kwong. 2011. From synchronous corpus to monitoring

corpus, LIVAC: The Chinese case. In Friedrich Laux and Lena Strömbäck, eds., *The Third International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2011)*, 175-180. St. Maarten: IARIA XPS Press.

Wang, Leslie Fu-mei. 2004. A corpus-based study of mandarin verbs of doing. *Concentric: Studies in Linguistics* 30.1: 65-85.

Wang, Li. 1985. *Zhongguo Xiandai Yufa* [Modern Chinese Grammar]. Beijing: Commercial Press.

Yan, Zhongsheng. 1998. Shuo "houxu dongcixing binyu dongci" [On "verbs taking verbal objects"]. *Journal of Hebei Normal University (Natural Science)* 2: 91-93.

Zhou, Gang. 1987. Xingshi dongci de cifenlei [Subdivision of dummy verbs]. *Chinese Language Learning* 1: 11-14.

Zhu, Dexi. 1985. Xiandai shumian hanyu li de xuhua dongci he mingdongci [Dummy verbs and nominal verbs in Modern Written Chinese]. *Journal of Peking University (Humanities and Social Sciences)* 5: 1-6.

Zhu, Jian-jun. 2004. "Da-V" zhi "da" de yufahua tanxi [The grammaticalization of *da* in the structure of *da-V*]. *Research in Ancient Chinese Language* 3: 38-44.

Zhuang, Huibin. 2014. Xiandai Hanyu Qingdongci "da" de laiyuan chuyi [On the etymology of the light verb *da* in Chinese]. *Language Teaching and Linguistic Studies* 3: 67-74.

Sensorimotor Enhanced Neural Network for Metaphor Detection

Mingyu Wan[1,2], **Baixi Xing**[3], **Qi Su**[1], **Pengyuan Liu**[3] **and Chu-Ren Huang**[2]

[1]School of Foreign Languages, Peking University
[2]Department of Chinese Bilingual Studies, The Hong Kong Polytechnic University
[3]School of Information Science Language, Beijing Language and Culture University
{wanmy,Liupengyuan,sukia}@pku.edu.cn
{xingbaixi}@gmail.com,{churen.huang}@polyu.edu.hk

Abstract

Detecting metaphors is challenging due to the subtle ontological differences between metaphorical and non-metaphorical expressions. Neural networks have been widely adopted in metaphor detection and become the main stream technology. However, linguistic insights have been less utilized. This work proposes a linguistically enhanced model for metaphor detection extending one published work (WAN et al., 2020) by incorporating the modality norms into attention-based Bi-LSTM. Results show that the current model outperforms most recent works by 0.5%-11% F1, indicating the effectiveness of using modality norms for metaphor detection. This work provides a new perspective to detect token-level metaphoricity by leveraging the modality mismatch between words.

1 Introduction

Metaphors are prevalent in our everyday language even without our consciousness of its presence as we speak and write. It induces the unknown using the known, explains the complex using the simple, and helps us to emphasize the relevant aspects of meaning resulting in effective communication.

In general, metaphor involves certain concept transfer from one domain (Source) to another (Target), as in 'sweet voice' (using taste to describe sound). Lakoff (1980) describes metaphor as a cognitive mechanism (a property of language) reflected by our conceptual system for structuring our understanding of the world. It is a fundamental way to relate our physical and familiar social experiences to a multitude of other subjects and contexts (Lakoff and Johnson, 2008).

As a popular linguistic device, metaphors encode versatile ontological information, which usually involve e.g. domain transfer (Ahrens et al., 2003; Ahrens and Jiang, 2020), sentiment reverse (Steen et al., 2010) or modality shift (Winter, 2019) etc. Therefore, detecting the metaphors in texts is essential for capturing the authentic meaning of the texts, which can benefit many natural language processing applications, such as machine translation, dialogue systems and sentiment analysis (Tsvetkov et al., 2014).

To better understand the intrinsic properties of metaphors and to provide an in-depth analysis to this phenomenon, we propose a linguistically-enriched deep learning model extending one published work (WAN et al., 2020) at ACL Figlang 2020 workshop by incorporating the modality norms into attention-based BiLSTM. As a continuation of their work, we conduct the current research to further testify the effectiveness of leveraging conceptual norms for metaphor detection. For standard reference, we adopt the dataset of the first and second shared tasks of metaphor detection on verbs of the VUA corpus (Klebanov et al., 2018)[1]. Details about the experiment are given in Sections 3-5.

2 Related Work

Research on metaphors have been mainly explored in the context of political communication, mental health, teaching, discourse analysis, assessment

[1]http://www.vismet.org/metcor/documentation/home.html

of English proficiency, among others (Ahrens and Jiang, 2020; Thibodeau and Boroditsky, 2011; Kathpalia and Carmel, 2011; Klebanov et al., 2008; Semino, 2008; Billow et al., 1997; Bosman, 1987).

Over the last decade, automated detection of metaphor has gained increasing research interest among the Natural Language Processing community. Many approaches have been proposed with systems such as traditional machine learning classifiers, deep neural networks and sequential models etc., trained on features of word vectors, n-grams, lexical information, semantic classes, concreteness, word associations, constructions and frames etc. (Hong, 2016; Rai et al., 2016; Do Dinh and Gurevych, 2016; Klebanov et al., 2014; Wilks et al., 2013; Bizzoni and Ghanimifard, 2018; Klebanov et al., 2015).

Early studies of metaphor detection tend to adopt feature-engineering in a supervised machine learning paradigm, which construct feature vectors based on concreteness and imageability, semantic classification using WordNet, FrameNet, VerbNet, SUMO ontology, property norms and distributional semantic models, syntactic dependency patterns, sensorial and vision-based features (Alnafesah et al., 2020; Klebanov et al., 2016; Shutova et al., 2016; Gutierrez et al., 2016).

Recently, deep learning methods have been explored and become the main stream technology for metaphor detection (Mao et al., 2019; Dankers et al., 2019; Gao et al., 2018; Wu et al., 2018; Rei et al., 2017; Gutierrez et al., 2017). To name a few advances, Brooks and Youssef (2020) build up an ensemble of RNN models with Bi-LSTMs and bidirectional attention mechanisms. Chen et al. (2020) employs BERT to obtain the sentence embeddings, and then a linear layer is applied with softmax on each token to make predictions. Maudslay et al.(2020) combines the concreteness of a word with its static and contextual embeddings as inputs into a deep Multi-layer Perceptron network for predicting metaphoricity. Gong et al.(2020) used RoBERTa to obtain word embeddings and concatenate it with linguistic features (e.g. WordNet, VerbNet) as well as other features (e.g. POS, topicality, concreteness), and then feed them into a fully-connected Feedforward network to make predictions.

Despite many advances in the above studies, metaphor detection remains a challenging task. The semantic and ontological differences between metaphorical and non-metaphorical expressions are often subtle and their perception may vary from person to person. These methods show different strengths on detecting metaphors, yet each has its respective disadvantages, such as having generalization problems or lack association of their results with the intrinsic properties of metaphors. In Wan et al. (2020)'s work, they use conceptual features of modality and embodiment norms for metaphor detection based on traditional classifiers (Logistic Regression), which demonstrates the effectiveness of using modality exclusivity information for predicting metaphoricity. The current work aims to merge both strengths of linguistic wisdom and deep learning power into one architecture with the modality enriched neural networks, as illustrated in Section 4.

3 Data Description

3.1 The VUA Corpus

The VU Amsterdam Metaphor Corpus (VUA) (Tekiroğlu et al., 2015)[2] is used in the experiment for training and testing. The dataset consists of 117 fragments sampled across four genres from the British National Corpus: Academic, News, Conversation, and Fiction. The data is annotated using the MIPVU procedure (Steen, 2010) with a strong inter-annotator agreement (k>0.8). This dataset has been used as the competition corpus for two shared tasks on metaphor detection (Leong et al., 2018; Leong et al., 2020), which is publicly available for standard reference.

Information about the size of the sub-genres is given in Table 1. The training and testing texts, sentences, tokens and percentage of metaphors breakdown of the VUA verb track[3] is given in Table 2.

Text Genres	No. of Tokens	No. of Fragments
Academic texts	49,561 tokens	16 fragments
Conversation texts	48,001 tokens	24 fragments
Fiction texts	44,892 tokens	12 fragments
News texts	45,116 tokens	63 fragments
TOTAL	187,570 tokens	115 fragments

Table 1: Data components of the VUA corpus

[2]http://www.vismet.org/metcor/documentation/home.html

[3]The prediction and evaluation in this paper focuses on the verbs tokens only.

Dataset	Training	Testing
#texts	90	27
#sents	12,123	4,081
#tokens	17,240	5,873
%M	29%	-

Table 2: Number of texts, sentences, tokens, and percentage of metaphors for the VUA corpus

3.2 The Modality Norms

The Lancaster Sensorimotor norms (hereinafter modality norms) collected by Lynott (2019) is used for constructing the linguistic features in the deep learning model. The data include measures of sensorimotor strength (0-5 scale indicating different degrees of sense modalities/action effectors) for 39,707 English words across six perceptual modalities: touch, hearing, smell, taste, vision and interception, and five action effectors: mouth/throat, hand/arm, foot/leg, head (excluding mouth/throat), torso.[4]. Examples of five random words and their six main modality scores are demonstrated in Table 3.

Word	A	G	H	V	O	I
Adopt	1.222	0.056	1.056	**1.889**	0.111	1.222
Big	0.944	0.167	2.722	**3.889**	0.111	0.333
Daze	0.455	0.000	0.000	1.953	0.000	**3.253**
Eat	1.263	**4.526**	2.158	2.632	2.421	2.474
Learn	**3.941**	0.765	1.765	3.882	0.588	1.529

A: Auditory; G: Gustatory; H: Haptic;
V: Visual; O: Olfatory; I: Interoceptive

Table 3: Examples of the Modality Norms

The modality with the highest scores (highlighted) among the six senses of the words marks the dominant sense modality for each word, such as '**Visual**' for words '*Adopt*' and '*Big*'. As sensorimotor information plays a fundamental role in cognition, these norms provide a valuable knowledge representation to the conceptual categories of the tokens in the corpus which may serve as salient features for inferring metaphors. Motivated by the above idea, we propose a modality enriched neural network to further testify its effectiveness.

[4] https://osf.io/7emr6/

4 The Modality Enriched Model

In the modality enriched model, words are processed with the integration of linguistic features and word embedding. We map the modality scores of the words to the norms and obtain modality representations and then use them as inputs to neural networks. The architecture of the modality enriched model is demonstrated in Figure 1.

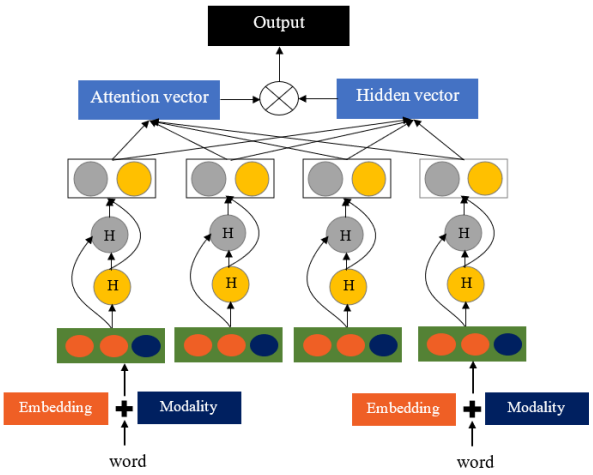


Figure 1: The Modality Enriched Model

Let H $\in$ R$^{d\times N}$ be a matrix consisting of hidden vectors $[h_1, h_2....h_N]$ that is produced by LSTM, where d is the size of hidden layers and N is the length of the given sentence. The attention mechanism will produce an attention weight α. The final sentence representation is given by:

$$h = H \times \alpha^T$$

We also add a additional Linear layer. The final probability distribution is:

$$y = softmax(W_s h + b_s)$$

Let y be the target distribution for sentence, $\hat{y}$ be the predicted sentiment distribution. Train to minimize the cross-entropy error between y and $\hat{y}$ for all sentences.

$$loss = -\sum_i \sum_j y_i^j log\hat{y}_i^j + \lambda \mid\mid \theta \mid\mid^2$$

We use glove embedding and modality vectors to represent the input data. The red circle denotes the

usual embedding, the gray circle represents the linguistics feature. We concatenate both representation to generate a new representation as the input of the next layer. LSTM layer produces a hidden status of each word in a sentence. We use these status to calculate an attention weight which will be multiplied with output of LSTM layer. Finally, we get a probability distribution of 0-1 label to train the model and as the prediction result.

5 Experimental Results

In order to evaluate the effectiveness of the proposed model for metaphor detection, we randomly select a development set (4,380 tokens) from the training set (17,240 tokens) in proportion to the Train/Test ratio of the task in Leong et al. (2020). The evaluation results are summarized in Table 4 below:

Category	Approach	P	R	F1
Baseline	uni-gram + LR	0.52	0.66	0.58
Linguistic	modality + linear	0.61	0.56	0.58
	modality + LSTM	0.70	0.68	0.69
Neural	Glove + LSTM	0.74	0.75	0.75
Enriched	modality + Glove + LSTM	**0.77**	**0.76**	**0.76**

Table 4: Evaluation Results of the System

In Table 4, the baseline of using unigram as features and logistic regression (LR) as the classifier is implemented for a basic comparison. It is a commonly adopted baseline in the tasks of metaphor detection. We also implement several sub-categories of approaches before trying the enriched model, including the linguistic and neural networks in separate and also in combination. The results show an 18% F1 improvement of the enriched model over the baseline, a 7% F1 improvement over pure linguistic model, a 1.5% F1 improvement over the pure neural network model, and this superiority is salient and consistent in terms of both P (Precision) and R (Recall).

To further demonstrate the effectiveness of our method, this following table presents the comparisons of our system to some highly related recent works on the same task. All the results are publicly available, as reported in Leong et al. (2020). The detailed results are displayed in Table 5 below:

Our method obtains very promising results: it outperforms 6/7 highly related works to a great extent (0.5%-11% F1 gain), also approaching a reachable performance (a 4% F1 discrepancy) to the Top 1 work in record (Su et al., 2020). Moreover, our results are consistently superior to the top baseline and other linguistically-based or deep learning approaches. This suggests the effectiveness of leveraging modality norms in neural networks for metaphor detection, echoing the hypothesis in Wan et al. (2020) that metaphor manifests a concept mismatch (modality shift in particular) between source and target.

6 Conclusions

We presented a linguistically enhanced method for metaphor detection of VUA verbs using modality features plus attention-based neural network in continuation of Wan et al. (2020)'s first implementation on using conceptual norms for metaphor detection. Inter- and cross-approach comparisons among state-of-the-arts all demonstrate the effectiveness of adding modality information into neural networks for enhancing the performance of metaphor detection. It reconfirms the hypothesis that metaphor manifests a concept mismatch (modality shift in particular) between source and target. Future work will expand the current experiment to predictions of all four lexical words across more datasets.

References

Kathleen Ahrens and Menghan Jiang. 2020. Source domain verification using corpus-based tools. *Metaphor and Symbol*, 35(1):43–55.

Kathleen Ahrens, Siaw Fong Chung, and Chu-Ren Huang. 2003. Conceptual metaphors: Ontology-based representation and corpora driven mapping principles. In *Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14*, pages 36–42. Association for Computational Linguistics.

Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210.

Richard M Billow, Jeffrey Rossman, Nona Lewis, Deberah Goldman, and Charles Raps. 1997. Observing expressive and deviant language in schizophrenia. *Metaphor and Symbol*, 12(3):205–216.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and bilstms two neural networks for sequential

Work	Method	F1
Wan et al. (2020)	modality + other features + LR	0.652
Kuo and Carpuat (2020)	Bi LSTM+Embeddings+Unigram Lemmas+Spell Correction	0.686
Kumar & Sharma (2020)	Character embeddings+Similarity Networks+Bi-LSTM+Transformer	0.717
Liu et al. (2020)	BERT, XNET + POS tags + Bi-LSTM	0.730
Li et al. (2020)	ALBERT + BiLSTM	0.755
Top base: Devlin et al. (2018)	BERT: Pre-training of deep bidirectional transformers	0.756
The current study	**modality + Glove + LSTM**	**0.761**
Top 1: Su et al. (2020)	Global and local text information+Transformer stacks	0.804

Table 5: Comparison of Results of Our System to State-of-the-art Works

metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101.

Jan Bosman. 1987. Persuasive effects of political metaphors. *Metaphor and Symbol*, 2(2):97–113.

Jennifer Brooks and Abdou Youssef. 2020. Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 244–249.

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. 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 2218–2229.

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*.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193.

E Dario Gutierrez, Guillermo A Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930.

Jisup Hong. 2016. Automatic metaphor detection using constructions and frames. *Constructions and frames*, 8(2):295–322.

Sujata S Kathpalia and Heah Lee Hah Carmel. 2011. Metaphorical competence in esl student writing. *Relc Journal*, 42(3):273–290.

Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. 2008. Lexical cohesion analysis of political speech. *Political Analysis*, pages 447–463.

Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20.

Beata Beigman Klebanov, Chee Wee Leong, E Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2018. A corpus of non-native written english annotated for metaphor. 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 86–91.

Tarun Kumar and Yashvardhan Sharma. 2020. Character aware models with similarity learning for metaphor

detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 116–125.

Kevin Kuo and Marine Carpuat. 2020. Evaluating a bi-lstm model for metaphor detection in toefl essays. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 192–196.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *Chicago, IL: University of Chicago*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 vua metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29.

Shuqun Li, Jingjie Zeng, Jinhui Zhang, Tao Peng, Liang Yang, and Hongfei Lin. 2020. Albert-bilstm for sequential metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 110–115.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, pages 1–21.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226.

Sunny Rai, Shampa Chakraverty, and Devendra K Tayal. 2016. Supervised metaphor detection using conditional random fields. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 18–27.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. *arXiv preprint arXiv:1709.00575*.

Elena Semino. 2008. *Metaphor in discourse*. Cambridge University Press Cambridge.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

Mingyu WAN, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109, Online, July. Association for Computational Linguistics.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44.

Bodo Winter. 2019. Synaesthetic metaphors are neither synaesthetic nor metaphorical. *Perception metaphors*, pages 105–126.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu ngn at naacl-2018 metaphor shared task: Neural metaphor detecting with cnn-lstm model. In *Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA*.

A Parallel Corpus-Driven Approach to Bilingual Oenology Term Banks: How Culture Differences Influence Wine Tasting Terms

Vincent Xian Wang
Department of English
University of Macau
vxwang@um.edu.mo

Xi Chen
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Department of English, University of Macau
yb77703@um.edu.mo

Songnan Quan
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
songnan.quan@connect.polyu.hk

Chu-Ren Huang
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
The HK PolyU-PKU Research Centre on Chinese Linguistics
churen.huang@polyu.edu.hk

Abstract

This paper describes the construction of an English-Chinese Parallel Corpus of wine reviews and elaborates on one of its applications – i.e. an E-C bilingual oenology term bank of wine tasting terms. The corpus is sourced from Decanter China, containing 1211 aligned wine reviews in both English and Chinese with 149,463 Chinese characters and 66,909 English words. It serves as a dataset for investigating cross-lingual and cross-cultural differences in describing the sensory properties of wines. Our log-likelihood tests revealed good candidates for the Chinese translations of the English words in wine reviews. One of the most challenging features of this domain-specific bilingual term bank is the dominant many-to-many nature of term mapping. We focused on the one-to-many English-Chinese mapping relations of two major types: (a) the words without a single precise translation (e.g. "palate") and (b) the words that are underspecified and involve 'place-holder' translation (e.g. "aroma"). Our study differs from previous bilingual CompuTerm studies by focusing on an area where cultural and sensory experiences favour many-to-many mappings instead of the default one-to-one mapping preferred in scientific and jurisprudential areas. This necessity for many-to-many mappings in turn challenges the basic design feature of many state-of-the-art automatic bilingual term-extraction approaches.

1 Introduction

The textual data of wines in the field of wine informatics are increasingly accessible to the

public through the Internet in this big data era. Wine reviews have been much criticised in terms of the use of metaphors for describing wine-tasting experience that often goes so free that it becomes "difficult to understand" (Demaeker, 2017, p.117). However, Croijmans, Hendrickx, Lefever, Majid and Van Den Bosch (2020) refuted this line of criticism by exhibiting high consistency in the use of wine terms in 76,410 wine reviews they gathered, which also effectively trained a classifier that automatically and rather accurately predicted the wine colour (red, white, or rose) and grape variety (n=30) in the wine reviews that were new to the classifier. Largely consistent with the results of Croijmans et al. (2020), López-Arroyo and Roberts (2014) found wine reviews used a limited repertoire of commonly-used words to convey specialised senses about wine tasting experience, while extending metaphorical applications of the words. The sensory experience of wine tasting was also studied in the frame of "motions" by Caballero (2017), who drew on cognitive linguistic research on motion events to examine the description of the aromas and flavours of wine "travelling" to sensory organs. Caballero gathered 12,000 wine notes in both English and Spanish and identified similarities and differences in the description of motions between the two languages.

Research in sensory sciences and informatics focuses on the extracting meaningful information from the wine reviews. For example, Valente, Bauer, Venter, Watson and Nieuwoudt (2018) introduced a new approach of using formal concept lattices to visualise the sensory attributes of Chenin blanc and Sauvignon blanc wines. Palmer and Chen (2018) employed a large-scale dataset of wine reviews to perform regression predictions on the grade and price of wines. In linguistic studies, wine reviews provide sensory descriptions for the research on language and cognition. Thus, comparative studies based on bilingual wine reviews, such as Chinese and English, underline the issue of how sensory cognition is encoded across different languages.

Another possible research application is to build English-Chinese parallel corpus based on wine reviews for domain-specific machine translation or translation studies. Such corpora should follow established guidelines (e.g. Chang, 2004) in order to be sharable. Such a corpus is crucial as terms loaded with rich cultural tradition tend to be considered 'untranslatable'. In computational term banks, it often leads to one-to-many (cf. Lim, 2018, 2019), many-to-one, or many-to-many mappings, although the mapping relations do not seem to have attracted in-depth research in computational terminology. In this paper, we propose a parallel corpus-driven approach to culturally bound bilingual terms discovery. In particular, we look at English-Chinese bilingual wine-tasting terminology. Since modern table wine culture and technology are mostly borrowed in the direction from the Western world to China, we focus on the one-to-many mapping of terms in E-C wine terminology. Therefore, this paper will address: a) how we are constructing the English-Chinese parallel corpus of wine reviews; b) the application of this parallel corpus in computational E-C oenology terminology.

Another important issue in computational terminology that our study will raise lies in the design criteria and evaluation metrics. The assumed ideal world criterion in bilingual term extraction is to achieve perfect one-to-one mapping. Previous studies on formal information dominant domains (sciences, technology, law etc.) worked well under this default criterion. However, what happens when the best terms in the target language vary in a wide range according to the context? Is there a better algorithm for this complex mapping issue?

Figure 1: A Wine Review at Decanter (English)

Figure 2: A Wine Review at Decanter (Chinese)

2 The Parallel Corpus

We describe our parallel corpus in terms of the source data (cf. 2.1) and our corpus construction method (cf. 2.2).

2.1 Source Data

Our data consists of bilingual reviews published on DecanterChina.com (醇鉴中国 chún jiàn zhōngguó), a website (www.decanterchina.com) presented by Decanter magazine, an international wine authority. Each wine review (酒评 jiǔ píng) is presented in both English and Chinese (cf. Figures 1 and 2). One of the present authors who is an accredited English-Chinese translator in China studied the bilingual reviews and confirmed that the Chinese reviews were the human translations of the English ones, ruling out the possibility that they were the outcome of automatic machine translation.

We crawled the data by means of "request" and "Beautiful Soup" of Python. The website contains data on thousands of wines and each wine has a separate introduction page. Each page displays the name, score, region, grape variety, producer, alcohol level, reference price, and reviews of the wine, which are the focus of this study. By targeting the English and the corresponding Chinese URLs with the use of the English/Chinese switch button on the top right corner of each webpage, we wrote the scripts to simulate the process of clicking on each wine page, and automatically collected the content in each page (Figure 3). We saved the content into data frame (Figure 4), and manually removed the noise and inequivalent pairs in the data.

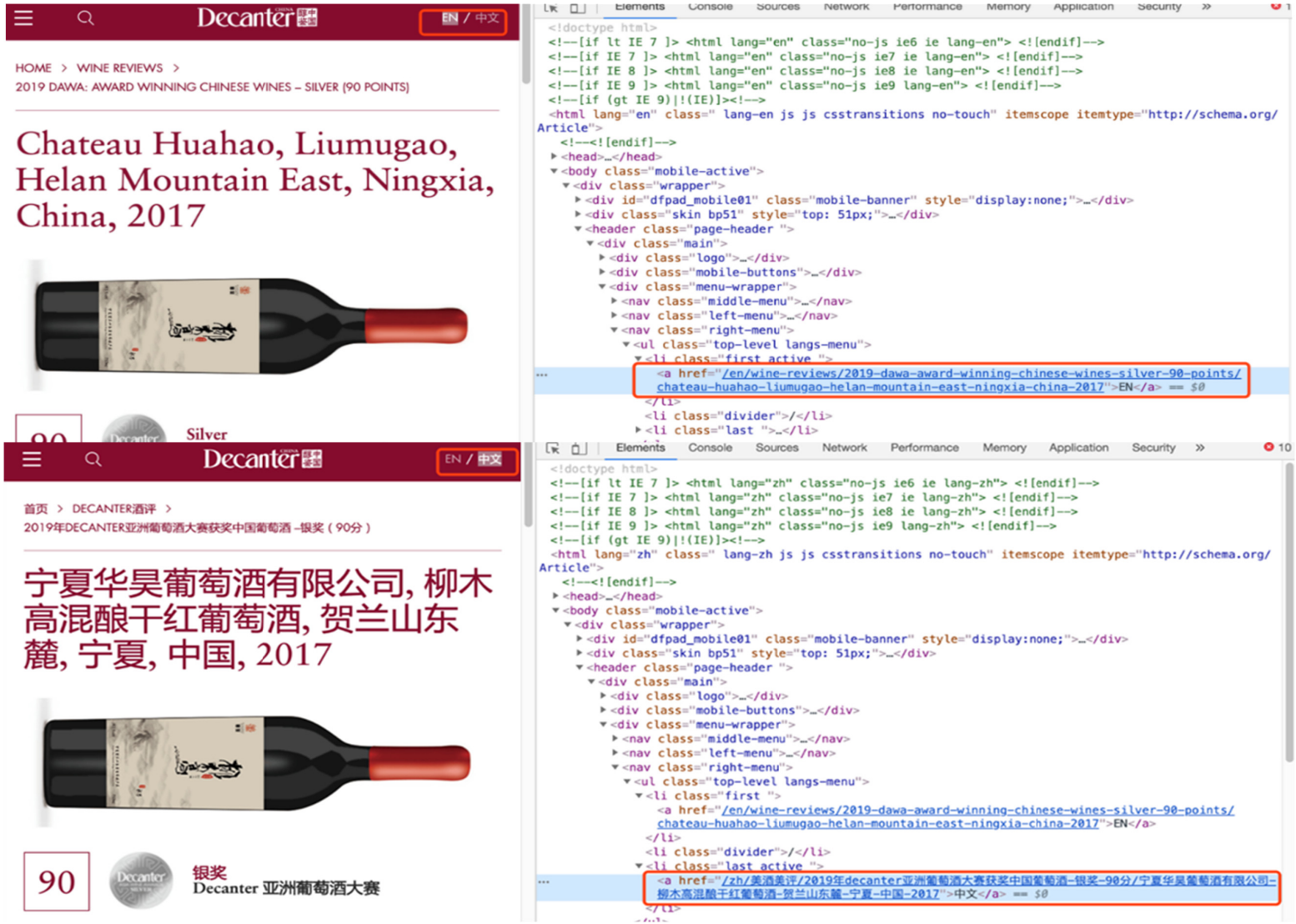
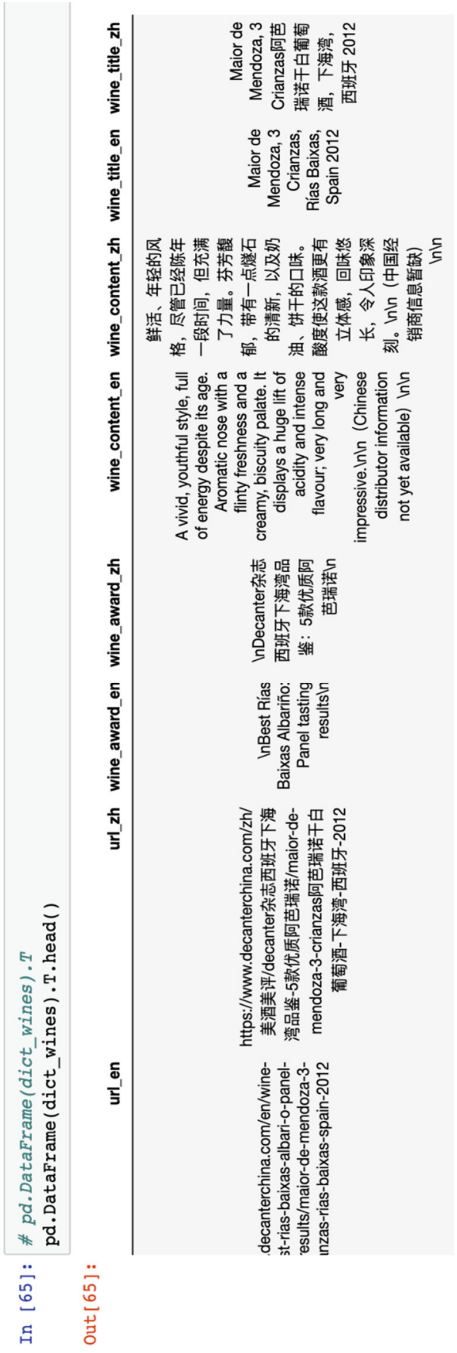


Figure 3: Data Crawl



```
In [65]: # pd.DataFrame(dict_wines).T
         pd.DataFrame(dict_wines).T.head()
```

Out[65]:

	url_en	url_zh	wine_award_en	wine_award_zh	wine_content_en	wine_content_zh	wine_title_en	wine_title_zh
	decanterchina.com/en/wine-st-rias-baixas-albari-o-panel-results/maior-de-mendoza-3-inzas-rias-baixas-spain-2012	https://www.decanterchina.com/zh/美酒美评/decanter杂志西班牙下海湾品鉴-5款优质阿芭瑞诺/maior-de-mendoza-3-crianzas阿芭瑞诺干白葡萄酒-下海湾-西班牙-2012	\nBest Rias Baixas Albariño: Panel tasting results\n	\nDecanter杂志西班牙下海湾品鉴：5款优质阿芭瑞诺\n	A vivid, youthful style, full of energy despite its age. Aromatic nose with a flinty freshness and a creamy, biscuity palate. It displays a huge lift of acidity and intense flavour, very long and very impressive.\n\n (Chinese distributor information not yet available) \n\n	鲜活、年轻的风格，尽管已经陈年一段时间，但充满了力量。芬芳馥郁，带有一点燧石的清新，以及奶油、饼干的口味。酸度使这款酒更有立体感，回味悠长，令人印象深刻。\n\n (中国经销商信息暂缺) \n\n	Maior de Mendoza, 3 Crianzas, Rias Baixas, Spain 2012	Maior de Mendoza, 3 Crianzas阿芭瑞诺干白葡萄酒，下海湾，西班牙 2012

Figure 4: Data Frame

2.2 Corpus Construction

The construction of our corpus is still in progress. The textual attributes of this corpus are from the title data we crawled, namely wine name, score, region, grape, producer, alcohol. These attributes will be valuable for future research on regression analysis of these attributes with wines. The word segmentation and POS tag were conducted by "Jieba" for Chinese texts and NLTK for English texts. So far, this parallel corpus is aligned at the paragraph level. We found it was difficult to establish the correspondence between the English texts and their Chinese translations at the sentence level because the wine reviews were rendered rather freely with translation methods like omission, addition, division and combination. We are seeking reliable means for sentence alignment in further studies. This corpus adopted the XML-based framework. The text head consists of the textual attributes and the text body is comprised of the wine reviews and the linguistic tags. This corpus contains 1211 aligned items of English-Chinese wine reviews with 149,463 Chinese characters and 66,909 English words up to now. Although Decanter China published the wine data on its website to the public, such commercial content is typically not easy for others to have the right for re-distribution. Thus, we are making an interface for people to access/search in the corpus for academic purpose only without openly sharing it. In addition, we are sharing one of our applications of this corpus, namely the English-Chinese bilingual oenology term bank (Chen, Quan, Wang, & Huang, 2020), which will be discussed in the following section.

3 The Application: Oenology Term Bank

3.1 Identifying the Key Words

We generated two word clouds of our parallel corpus of wine reviews in Chinese and English separately by Nvivo 12 Plus (Figures 5 and 6). The full lists of Chinese and English top 100 frequent words are in Appendices (cf. Tables 2 and 3). Figures 5 and 6 show that a number of the most frequently used words in the two languages do not match. For instance, there is not a single corresponding item in the Chinese word cloud for "Palate" in English, although, physically, "Palate" refers to 腭 è in Chinese.

Figure 5: Chinese Word Cloud

Figure 6: English Word Cloud

3.2 Key Words Translation

In order to examine the English-Chinese translation correspondence of certain words for the wine reviews, we used an alignment method to detect the word-pairs by the log-likelihood ratio estimation demonstrated in Rapp (1999). This calculation is based on the assumption according to the Distributional Hypothesis (Harris, 1954) – i.e. word meaning depends on its textual context. Hence, in the parallel corpus, if an English word and a Chinese word co-occur frequently in the parallel sentences, they are potentially good translation candidates for each other (Samuelsson & Volk, 2007).

English Word	Acceptable Translation	Top 10 Scored Candidates
Palate	风格 fēng gé (style); 口味 kǒu wèi (taste)	风格 fēng gé (style)；现在 xiàn zài (now)；舌尖 shé jiān (tongue tip)；非常 fēi cháng (very)；很 hěn (very)；口味 kǒu wèi (taste)；混酿 hùn niàng (blend)；这款 zhè kuǎn (this type)；带 dài (with)；酿造 niàng zào (brew)
Nose	香气 xiāng qì (scent)；鼻腔 bí qiāng (nasal cavity)	香气 xiāng qì (scent)；推荐人 tuī jiàn rén (recommender)；非常 fēi cháng (very)；款 kuǎn (type)；黑比诺 hēi bǐ nuò (Pinot Noir)；酒庄 jiǔ zhuāng (winery)；鼻腔 bí qiāng (nasal cavity)；潜力 qián lì (potential)；这款 zhè kuǎn (this type)；口味 kǒu wèi (taste)
Notes	nil	几年 jǐ nián (several years)；不满 bú mǎn (dissatisfied)； 木槿花 mù jǐn huā (hibiscus)；符合 fú hé (correspond)；水平 shuǐ píng (level)；月刊 yuè kān (monthly)；杂志 zá zhì (magazine)；明星 míng xīng (star)；坚固 jiān gù (firm)；甜椒 tián jiāo (sweet pepper)
Aromas	芬芳 fēn fang (fragrance)	强壮 qiáng zhuàng (strong)；黑色 hēi sè (black)；芬芳 fēn fang (fragrance)；黄油 huáng yóu (butter)；非常 fēi cháng (very)；完美 wán měi (perfect)；平衡 píng héng (balanced)；红果 hóng guǒ (red fruit)；十分 shí fēn (very)；草木 cǎo mù (vegetation)

Table 1: The Word with its Top 10 Candidates

Since the corpus is already aligned in terms of English-Chinese pairs at the paragraph level, we can directly process the word alignment part. The word-level co-occurrence frequency was calculated and a statistical test for the log-likelihood ratio was launched. We desired to do the sentence alignment, but the way to segment a paragraph into the sentences is often different between Chinese and English. We excluded the stop words in both English and Chinese for the purpose of better decreasing the noise. We calculated the log-likelihood ratio for every possible pair of English-Chinese words in the corresponding wine review and sorted them according to the log-likelihood score.

As a result, we automatically extracted a bilingual lexicon from the parallel corpus on wine reviews. Four most frequently used words in the word clouds – i.e., "palate", "nose", "notes" and "aromas" – are presented in this section with their translation equivalents. Plural forms are used for "notes" and "aromas" here, since their singular forms occur at very low frequencies in our data. **Table 1** lists each word with its top 10 scored candidates, and we manually selected the 'accepted translations' from the top 10 based on their potential to serve as optional translations in wine reviews. The full list of our result, i.e. the English-Chinese bilingual oenology term bank (Chen et al., 2020) can be viewed at https://drive.google.com/file/d/1LlDuU0euWKszq_WE1eUdzua25m6v3J4X/view?usp=sharing.

First, the literal translation of "palate" is "腭" è in Chinese, which sounds odd for reviewing wines in Chinese culture as native speakers do not directly mention this sensory part to describe their wine-tasting experience. There are two acceptable translations of "palate" in the candidates, namely "风格" fēng gé (style) and "口味" kǒu wèi (taste), which are rather free renditions. The differences in cultural and sensory experiences between English and Chinese favour this one-to-many mapping instead of the default one-to-one mapping. Second, the word "nose" can be either rendered very generally into "香气" xiāng qì (scent), or translated literally into "鼻腔" bí qiāng (nasal cavity), which preserves the semantic meaning of nose. Third, there are no acceptable translations of "notes" from the top candidates, and the log-likelihood scores of "notes" are not ideal. This points to a void in the Chinese lexis that corresponds to the meaning of "notes". Finally, "aromas" tends to be rendered into "芬芳" fēn fang (fragrance), a very acceptable translation that beautifully conveys the meaning of aroma/s.

Our appliance of the log-likelihood leads to rather effective identification of good translation candidates for the English words in wine reviews, e.g., the translations for "palate", "nose" and "aromas". However, there were also cases in which not a single acceptable translation can be found – e.g., "notes" – which strongly suggest cross-linguistic and cross-cultural differences in word choices for wine reviews. The log-likelihood tests demonstrate that, based on the key words in wine reviews, translation candidates can be generated that are potentially useful for rendering the terms from English into Chinese (cf. 'acceptable translations' in Table 1). The translation candidates need to be manually selected to suit various oenological contexts though. Moreover, this method can be applied to the studies of the one-to-many and further many-to-many bilingual term extraction in the domain of oenology regarding the cross-lingual and cross-cultural differences. It can also involve translation studies that look into translation strategies – e.g. literal versus liberal, semantic versus communicative, or foreignised versus domesticated translation – in dealing with the specific texts of wine reviews and the manipulation of translators.

Based on the automatic extraction of mapping terms, two major types of mapping of words across the languages emerged. The first type of mapping pertains to the words that have no precise translation equivalent/s (e.g. "palate" and

"nose"), and therefore paraphrasing and other freer translation methods tend to be used. The second type involves 'place-holder' translation, while the multiple mappings are mostly dependent on the *modifiers* of the word in question to express different meanings. The second type exhibits two sub-types. The first sub-type is those that have null term mapping (e.g. "notes"). The term is so generic and flexible to collocate with a rich repertoire of modifiers that it is considered as a noun that is semantically bleached and is usually not translated, since the meaning is conveyed by the modifier/s of notes. The second sub-type (e.g. "aromas") is still treated as semantically bleached but there is a corresponding 'light noun' – i.e., 芬芳 fēn fang 'fragrance' – for direct (one-to-one) term mapping. Our subsequent task is to sort out a solution to automatically classify the three different types and represent these three different types of bilingual term mapping in a term bank.

4 Conclusion

In this paper, we introduced our English-Chinese parallel corpus of wine reviews and described our preliminary attempt for the extraction of bilingual oenology term bank. Our study showed that the log-likelihood approach we chose can deal with the many-to-many mapping challenge posed by the nature of 'untranslatable' terms. Yet it does require significant human intervention – i.e., manual selection of the useful translation candidates that suit different oenological contexts. On the other hand, the current corpus size is too small to support deep learning approaches. In the subsequent studies we will enlarge the corpus and also adopt a sensory domain based (rather than term-based) mapping to attempt more revealing findings.

Acknowledgements

We are thankful to the two anonymous reviewers of this article for their valuable comments and suggestions. The authors would like to acknowledge the support of the research projects "A Comparative Study of Synaesthesia Use in Food Descriptions between Chinese and English" (G-SB1U) of The Hong Kong Polytechnic University and MYRG2018-00174-FAH of the University of Macau.

References

Caballero, R. (2017). From the glass through the nose and the mouth: Motion in the description of sensory data about wine in English and Spanish. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 23*(1), 66-88. https://doi.org/10.1075/term.23.1.03cab

Chang, B. (2004). Chinese-English parallel corpus construction and its application. In H. Masuichi, T. Ohkuma, K. Ishikawa, Y. Harada, & K. Yoshimoto (Eds.), *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation* (pp. 283-290). Tokyo, Japan: Logico-Linguistic Society of Japan.

Chen, X., Quan, S., Wang, X., & Huang, C. (2020). An English-Chinese bilingual oenology term bank, ISLRN 851-636-882-375-0.

Croijmans, I., Hendrickx, I., Lefever, E., Majid, A., & Van Den Bosch, A. (2020). Uncovering the language of wine experts. *Natural Language Engineering, 26*(5), 511-530. https://doi.org/10.1017/S1351324919000500

Demaecker, C. (2017). Wine-tasting metaphors and their translation: A cognitive approach. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 23*(1), 113-131. https://doi.org/10.1075/term.23.1.05dem

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Lim, L. (2018). A corpus-based study of braised dishes in Chinese-English menus. In S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang, & Y. Yao (Eds.),

Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 25th Joint Workshop on Linguistics and Language Processing (pp. 887-892). Association for Computational Linguistics.

Lim, L. (2019). Are TERRORISM and *kongbu zhuyi* translation equivalents? A corpus-based investigation of meaning, structure and alternative translations. In R. Otoguro, M. Komachi, & T. Ohkuma (Eds.), *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation* (pp. 516-523). Association for Computational Linguistics.

López-Arroyo, B., & Roberts, R. P. (2014). English and Spanish descriptors in wine tasting terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 20*(1), 25-49. https://doi.org/10.1075/term.20.1.02lop

Palmer, J., & Chen, B. (2018). Wineinformatics: Regression on the grade and price of wines through their sensory attributes. *Fermentation*, *4*(4), 84-93. https://doi.org/10.3390/fermentation4040084

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 519-526). College Park, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/1034678.1034756

Samuelsson, Y., & Volk, M. (2007). Alignment tools for parallel treebanks. *GLDV Frühjahrstagung*. Tübingen, Germany: Zurich Open Repository and Archive, University of Zurich. https://doi.org/10.5167/uzh-20448

Valente, C.C., Bauer, F.F., Venter, F., Watson, B., & Nieuwoudt, H.H. (2018). Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. *Scientific Reports*, *8*(4987), 1-13. https://doi.org/10.1038/s41598-018-23347-w

Appendices

The 100 most frequently occurring wine-tasting words in both Chinese and English reviews.

Appendix A. The Chinese Top 100 Frequent Words

Word	N of characters	Frequency	Weighted percentage (%)
风味	2	674	2.21
水果	2	504	1.65
香	1	468	1.53
酒	1	464	1.52
黑	1	413	1.35
款	1	387	1.27
非常	2	367	1.20
味	1	348	1.14
橡木	2	346	1.13
口感	2	331	1.08
单	1	324	1.06
浓郁	2	319	1.04
香气	2	300	0.98
味道	2	291	0.95
气息	2	290	0.95
成熟	2	277	0.91
芬芳	2	269	0.88
很	1	239	0.78
莓	1	235	0.77
酸度	2	229	0.75
中	1	219	0.72
余味	2	215	0.70
果	1	211	0.69
人	1	208	0.68
口味	2	208	0.68
葡萄酒	3	196	0.64
樱桃	2	190	0.62
般	1	181	0.59
黑色	2	179	0.59
风格	2	175	0.57
十分	2	164	0.54
感	1	162	0.53
香料	2	158	0.52
优雅	2	156	0.51
清爽	2	152	0.50
体	1	142	0.46

充满	2	142	0.46
复杂	2	140	0.46
气	1	133	0.44
甜美	2	133	0.44
清新	2	130	0.43
柔和	2	129	0.42
平衡	2	128	0.42
醋栗	2	128	0.42
好	1	124	0.41
李子	2	122	0.40
细腻	2	122	0.40
不	1	120	0.39
丝	1	119	0.39
咸	1	118	0.39
甜	1	117	0.38
滑	1	114	0.37
熏	1	113	0.37
辛	1	113	0.37
汁	1	110	0.36
令	1	106	0.35
分	1	106	0.35
烟	1	103	0.34
红	1	102	0.33
香草	2	99	0.32
迷人	2	98	0.32
纯净	2	97	0.32
淡淡	2	95	0.31
丰满	2	93	0.30
红色	2	93	0.30
美	1	93	0.30
带有	2	92	0.30
苹果	2	92	0.30
丰富	2	91	0.30
饱满	2	89	0.29
矿物	2	86	0.28
结构	2	86	0.28
活泼	2	85	0.28
出	1	82	0.27
巧克力	3	82	0.27
胡椒	2	81	0.27
还	1	81	0.27
陈年	2	81	0.27
柠檬	2	77	0.25
带来	2	76	0.25
更	1	76	0.25
紧	1	76	0.25
强劲	2	74	0.24
经典	2	74	0.24
具有	2	73	0.24
滋味	2	73	0.24
柑橘	2	72	0.24
氛	1	72	0.24
起来	2	71	0.23
会	1	70	0.23
口腔	2	70	0.23
绵长	2	70	0.23
饮	1	70	0.23
果香	2	69	0.23
回味	2	68	0.22
圆润	2	68	0.22
悠长	2	68	0.22
霞	1	68	0.22
完美	2	67	0.22
酿	1	66	0.22

Table 2: The Chinese Top 100 Frequent Words

Appendix B. The English Top 100 Frequent Words

Word	N of characters	Freq	Weighted percentage (%)
palate	6	692	3.01
fruit	5	588	2.56
nose	4	482	2.10
aromas	6	334	1.46
wine	4	311	1.35
oak	3	299	1.30
finish	6	294	1.28
ripe	4	266	1.16
tannins	7	264	1.15
black	5	232	1.01
notes	5	232	1.01
acidity	7	217	0.95
fresh	5	210	0.91
sweet	5	193	0.84
red	3	171	0.74
well	4	164	0.71
fruits	6	156	0.68
cherry	6	151	0.66
flavours	8	149	0.65
spice	5	144	0.63
style	5	135	0.59
dark	4	125	0.54
juicy	5	120	0.52
long	4	118	0.51

rich	4	115	0.50
elegant	7	104	0.45
cassis	6	103	0.45
savoury	7	103	0.45
vanilla	7	99	0.43
plum	4	98	0.43
fine	4	97	0.42
hints	5	92	0.40
lovely	6	91	0.40
full	4	88	0.38
bright	6	82	0.36
good	4	82	0.36
smoky	5	82	0.36
blackberry	10	79	0.34
character	9	78	0.34
apple	5	75	0.33
chocolate	9	75	0.33
clean	5	75	0.33
soft	4	74	0.32
texture	7	74	0.32
intense	7	72	0.31
floral	6	71	0.31
pepper	6	70	0.30
firm	4	69	0.30
touch	5	69	0.30
made	4	68	0.30
citrus	6	67	0.29
mineral	7	66	0.29
liquorice	9	64	0.28
spicy	5	64	0.28
balanced	8	63	0.27
bodied	6	62	0.27
complex	7	61	0.27
great	5	60	0.26
characters	10	59	0.26
shows	5	59	0.26
freshness	9	57	0.25
attractive	10	56	0.24
hint	4	56	0.24
peach	5	56	0.24
blackcurrant	12	55	0.24
green	5	55	0.24
structure	9	54	0.24
white	5	54	0.24
crisp	5	53	0.23
yet	3	53	0.23
creamy	6	52	0.23
dried	5	52	0.23
lemon	5	51	0.22
structured	10	49	0.21
berry	5	48	0.21
light	5	48	0.21
concentration	13	47	0.20
easy	4	47	0.20
pure	4	47	0.20
powerful	8	46	0.20
blueberry	9	45	0.20
herbs	5	45	0.20
lively	6	45	0.20
medium	6	45	0.20
cabernet	8	44	0.19
classic	7	44	0.19
concentrated	12	44	0.19
delicate	8	44	0.19
mouth	5	44	0.19
tannin	6	44	0.19
toasty	6	44	0.19
cherries	8	43	0.19
complexity	10	43	0.19
herbal	6	43	0.19
integrated	10	43	0.19
length	6	43	0.19
lime	4	43	0.19
spices	6	43	0.19
followed	8	42	0.18
dry	3	41	0.18

Table 3: The English Top 100 Frequent Words

Corpus-based Comparison of Verbs of Separation "Qie" and "Ge"

Nga-In Wu[1], Chu-Ren Huang[2], and Lap-Kei Lee[3]

[1]College of Professional and Continuing Education, Hong Kong Polytechnic University
ngain.wu@cpce-polyu.edu.hk
[2]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
churen.huang@polyu.edu.hk
[3]School of Science and Technology, The Open University of Hong Kong
lklee@ouhk.edu.hk

Abstract

This paper predicts that Chinese Synonyms *qie* and *ge* are verbs of separation and uses a variety of Chinese Word Sketch (CWS) functions to distinguish them. Several subtle differences are demonstrated in modifying relation and noun-verb relation, showing that the use of the two target words differ mainly in terms of the purpose of separation. Developing history is one of the factors why the use of *qie* and *ge* differ across the straits. These findings are more detailed when comparing with the work focusing on dictionary study. Obviously, traditional dictionary is no longer enough to Chinese language learners. This study is expected to provide some insights for Chinese dictionary editors and hence Chinese teachers.

1 Introduction

Many studies were done on near synonyms in Mandarin Chinese, and verb has been particular interesting to scholars (Wang and Huang, 2018). *Qie* 切 and *ge* 割 is one of the interesting pair of near synonyms. As a native speaker, semantic difference between the two words is not clear at the first glance. It is interesting to note that 切割 is acceptable, while 割切 sounds strange. This implies that there should be a semantic difference between two words, because pure coordination usually allows reversed order.

Studies on *qie* 切 and *ge* 割 have been done by Lian (2005) based on dictionary before. Yet, she failed to identify unique features of the two words. It may be because polysemy of words is not supported in a traditional dictionary. If we simply look at the definitions provided, it is not feasible to distinguish the difference between their usages, especially in different part of speech (Fillmore and Atkins, 1992). To fill the gap, Chinese Word Sketch (CWS) will be used in this article; CWS is a combination of Word Sketch Engine (Kilgarriff et al., 2005) and Chinese GigaWord Corpus (Huang et al., 2005). With the help of "computer-aided armchair linguistics" (Fillmore, 1992), it is believed that some common and unique features of the two words will be found, as the observations are based on large amount of authentic data. This method should be more efficient than relying on researchers' background knowledge merely to process the data (Li et al., 2018) and more reliable than studying the dictionary.

Our contribution. This paper tries to find out grammatical and collocational relations of *qie* 切 and *ge* 割, hoping to identify the differences and similarities between these two synonyms so as to figure out unique features and core meanings of the two words. Cross-strait comparison is also done, which aims to see how the use of two words differ in Mainland and Taiwan in view of different developing history and time. We expect that this study will provide insights to dictionary editing and writing.

Organization of paper. Section 2 states the research questions. Section 3 examines the meanings of the two words in dictionaries, the significant claim form Lain (2005), the classification suggested by Lian (2005) based on dictionary, and the frequency distribution in the Chinese Gigaword corpus. Section 4 and 5 are a cross-strait comparison and a summary.

2 Research Questions

This paper explores the research questions below:
(1) What are the grammatical and collocational relations of the target words found based on the Chinese Word Sketch results?
(2) Are there any unique features and core meanings for the two words? If yes, what are they?
(3) What are the differences on the usage of the two words in Mainland and Taiwan?

3 Dictionary-based and Corpus-based analysis

Dictionary-based analysis. As shown in Table 1, both *qie* 切 and *ge* 割[1]mean "to cut" in English. Ambiguity between two words is found when we refer to the definitions of the Contemporary Chinese Dictionary.

	新時代漢英詞典 New Age Chinese-English Dictionary	現代漢語詞典(第6版) The Contemporary Chinese Dictionary (The 6th Edition)
切 *qie*	Pronunciation 1: *qiē* to cut / to slice Pronunciation 2: *qiè* definitely / absolutely (not) / (scoffing or dismissive interjection)	Pronunciation 1: qiē 用刀把物品分成若干部份；真綫與圓、直綫與球、圓與圓、平面與球或球與球只有一個交點叫作切 Pronunciation 2: qiè 符合 / 貼近；親近 / 急切；殷切 / 切實；務必
割 *ge*	to cut / to cut apart	用刀截斷；分割；捨棄

Table 1: Explanation of *qie* and *ge* in dictionary

[1] According to the Contemporary Chinese Dictionary, there are two pronunciations for the word *qie* 切. Only *qie* 切 in the first tone giving similar meaning as 割 *ge* will be discussed in this paper.

Lian (2005) has tried to figure out the features of the two words (see Table 2) based on dictionary. However, the study failed to give real explanation to the two words. Lian (2005) used other near synonyms to explain and distinguish *qie* 切 and *ge* 割; Lain (2005) used *fen* 分 to paraphrase *qie* 切 and *zhe* 截 to paraphrase *ge* 割. Clearly, definitions in dictionary are not sufficient to tell the unique features of the two words.

Although Lain (2005) failed to give the real explanation of the two verbs, her claim gives a great implication to this paper (i.e. distinguishing meaning of words by using different paraphrases). When we paraphrase the verb *duan* 斷 in the Chinese classical poem *choudao duanshui shui gen liu* 抽刀斷水水更流 , *ge* 割 is acceptable. It is found that water is not really cut by knife, but separated. Therefore, this paper predicts that *qie* 切 and *ge* 割 are verbs of separation instead of just verb of cutting.

To see how the corpus data is useful on capturing the features so as to modify the definitions, the classification proposed by Lian (2005) will be adopted and discussed in this paper.

切 *qie*	割 *ge*
1. Tool used for the action – *Dao* 刀 *knife*	1. Tool used for the action – *Dao* 刀 *knife*
2. Process involved in the action – *Fen* 分 *separate*	2. Process involved in the action – Zhe 截 *cut*
3. Final state of object being cut because of the action – *Cheng ruogan bufen* 成若干部份 become several pieces	3. Final state of object being cut because of the action – *Duan* 斷 *separate*
4. Object being cut in the action – *Wupin* 物品 *product*	

Table 2: Analysis of *qie* and *ge* in the work of Lian (2005)

Gigaword corpus via CWS is used in this paper. We will present details of Gigaword corpus and the frequency distribution of the two words in this section.

Corpus-based analysis. Chinese Gigaword corpus data consists of three sub-corpora which are corpora coming from Central News Agency in Taiwan (CNA, 501,456,000 words), Xinhua News Agency in Mainland (XIN, 311,660,000 words) and Lianhe Zaobao in Singapore (Gigaword2zbn, 18,632,000 words). Table 3 shows the overall frequency and frequency of the two words in Gigaword2cna and Gigaword2xin. It is found that the overall frequency of *qie* 切 per million words is almost four times higher than *ge* 割. Also, the frequency of *qie* 切 is four times higher than *ge* 割 in Mainland and China. Based on the results, it is found that the use of *qie* 切 is dominant across the straits. Mainland and Taiwan share the same preference on the usage of *qie* 切.

	ge 割		*qie* 切	
Corpora	freq.	freq./ million	freq.	freq./ million
Gigaword2all	1352	1.63	831	6.09
Gigaword2cna	540	1.08	2019	4.03
Gigaword2xin	750	2.41	2821	9.07

Table 3: Frequency of *ge* and *qie* in corpora

The following sections find out the similarities and differences between *qie* 切 and *ge* 割 in terms of lexical grammatical relations, and the features are discussed and categorized according to the classification proposed by Lian (2005).

4 Grammatical Patterns Through Word Sketch

The Word Sketch function helps to illustrate the relations the target word has and the salient words within the relation. The *minimum frequency* is set at 5. Clicking *Show Word Sketch* and then inputting each word generate the result in Table 4.

	PP_給	Subject	Object	SentObject_of	Modifier	Modifies
qie	✓	✓	✓	✓	✓	✓
ge		✓	✓	✓	✓	

Table 4: Grammatical patterns of *qie* and *ge*

As shown in Table 4, there are more grammatical patterns for *qie* 切 than *ge* 割. This may explain why the frequency of the use of *qie* 切 is higher than the use of *ge* 割 as mentioned in the corpus-based analysis in Section 3. It is found that PP_給 and modifiers are relations absent for *ge* 割. When we set *minimum frequency* to 2, only *youshouwan* 右手腕 *right wrist* appears in the modifies relation for *ge* 割. Yet, PP_給 still does not appear in the prepositional relation for *ge* 割.

It is interesting to note that PP_將 appears for *ge* 割 at the *minimum frequency* of 2. The instance given below suggests that *ge* 割 is done for a particular goal. Feeding cow is the purpose. This observation suggests that feature "purpose of the action" should be added to the classification proposed by Lian (2005).

(1) 乾脆 分批 分期 將 麥子 割 了 餵 牛。
gān cuì fēn pī fēn qī jiāng mài zi gē le wèi niú
Simply in batches JIANG wheat cut ASP fed cow
'Simply harvest wheat by stages to feed cow.'

PP_ 到 and PP_ 把 are the other two prepositional relations appearing for *ge* 割 at the *minimum frequency* of 2. However, it is found that two instances containing PP_把 generated by the SkE are mismatches to *ge* 割. *Ba* 把 *set* does not appear as a prepositional relation collocating with *ge* 割, *Ba* 把 *set* is a classifier modifying *dao* 刀 *knife*. The two instances generated are as follows.

(2) 如同 三把 刀 強行 割 佔 了 中國 1 0 0 0 多 平方
Rú tóng sānbǎ dāo qiáng xíng gē zhàn le zhōng guó 1 0 0 0 duō píng fang
Like three CLASSIFER knife forcibly cut ASP China 1000 square meters.
Take 1000 square meters from China forcibly like three knives.

(3) 他們 再 用 一把 刀 割 我 的 長褲。
Tā men zài yòng yī bǎ dāo gē wǒ de cháng kù

They again use one CLASSIFER knife cut I DE trouser
They cut my trousers with a knife again.

5 Common Patterns and Only Patterns via Sketch Diff

The *Sketch Diff* function can compare and contrast two words in one time. It can help to find the common patterns and exclusive patterns of the pair of words. Table 5 is the Word Sketch Differences Entry Form. The default setting is used: the minimum frequency is 5; the maximum number of items in a grammatical relation of the common block is 25; the maximum number of items in a grammatical relation of the exclusive block is 12. After generating the results, this section will further explore the unique and common features of target words. To ensure the accuracy, mistakes like instances of PP_把 in Section 4 are removed from the results.

Corpus:	gigaword2all
First lemma:	割
Second lemma:	切
Sort grammatical relations:	
Separate blocks:	all in one block / common/exclusive blocks
Minimum frequency:	5
Maximum number of items in a grammatical relation of the common block:	25
Maximum number of items in a grammatical relation of the exclusive block:	12
	Show Diff

Table 5: Word Sketch Differences Entry Form

After inputting *qie* 切 than *ge* 割 and clicking *Show Word Diff*, the common and exclusive patterns of *qie* 切 than *ge* 割 are generated as shown in Table 6 and Table 7 below.

6 Common patterns of *qie* and *ge*

The colour chain generated can show the tendency (see Table 6). The words are highlighted in red and green. The greener the word means it has a higher tendency to collocate with *qie* 切. While the words are in red colour, it means that they have higher tendency to collocate with *ge* 割.

切 21 14 7 0 -7 -14 -21 割

英

Patterns	Frequency		Salience[2]	
	qie	*ge*	*qie*	*ge*
Subject	823	241	6.3	4.6
刀	10	31	23.6	45.9
人	9	10	4.2	9.0
Object	2065	880	4.0	4.2
塊	156	5	48.8	10.6
刀	69	30	48.0	38.9
傷口	6	6	15.4	17.8
手術	6	12	8.7	17.1
他	51	5	12.7	2.0
Modifier	329	229	1.9	3.3
能	9	18	10.6	18.0
也	21	17	15.2	14.8
要	8	14	9.2	15.0
不	20	7	14.8	8.0
再	6	6	9.0	10.0

Table 6: Common patterns of *qie* and *ge*

As shown in Table 6, *qie* 切 and *ge* 割 are similar in three aspects, i.e. they can have *ren* 人 *person* as a subject; they can have *shangkou* 傷口 *wound* as an object; and they can be modified by *ye* 也 *also*, *yao* 要 *necessity*, *bu* 不 *negation*, and *zai* 再 *again*. These results imply that the two words share a core meaning (i.e. a wound made by someone).

7 Only patterns of *qie* and *ge*

After focusing on the common pattern of the two words, this section focuses on exclusive patterns to figure out the unique features of the two words. As shown in Table 7, it is noticed that *qie* 切 and *ge* 割 differ in five grammatical relations including subject, object, modifier, sentObject_of and modifies. To facilitate the analysis, the five grammatical relations are categorized into two categories which are noun-verb relation and modifying relation. The following section focuses on noun-verb relation first.

Table 7: Only patterns of *qie* 切 and *ge* 割

7.1 Noun-verb relation

This category of grammatical relation refers to the relation between collocated words and target words which are served as subjects or objects and verbs respectively. As mentioned earlier, the modified Lian (2005)'s classification (i.e. with the new feature) is adopted for further analysis in the following sections.

Tool used for the Action. As shown in Table 6, it is found that *ge* 割 tends to collocate with knifes as a subject such as *ren* 刃 *blade*, *liandao* 鐮刀 *sickle* and *meigongdao* 美工刀 *utility knife*. It implies that *ge* 割 is used especially with knifes for specific purpose. In contrast, we only know that *qie* 切 collocates with *dao* 刀 *knife* as a subject according to Table 6, and it collocates with *yonglai* 用來 *used for* for the SentObject_of relation suggesting that a tool should be used. It implies that *qie* 切 can be used with any knifes as a subject.

Process involved in the action. It is noticed that food such as *dangao* 蛋糕 *cake*, *cai* 菜 *vegetable and shougao* 壽糕 *birthday cake* are objects collocated with *qie* 切. These objects imply that *qie* 切 refers to a fixed cutting method. On the other hand, organs or tissues (e.g. *shuangyanpi* 雙眼皮 *double eyelid*, *baopi* 包皮 *foreskin* and *hou* 喉 *throat*) and crops (e.g. *maizi* 麥子 *wheat* and *daozi* 稻子 *paddy*) are objects tend to collocate with *ge* 割. *Shuangyanpi* 雙眼皮 *double eyelid* and *hou* 喉 *throat* suggests that the cutting method is flexible which can be horizontal cutting, ring cutting and diagonal cutting.

Object being cut in the action. Meanwhile, the collocates (e.g. *cai* 菜 *vegetable and shougao* 壽糕 *birthday cake*) with *qie* 切 which are served as objects imply that the target being cut should be placed horizontally on a surface. The target of cutting should not be too small, as they can be cut into several pieces. In contrast, the collocates such as *shuangyanpi* 雙眼皮 *double eyelid*, *baopi* 包皮 *foreskin* for *ge* 割 acting as objects suggest that the cutting target can be small. *Maizi* 麥子 *wheat* and *daozi* 稻子 *paddy* suggests that *ge* 割 can be used when the target of cutting is standing upright.

Final state of object being cut. *Dangao* 蛋糕 *cake*, *cai* 菜 *vegetable and shougao* 壽糕 *birthday cake* act as object for *qie* 切 suggest that the actual amount of the cutting target should remain unchanged after the cutting process. The cutting target is cut into several pieces. In contrast, *shuangyanpi* 雙眼皮 *double eyelid*, *baopi* 包皮 *foreskin* served as objects for *ge* 割 suggest that a part is be removed and taken away. The actual amount of the cutting target should be different after the process.

Purpose of the action. Based on findings above, obviously, *qie* 切 is especially used in cooking context aiming to cut the target into several pieces and *qie* 切 can be done with any

kinds of knifes or tools with a fixed cutting method. *Qie* 切 is a verb which is result-oriented. On the contrary, it is found that *ge* 割 should be done with specific knifes for a particular purpose. Also, the aim of *ge* 割 can refer to the removal of a small part of the cutting target with a more flexible cutting method. The verb *ge* 割 is purpose-oriented. We can clearly see that this is a unique feature to the two target words, and it is rather abstract which can be explained and supported by other four features.

	Subject		Object		SentObject_of
qie	People	大家 We	Food	蛋糕 cake, 菜 vegetable, 壽糕 birthday cake	用來 Used for
	Proper nouns	費斯 Fisichella	Proper nouns	梅爾金 Chemerkin, 戈夫州 Chernihiv oblast,	
ge	鐮刀 sickle,美工刀 Utility knife, 刃 blade, 刀子 knife		雙眼皮 double eyelids, 麥子 wheat, 腕 wrist,喉 throat, 包皮 foreskin, 尾巴 tail, 盲腸 cecum, 稻子 paddy, 肉 meat, 麥 wheat		助 help

Table 8: Only patterns of noun-verb relation

7.2 Modifying relation

Different features can be found from their modifier and collocated modifies listed in Table 7. The features of the two words are as shown below.

Process involved in the action. Two collocates appear in the modifiers relation for *qie* 切, which are *yiqi* 一起 *together* and *yitong* 一同 *together* implying that the process can be done by more than one person at the same time, while *ge* 割 does not have such words in the modifiers relation. On the other hand, *qiangxing* 強行 *forcibly* appear in the modifier relation whereas *qie* 切 does not have such words in the modifiers relation. It implies that *qie* 切 is a more well accepted action, while people are forced to do the action when *ge* 割 is used. Moreover, *ge* 割 has a collocated modifier *yizhi* 一直 *continue* which implies that *ge* 割 is a continuing or repeating action, while *qie* 切 does not have such words. It implies that it is an action completed at once.

Purpose of the action. Similar to the previous section, it is found that *qie* 切 should be a result-oriented verb while *ge* 割 is a purpose-oriented verb. *Qu* 去 *for* appears in the modifier relation whereas *qie* 切 does not have such words in the modifiers relation. It infers that people do the action for a particular goal when *ge* 割 is used.

For modifies, while *gesheng* 歌聲 *song*, *xin* 信 *letter*, *banfa* 辦法 *method* and *qiuyuan* 球員 *player* are present in the modifies relation, *ge* 割 does not have this relation. This result is consistent with the findings in Section 4.

	Modifier	Modifies
qie	一起 together, 一同 together, 愈 more, 更 more, 共 sum, 各 separate, 並 also, 沒有 no	歌聲 song, 信 letter, 辦法 method, 球員 player
ge	著 continue, 強行 forcibly, 一直 continue, 去 go, 先 before, 就 so, 可以 can, 一 one, 都 also, 可 can	

Table 9: Only patterns of modifying relation

8 Cross-strait Comparison of *ge* 割 and *qie* 切 in CNA and XIN

Due to the frequent communication across straits, cross-strait comparison is worth discussing, and (Hong and Huang 2008; Hong and Huang, 2007) have already done some related studies. The use of vocabularies always depends on the context of texts, and the context may differ because of different culture, history, living habit and customs across the straits. To examine the actual use of

vocabularies, we can make use of Word Sketch function. As mentioned in Section 3, Chinese Gigaword corpus data is composed of three sub-corpora. Now, we would like to make use of two of them. They are corpora coming from Central News Agency in Taiwan and Xinhua News Agency in Mainland. First, we input *ge* 割 and *qie* 切 by using CNA and set the *minimum frequency* at 5. Then, we click *Show Word Sketch*. After that, same procedures are done again using XIN. The findings are then generated as shown in Table 9 and 10.

It is noticed that *ge* 割 and *qie* 切 differ mainly in collocations of the noun-verb relation. Therefore, only noun-verb relation is discussed in this section. This section focuses on *qie* 切 first. It is found that many subjects and objects of *qie* 切 in Mainland are transliteration; therefore, they will not be discussed in this paper. On the contrary, there is an intriguing finding which deserves more detailed analysis and explanation. *Shengyupian* 生魚片 *sashimi* is a dish, hence *ge* 割 should be preferred as this action is purpose-oriented. Yet, *qie* 切 includes *shengyupian* 生魚片 *sashimi* acted as object. It is believed that it may be because part of the noun (i.e. *pian* 片 *slice)* seems to require *qie* 切. *Qiepian* 切片 is acceptable, while *gepian* 割片 is unacceptable. There is a similar usage such as *qiezhangyupian* 切章魚片 *sashimi*. *Pian* 片 *slice* also requires *qie* 切 in this case

	Mainland (XIN)	Taiwan (CNA)
Noun-verb relation	Object: 梅爾金 Chemerkin, 戈夫州 Chernihiv oblast Subject: 費斯 Fisichella	Object: 蛋糕 cake, 菜 vegetable, 塊 piece, 壽糕 birthday cake, 麵店 noodle shop, 生魚片 sashimi, Subject: 她 she, 他 he, 人 person

Table 10: Comparison of *qie* between Mainland (XIN) and Taiwan (CNA)

As for *ge* 割, objects of it in Mainland are more diverse and much richer than Taiwan (see Table 10). The additional collocations in Mainland acted as objects are crops and classifier for fields (i.e. *mu* 畝 *classifier for fields*, *maizi* 麥子 *wheat*, *daozi* 稻子 *paddy* and *jiucai* 韭菜 *leek*), place (i.e. *Taiwan* 台灣 *Taiwan*), ideology (i.e. *weiba* 尾巴 *ideology*) and love (i.e. *qinqing* 親情 *family love* and *qing* 情 *love*). However, *ge* 割 includes only tissues or organs acted as objects: *hou* 喉 *throat*, *wan* 腕 *wrist*, *baopi* 包皮 *foreskin*, *mangchang* 盲腸 cecum, *shetou* 舌頭 *tongue* in Taiwan. *Shuangyanpi* 雙眼皮 *double eyelid* is the only collocation shared by both of them.

For subjects, *liandao* 鐮刀 *sickle* is the most frequent subject in Mainland, whereas it is *meigongdao* 美工刀 *utility knife* in Taiwan. *Liandao* 鐮刀 *sickle* is the tool for harvesting corps, but *meigongdao* 美工刀 *utility knife* are tools mostly used for crafts. It reflects that love, ideology, One-China principle, harvest and farming are still very important to Mainland. On the other hand, Taiwan people focus more on arts development and they are more open-minded. They are more willing to accept doing surgery. The differences can be attributed to the different political history, developing time and political stands of Mainland and Taiwan. Mainland is still more traditional, while Taiwan is more commercial.

Also, it is interesting to note that a new feature of *ge* 割 is found when place (i.e. *Taiwan* 台灣 *Taiwan*), ideology (i.e. *weiba* 尾巴 *ideology*) and love (i.e. *qinqing* 親情 *family love* and *qing* 情 *love*) acted as subjects in Mainland. Clearly, *ge* 割 is used for a particular purpose. More important, it is observed that the separation is not done by knife when these collocations appear. It simply implies that a part is separated from one entity. This result echoes to the prediction we make in section 3.

	Mainland (XIN)	Taiwan (CNA)
Noun-verb relation	Object: 麥子 wheat, 尾巴 tail, 雙眼皮 double eyelid, 肉 meat, 稻子 paddy, *qie* 切, 韭菜 leek, 畝 classifier for fields, 情 love, 台灣 Taiwan Subject: 鐮刀 sickle	Object: 雙眼皮 double eyelid, 喉 throat, 腕 wrist, 包皮 foreskin, 盲腸 cecum,舌頭 tongue Subject: 美工刀 utility knife

Table 21: Comparison of *ge* between Mainland (XIN) and Taiwan (CNA)

9 Conclusion

Our study shows that CWS is a powerful tool which can help discriminate the pair of synonyms *ge* 割 and *qie* 切. Huge amount of authentic linguistics data are generated by using various functions of SKE in an efficient way. Because of the rich data, detailed analysis and cross-straits comparison can be demonstrated in this paper. To conclude, it is predicted that *ge* 割 and *qie* 切 are verbs of separation implied by Lian (2005). Several differences are demonstrated in noun-verb relation and modifying relation. The subtle differences can show that the two target words differ mainly in terms of the purpose of separation. *Ge* 割 is a purpose-oriented verb, while *qie* 切 is a result-oriented verb. For similarities, both can be used when a wound is made by someone. These findings are far better than Lian (2005)'s work which focuses on dictionary study. Furthermore, the cross-strait comparison can help understand how Mainland and Taiwan differ from each other. It is believed the meaningful results obtained will facilitate cross-strait communication. Obviously, traditional dictionary is no longer enough to Chinese language learners. This study is expected to provide some insights for Chinese dictionary editors, and hence Chinese teachers. One possible future work direction is to figure out ways to improve the accuracy of results generated by CWS so as to provide more reliable sources for analysis and show how these findings are organized in a modern dictionary. The other possible way is to compare the results generated by CWS and SKE obtained in this study with the interesting results achieved with the semantic decomposition approach stated by Gao (2001).

References

Fillmore, C. J. (1992). "Corpus Linguistics" or "Computer-aided armchair linguistics". In J. Svartvik (Ed), *Directions in corpus linguistics.* (pp. 35-60). Berlin, Germany: Walter de Gruyter.

Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer & E. F. Kittay (Eds), *Frames, fields, and contrasts: New essays in semantic and lexical organization*, (pp. 75-102). New York, NY: Routledge.

Gao, H. (2001). *The physical foundation of the patterning of physical action verbs*. (Doctoral dissertation, Lund University, Sweden). Retrieved from https://lup.lub.lu.se/search/publication/693c9995-4857-4a77-b972-8852a3c93be3

Hong, J. F., & Huang, C. R. (2008). A corpus-based approach to the discovery of cross-strait lexical contrasts. *Language and Linguistics*, *9*(2), 221-238.

Hong, J. F., Huang, C. R., & Xu, M. (2007). *A study of lexical differences between China and Taiwan based on the Chinese gigaword corpus*. Paper presented at 19th Conference on Computational Linguistics and Speech Processing, ROCLING 2007, Taipei, Taiwan.

Huang, C. R., & Hong, J. F. (2005). Deriving conceptual structures from sense: A study of near synonymous sensation verbs. *Journal of Chinese Language and Computing*, *15*(3), 125-136.

Kilgarriff, A., Huang, C.-R., Rychly, P., Smith, S., & Tugwell, D. (2005). Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.

Li, L., Huang, C. R., & Gao, X. (2018, May). A SkE-assisted comparison of three "prestige" near synonyms in Chinese. In J. F. Hong, Q. Su & J.S. Wu (Eds.). *Chinese Lexical Semantics*, (pp. 256-266). Chiayi, Taiwan: Springer, Cham.

Lian, X. (2005). To understand the semantics fuzziness of verbs from "qie, ge, jie". *China Academic Journal*, *19*(3), 62-64.

Wang, X., & Huang, C. R. (2018, May). From Near Synonyms to Power Relation Variations in Communication: A Cross-Strait Comparison of "Guli" and "Mianli". *Chinese Lexical Semantics* (pp. 155-166). Chiayi, Taiwan: Springer, Cham.

Association between declarative memory and language ability in older Chinese by education level

Xie Chenwei
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
Hong Kong Polytechnic University
chenwei7.xie@connect.polyu.hk

Feng Yun
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
Hong Kong Polytechnic University
yunr.feng@connect.polyu.hk

William Shi-Yuan Wang
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
Hong Kong Polytechnic University
Department of Electronic Engineering
Chinese University of Hong Kong
shiyuan.w.wang@connect.polyu.hk

Abstract

Episodic memory and semantic memory are two subsystems of declarative memory which is considered to be related to language ability. We investigated the interdependence of the episodic and semantic memory and their association with syntactic complexity in two groups of Chinese older adults with higher and lower education levels. The results indicate that episodic memory and semantic memory are significantly correlated with each other, but only episodic memory shows a significant relationship with syntactic complexity. Educational attainment has a substantial influence on the performance of memory tasks, but no influence on the syntactic complexity. The study provides new evidence from Chinese older adults for clarifying the association between declarative memory and syntactic processing and the benefits from education on memory preservation in later life.

1 Introduction

Nowadays, China is facing an aging population. According to the general definition of an aging society published by the United Nations in 1956[1], when a country or region's population aged 65 and above accounts for 7% of the total population, the country or region enters the aged stage. China has around 88 million elderly people aged 65 and above as early as 2000, accounting for 7% of the total population, which means that China officially entered the aging society two decades ago. For the past over ten years, according to the National Bureau of Statistics of China[2], this number has continued to climb and exceeded 160 million by the latest 2018, which has nearly doubled from 2000 and occupied 11.9% of the entire population. Previous studies suggested that older adults are vulnerable to memory decline (for a review, see Buckner, 2004) and language attrition (for a review, see Burke and Shafto, 2011). However, since most of these studies have been based on the WEIRD population (Henrich et al., 2010a; Henrich et al., 2010b), it remains to be established what changes

[1] Department of Economic and Social Affairs, United Nations, *The Aging of Populations and Its Economic and Social Implications* (New York: United Nations, 1956), 7.
[2] For details, see http://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0305&sj=2017.

Chinese older adults undergo in terms of memory ability and language usage, and how memory ability and language usage interact with each other in older Chinese.

1.1 Declarative and Procedure Model

Memory is not a single process but has at least two major forms, namely, declarative memory and procedure memory which are both associated with language ability, as proposed in the declarative and procedure model. (Ullman, 2001a, 2001b, 2004). Declarative memory is an explicit and conscious recollection of arbitrarily related information, such as the people, places, and objects in specific events and the general knowledge about facts. On the contrary, procedure memory functions implicitly and unconsciously, in which habits and skills are computed automatically. In terms of language usage, the declarative memory serves the mental lexicon of memorization of the word-specific knowledge, whereas the procedure memory assists the mental grammar of the rule-governed complex representations. Besides, with regard to anatomical structures, long-term storage of procedure memory involves in the frontal cortex and basal ganglia circuits, especially the striatum, while long-term storage of declarative memory requires the hippocampus and the medial temporal lobe of the neocortex which also subserve language processing (Piai et al., 2016; Ullman, 2014).

Declarative memory and procedure memory alter in different trajectories across the lifespan (for a review, see Hedden and Gabrieli, 2004). The procedure memory barely exhibits age-related declines, whereas declarative memory, particularly encoding new memories of episodes or facts, declines evidently in the late years. A piece of robust evidence is that the preliminary phases of Alzheimer's disease are hallmarked by shortfalls in declarative memory, for instance, the older adults are only able to store fewer number of objects or specific events than the younger adults for a long period of time (Buckner, 2004; Lai and Lin, 2013). Besides, brain regional changes in volume are not uniform pertaining to declarative memory and procedure memory. Striatal volume declines by about 3% per decade at a span between 20 and 80 years of age (Gunning-Dixon et al., 1998), whereas hippocampus and the parahippocampal gyrus shows a 2–3% per decade decline in the volume (Raz et al., 2004), and might extend to a 1% per annual decline after the age of 70 (Jack et al., 1998), which would likely deteriorate the neural mechanical foundation of declarative memory and related language ability.

What's more, regarding the function in language processing, declarative memory and procedure memory implicate different alterations in older adults. Kempler et al. (1987) demonstrated that syntactic processing and semantic processing are separately operating in patients with Alzheimer's disease. They required subjects to perform spontaneous speech tasks and write to dictation tasks, and analyzed their syntactic complexity and errors, as well as semantic errors. Eventually, they found that compared to the semantic ability which is likely linked to the declarative memory, the syntactic ability is selectively conserved due to its automatic nature which is related to the procedure memory. The syntactic ability seems to remain intact in the late years of life, but it lacks close inspections in the Chinese population.

In addition to the distinct functions of declarative memory and procedure memory, it is noted that they could also cross the boundary in language processing (Ullman, 2014). Complex structures, such as "walked" can be an example to illustrate this point. In procedure memory, "walked" could be computed into the composition of the original word "walk" and English regular past tense "-ed", which can be stored based on its rule-based grammatic foundation. However, it could also be acquired and stored as a chunk in the declarative memory. This is the so-called compensatory role of declarative memory to procedure memory (Ullman et al., 2020; Ullman and Pierpont, 2005), which is widely found in patients with neurodevelopmental disorders (Evans and Ullman, 2016; Lee et al., 2020; Lum et al., 2012; Ullman and Pullman, 2015; Walenski et al., 2014), but scarcely investigated in the aging population (Rieckmann and Bäckman, 2009). In the current research, we aimed to investigate whether and how declarative memory associates with the rule-based syntactic process, especially the syntactic complexity in Chinese older adults via delving into its two component subcategories, namely episodic memory and semantic memory as described below.

1.2 Episodic Memory and Semantic Memory

Declarative memory can be divided into episodic memory and semantic memory (Greenberg and Verfaellie, 2010; Tulving, 1972). Episodic memory refers to the collection of specific events or experiences which entails the personal realization of what was happening and can be reconstructed vividly with many details from the individuals' perspective (Garrard et al., 1997). That is to say, episodic memory includes the exact contextual information, such as times, locations, persons, and related feelings, and most individuals regard themselves as the actors or contemporary witnesses in these autobiographic events. Therefore, episodic memory involves the emotional charge and the whole context surrounding an occurrence, rather than the bare facts of the occurrence itself. By contrast, semantic memory is a more structured record and embraces the recollection of general information of facts and concepts about the external world which shares with others and involves no personal emotional charge (Squire and Zola, 1998). Much of semantic memory is abstract and is connected with the meaning of verbal symbols.

The story recall task is considered as a common method to assess the episodic memory (Baudic et al., 2006; De Anna et al., 2008; Hertzog et al., 2003), whereas the verbal fluency task has been widely used to test the semantic memory (Chertkow and Bub, 1990; Henry et al., 2004). Moreover, Kavé and Sapir-Yogev (2020) put forward shared mechanisms for both delayed retrieval of story information on story recall tasks and retrieval of words on verbal fluency tasks, which also functioned in the older adults. They asked subjects to complete tasks of verbal fluency and recollect the story that they heard after a delay of around 30 min. The results showed that there are significant correlations between semantic fluency and delayed story recall, and the delayed story recall significantly contributes to semantic fluency performance when conducting the regression analysis. Episodic and semantic memory seems to be interdependent and affect each other both at encoding and at retrieval (Greenberg and Verfaellie, 2010), but this is less examined in the Chinese population.

Furthermore, recent neural research exhibits that hippocampal theta oscillations are discovered during online language processing in addition to memory function, which suggests that the hippocampal complex contributes to a shared neurophysiological mechanism between language and episodic memory (Piai et al., 2016; Pu et al., 2020). However, it remains unclear whether and how episodic memory and semantic memory interact with language usage, and in the present study, we aimed to elucidate the association between episodic memory and syntactic complexity, and the relationship between semantic memory and syntactic complexity as well.

1.3 Education Effects

Although all humans cannot escape from aging, the trajectories of both their memory decline and language attrition proceed dissimilarly in which education plays an important role. Anatomically, education has a profound influence on the aging of specific hippocampal subfields, for instance, there are significant negative correlations between educational attainment and the atrophy of hippocampal CA2/3 in older men (Jiang et al., 2019). On the other hand, education appears to have positive effects on declarative memory in older adults (Reifegerste et al., 2020). For instance, when performing retention, recall, and recognition tasks in word memory tests, less educated healthy older adults showed significantly decreased scores and lower response rates in all tests, compared to the higher-educated equivalents (de Azeredo Passos et al., 2015). This also happened in patients with Alzheimer's disease that those attained lower education levels obtained fewer scores of recall tasks in the mini-mental state examination test (MMSE) (Delpak and Talebi, 2020). And language ability is one of the most impaired domains for these patients with low education level in the above MMSE test. Besides, educational attainment is also associated with semantic memory (Gladsjo et al., 1999) and episodic memory (Zahodne et al., 2019) among non-Chinese speaking older adults.

However, it is still unclear whether a higher education level is a protective factor against age-related memory decline and language attrition for the Chinese population, and how educational attainment affects the trajectories of memory decline and language attrition in Chinese older adults due to the scarce research. Therefore, the current study aimed to seek whether and how Chinese older adults with higher and lower

education levels perform differently on memory processing and language usage.

1.4 The Present Study

The current study investigated whether the language ability of Chinese older adults, especially the ability of syntactic processing, is affected by declarative memory, or put it in another way, by the episodic memory and semantic memory. In order to answer this question, delayed story recall tasks and verbal fluency tasks were utilized. Moreover, we also explored whether the performance of delayed story recall is related to verbal fluency on Chinese older adults, since both episodic memory and semantic memory belong to declarative memory. Besides, educational attainment was designed as a controlling factor in order to explore whether and how education level impacts the memory decline and language attrition on Chinese older adults.

2 Methods

2.1 Participants

We recruited 36 participants, of whom eighteen are from Shenzhen (SZ participants) and speak Mandarin daily. The other eighteen participants were recruited in Chenzhou (CZ participants), and speak Chenzhou dialect[3] in daily life. There were 6 males and 12 females in SZ cohorts, and 11 males and 7 females in CZ cohorts. SZ participants and CZ participants were matched on age, but differed in education level. The SZ participants varied in age from 61 to 74 years with an average age of 66.0 ± 4.9 years, while CZ varied in age from 63 to 79 years with an average age of 69.4 ± 4.4 years. There was no statistically significant difference between SZ and CZ participants on age (t = -1.637, p = .119). By contrast, the SZ participants varied from 15 to 21 years with a mean of 16.3 ± 1.8 years on the level of education, and the CZ participants varied from 2 to 11 years with a mean of 7 ± 2.8 years on the educational attainment. There was a statistically significant difference between SZ and CZ participants on the level of education (t = 8.9078, p < .001). Besides, for both two groups of participants, if they had a past history of brain traumatic injury, stroke, clinical depression, alcoholism, and vision and hearing problems, they were not included in the study. The experiments were approved by The Hong Kong Polytechnic University's Human Subject Ethics Subcommittee. Prior informed consent was obtained from all participants.

2.2 Psycholinguistic Measures

Story Recall: The delayed story recall is an efficient assessment for episodic memory, and it is also included in the frequently uesd Wechsler Memory Scale-Fourth Edition of Chinese version (adult battery)[4]. In that battery, subjects are required to recollect the story that the experimenter read for them. That is to say, subjects might mimic the sentences and syntactic structures that the experimenter used. However, in the current study, the syntactic ability is what we plan to survey. Thus, recalling a written story is not suitable to inspect participants' syntactic ability. Instead, we asked participants to watch a six-minute movie of pear story and then verbalize what they have remembered, which requires subjects to create their own speech production to describe the story, rather than simply recall the sentences appeared in the story essay. In this way, the subjects' language usage and memory ability can be measured at the same time.

The six-minute movie is developed by Wallace Chafe in 1975, and it shows a farmer harvests pears in the morning, and a boy who passes on a bike steals a basket of pears. And then the boy experiences falling from the bike and getting help from other children, before the farmer finds he loses some pears (Chafe, 1980). After subjects finished watching the movie, they were asked to recall the story immediately. And after 30 minutes, they were asked to recall the story again. The syntactic complexity of their speech production and the number of information units they recalled were analyzed in both immediate recall and delayed recall.

Verbal Fluency: Verbal fluency tasks were applied to measure participants' semantic memory (Kavé and Sapir-Yogev, 2020), in which participants were asked to produce as many words as possible within one minute on one semantic

[3] Chenzhou dialect is subordinate to Southeastern dialect, which is similar to Mandarin Chinese in syntax (Shan 单泽周, 1997).

[4] Refer to Wang 王健 et al. (2015) for more details.

category. There were four different semantic categories, namely, fruit, tool, Chinese city, and profession. The first two categories belong to the concrete domain, while the latter two categories belong to the abstract domain. The total number of items of four categories was counted, and the unrelated and repeated words were excluded.

2.3 Procedure

Participants first provided their basic demographic information, including their age, language experience and education level, and then finished the Montreal Cognitive Assessment test (Nasreddine et al., 2005) which implies their mental states. Then participants began to watch the pear story film and immediately verbalized what they had remembered. After around 30 minutes, they were asked to complete the delayed recall tasks. During the interval, participants worked on some other nonverbal cognitive tests to distract their attention from the story recall task. Afterward, participants started to complete the verbal fluency tasks. The whole testing session lasted around 1 hour. Participants completed all tasks with an experimenter individually and had a rest when feeling tired.

2.4 Data Analysis

Information Unit: Information unit is an index to measure the participants' episodic memory, which involves "actors, activities, objects, numbers, days, times, or places" (Kavé and Sapir-Yogev, 2020), and there are 79 information units in total in the pear story film[5]. Participants' recalled stories were transcribed verbatim, and all information units were screened based on the uniform standard. At last, five types of information units were obtained, which are valid information unit, missing information unit, mismatch information unit, unrelated information unit and wrong information unit, and they have different states in scoring, as shown in appendix A of the data from one participant.

The first type is the valid information unit that correctly describes the sceneries in the pear story film. For instance, the participant recollected the information unit that the rooster crowed both in immediate recall and delayed recall, so the participant can obtain 1 score both in immediate recall and delayed recall[6]. The second type is the missing information unit which is the plot that the participant failed to mention in both immediate recall and delayed recall. For example, in both immediate recall and delayed recall, the participant missed the information unit of two baskets of pears which is quite an important clue, so in both of them, the participant obtained 0 scores.

The third and fourth situations in appendix A demonstrate the mismatch information unit which means the participant recollected the sceneries correctly only in one version of recall or missed the plots in one recall task, so only the recall task that the participant succeeded to evoke and describe accurately can obtain 1 score. In the third situation, the participant recollected the information unit that the boy who takes pears turns his head back in the immediate recall but missed it in the delayed recall, so this participant obtained 1 scores in the immediate recall but 0 score in delayed recall. On the contrary, in the fourth situation, the participant retrieved the information unit of how the boy left only in the delayed recall but failed to recall this information unit in the immediate recall and replaced it with the unrelated information unit of saying thank you. The last two types are the unrelated information unit and the wrong information unit, and both of them obtain 0 scores, since both of them reflect that the participants formed nonsensical episodic memory. For instance, the participant utilized unrelated information unit of saying thanks and wrong information unit of riding the bike to mistakenly substitute the correct information unit of pushing the bike.

Syntactic Complexity: In order to assess the syntactic ability of the participants, the syntactic complexity of speech production from the story recall task was analyzed. After meticulously transcribing all the speech utterances, 20 sentences were selected for each participant, in which 10 sentences were chosen from the immediate recall task and the other 10 sentences were chose from the delay recall task and both of them were scanned and selected from the beginning of the

[5] The participants' recall data were taken into consideration when counting the total information units of the pear story.

[6] Actually, the participant can obtain 2 scores for both immediate recall and delayed recall, as he mentioned the information unit of one morning in the countryside. Here we marked one score for simplistic demonstration and applied the same way on the following illustration.

recall production. We selected these 20 sentences, rather than all sentences, in order to make sure that the same number of sentences were finally obtained to be analyzed, since participants produced inequable number of sentences in their recalls. Furthermore, incomprehensible sentences were excluded, such as sentences with fatal syntactic or semantic errors. Before performing syntactic analysis on these transcribed sentences, the redundant elements, such as phonetic error, filler, and nonsense repetition, were deleted from the original sentences, as shown in the below.

Original sentence:

呃	那个	果农	摘	梨	的	那个
filler	that	farmer	pick	pear	auxiliary word	that
[s]	那个	那个	方法	挺	有意思	。
phonetic error	repetition	repetition	method	quite	interesting	.

Refined sentence:

那个	果农	摘	梨	的	那个
that	farmer	pick	pear	auxiliary word	that
方法	挺	有意思	。		
method	quite	interesting	.		

Translation:

The method that the farmer applies is quite interesting.

After dispelling the redundant elements in the original sentence, we obtained the refined sentence to conduct syntactic complexity analysis. According to Chao (1968)'s theory, there are five syntactic types of smaller constructions within one sentence, which are coordination, subordination, verb-objection(V-O) construction, verbal expressions in series, verb-complement(V-R) construction. These basic syntactic constructions can be "iterated and/or combined" to generate a more complicated structure. Based on these criteria, in the above example, 那个果农, 那个方法, 摘梨的那个方法 and 挺有意思 are four subordinations, and 摘梨 is a V-O construction, and eventually, this sentence contains five syntactic constructions and scores 5 points in total.

Speech Rate: In addition to the syntactic complexity, we also analyzed the speech rate of speech production in story recall tasks. Speech rate reflects how fast the participants speak and it is positively related to the number of sinograms and negatively related to the duration of their speech production. We firstly assessed the total number of sinograms the participants produced in both immediate recall and delayed recall and how much time they spent, and then calculated the speech rate by dividing the total number of sinograms by the total duration.

Information unit analysis and syntactic complexity analysis were performed by one trained rater who is a native speaker of both Mandarin and Chenzhou dialect. All statistical analyses were conducted by R program language.

3 Results

3.1 Performance Comparison between Two Groups

Firstly, we calculated the mean scores of SZ and CZ participants' performance in all tasks, as shown in table 1.

Group	Verbal Fluency	Duration(S)	Sinograms	Speech Rate	Immediate Recall	Delayed Recall	Syntactic Complexity
SZ	65.4	199.4	814	4.1	28.4	30.8	5.72
CZ	38.6	52.6	204	3.9	14.9	12.8	5.26

Table 1. Average scores of SZ and CZ participants

Through independent t-test, we found that there were significant differences between SZ and CZ participants in verbal fluency performance (t = 7.0224, p < .001), duration of story recall (t = 4.535, p < .001), the number of sinograms in story recall (t = 3.778, p = .003), the number of information units in immediate recall (t = 4.421, p < .001), and the number of information units in delayed recall (t = 5.895, p < .001). In contrast, no significant difference was found between these two groups in speech rate of story recall (t = 0.791, p < .441) and syntactic complexity (t = 1.255, p = .2264). That is to say, education level would affect subjects' results of verbal fluency tasks and delayed recall tasks but have limited influence on the speech production pertaining to the speech rate and syntactic complexity of their utterances.

3.2 Correlation Analysis

In order to examine the association between semantic memory and episodic memory, we conducted a correlation analysis on the results of verbal fluency tasks and story recall tasks. Also, to examine the relationship between syntactic processing and declarative memory, and the role of educational attainment, we added syntactic complexity and education years in the correlation analysis, as shown in table 2.

Task	Verbal fluency		Syntactic complexity		Education years	
	r	*p*	*r*	*p*	*r*	*p*
Immediate recall	0.829	< .001***	0.577	0.008**	0.788	< .001***
Delayed recall	0.847	< .001***	0.498	0.025*	0.836	< .001***
Verbal fluency	-	-	0.382	0.097	0.863	< .001***
Syntactic complexity	0.382	0.097	-	-	0.369	0.1094

* $p < .05$; ** $p < .01$; *** $p < .0001$.

Table 2. Correlations between verbal fluency tasks and story recall tasks

It is noted that verbal fluency performance has a significant strong positive correlation with the numbers of information units in both immediate recall and delayed recall, but not significantly related to syntactic complexity. However, syntactic complexity is significantly correlated with the numbers of information units in both immediate recall and delayed recall, which shows a moderate positive corelation. Besides, education level has a significant strong positive correlation with the verbal fluency performance and the numbers of information units in both immediate recall and delayed recall, but is not significantly related with syntactic complexity (r = 0.369, p = .1094), as mentioned in the table 1 before that SZ participants and CZ participants show no significant difference in syntactic complexity.

3.3 Regression Model of Syntactic Complexity

In order to further explore the association between episodic memory and syntactic processing, a simple linear regression was run to predict participants' syntactic complexity based on their information units recollected in delayed recall[7], as shown in the table 3.

	β	*SE*	*t*	*p*	R^2
Intercept	4.696	.366	12.841	< .001***	.2479
Delayed Recall	.036	.015	2.436	.0255*	

* $p < .05$; ** $p < .01$; *** $p < .0001$.

Table 3. linear regression predicting syntactic complexity

It is obvious that a significant regression equation was found (F (1, 34) = 5.932, p = .02549), with an R^2 of 0.2479. Participants' predicted syntactic complexity is equal to 4.696 +0.036 (information units). Participants' average syntactic complexity increased by 0.036 for each information unit, which means the information unit can significantly predict the variation of syntactic complexity of speech production in Chinese older adults.

4 Discussion

Our results show significant associations between performance on the verbal fluency tasks and the delayed story recall tasks, which is consistent with the results of previous studies (Kavé and Sapir-Yogev, 2020). Verbal fluency tasks and delayed story recall tasks are two measurements reflecting semantic memory and episodic memory respectively. As both semantic memory and episodic memory are subordinate to declarative memory (Squire and Zola, 1998; Tulving and Schacter, 1990), and both of them vary along a continuum of processing and stored by the similar brain regions (Rajah and McIntosh, 2005), it is not surprising that they are correlated with each other. Besides, episodic memory can convert into semantic memory after accumulated repetition (Garrard et al., 1997), which also underlines their connection.

Previous studies have demonstrated that declarative memory is considered to be less connected to syntactic processing (Ullman, 2001a, 2001b, 2004, 2013, 2014, 2016), which is

[7] The independent factor of the results of the immediate recall was excluded, because it is partly confounded with the status of working memory (Reifegerste et al., 2020).

consistent with the results of semantic memory performance but inconsistent with the results of episodic memory performance in the current study, as the number of information units in delayed recall contributes to the prediction of the syntactic complexity in online speech production, but dissociates from the number of generating words in verbal fluency tasks. This is probably because most previous studies focus on syntactic correctness judgment, which is, to some extent, distinct from online speech production in the present study. For instance, Pu et al. (2020) found that hippocampal theta power which is connected with episodic memory specifically associates with semantic errors rather than syntactic errors in perception modal. Indeed, readers might ignore the syntactic incorrectness in a sentence in order to obtain the whole meaning of the sentence, but it seems not to prove that they are not able to produce syntactically complicated natural sentences in language tasks. Another point is that even though declarative memory plays a compensatory role to procedure memory in syntactic processing (Ullman et al., 2020; Ullman and Pierpont, 2005), it appears to be only applied to the episodic memory. It needs more meticulous studies to untangle the complicated relationship between episodic and semantic memory and syntactic ability.

Another interesting phenomenon is that SZ participants and CZ participants showed no significant differences in syntactic complexity in the story recall utterances, although they have significant differences in education level. That is to say, educational attainment does not affect speakers' syntactic manipulation in speech production. Language enables its users to exploit syntactic processing equally, and syntactic ability keeps intact from other external factors. However, it does not mean all speakers can take full advantage of this ability to complete relevant tasks perfectly. Previous studies in Chinese adults have claimed that different educational attainments would lead to different patterns of neural activation associated with language tasks (Li et al., 2006), and different results in verbal fluency tasks (Mok et al., 2004). In verbal fluency tasks, CZ participants who attained lower education levels generated significantly fewer items than SZ participants. Low-educated participants appear to experience semantic memory deficits (Yang et al., 2006). Besides, in story recall tasks, SZ participants verbalized longer descriptions and more sinograms in their recall than CZ participants, as CZ participants sometimes failed to articulate some plots in recall and caused fewer number of sinograms. Even though both two groups had similar syntactic complexity in their utterance, they performed differently in language tasks due to their disparate educational attainments. It suggests that education plays an important role in moderating the weakness of age-related cognitive decline, especially memory decline. The limitations of this study was the wide range of age for both groups as memory loss is vulnerable to age differences, and that we did not correlate language and memory performance with their occupations as different professions involve in different magnitude of cognitive activities and affect their performance.

5 Conclusion

The current study recruited two groups of Chinese older adults with different education levels to perform verbal fluency tasks and story recall tasks in order to investigate the association between declarative memory and language usage. The results show that verbal fluency performance is significantly related to the number of recollected information units in both immediate recall and delayed recall, which suggests that the two subcategories, semantic memory, and episodic memory are closely connected with each other. However, only episodic memory shows a significant association with syntactic complexity in the story recall utterances. Furthermore, education level plays a large role in verbal fluency tasks and story recall tasks for Chinese older adults, but has little influence on the syntactic complexity of their online speech production. Higher educational attainment appears to yield considerable benefits to the resistance of age-related declarative memory decline for Chinese older adults.

Acknowledgments

We thank Hui Nga-Yan and Yuan Mingyu for their help in data collection. This work was supported by HKRGC-GRF 15601718 awarded to W.S.Wang, and a Joint Supervision Scheme and a HK postgraduate studentship to C.XIE, and to a HK postgraduate studentship Y.FENG.

References

Baudic, S., Barba, G. D., Thibaudet, M. C., Smagghe, A., Remy, P., & Traykov, L. (2006). Executive function deficits in early Alzheimer's disease and their relations with episodic memory. *Archives of Clinical Neuropsychology, 21*(1), 15-21.

Buckner, R. L. (2004). Memory and Executive Function in Aging and AD: Multiple Factors that Cause Decline and Reserve Factors that Compensate. *Neuron, 44*(1), 195-208.

Burke, D. M., & Shafto, M. A. (2011). Language and aging. In *The handbook of aging and cognition* (pp. 381-451): Psychology Press.

Chafe, W. L. (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. New Jersey: Alex: Norwood.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chertkow, H., & Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type: what do various measures measure? *Brain, 113*(2), 397-417.

De Anna, F., Attali, E., Freynet, L., Foubert, L., Laurent, A., Dubois, B., & Dalla Barba, G. (2008). Intrusions in story recall: When over-learned information interferes with episodic memory recall. Evidence from Alzheimer's disease. *Cortex, 44*(3), 305-311.

De Azeredo Passos, V. M., Giatti, L., Bensenor, I., Tiemeier, H., Ikram, M. A., de Figueiredo, R. C., . . . Barreto, S. M. (2015). Education plays a greater role than age in cognitive test performance among participants of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). *BMC Neurology, 15*(1), 191.

Delpak, A., & Talebi, M. (2020). On the Impact of Age, Gender and Educational Level on Cognitive Function in Alzheimer's Disease: A Quantitative Approach. *Archives of Gerontology and Geriatrics*, 104090.

Evans, T. M., & Ullman, M. T. (2016). An Extension of the Procedural Deficit Hypothesis from Developmental Language Disorders to Mathematical Disability. *Frontiers in Psychology, 7*, 1318.

Garrard, P., Perry, R., & Hodges, J. R. (1997). Disorders of semantic memory. *Journal of Neurology, Neurosurgery and Psychiatry, 62*(5), 431.

Gladsjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., & Heaton, R. K. (1999). Norms for Letter and Category Fluency: Demographic Corrections for Age, Education, and Ethnicity. *Assessment, 6*(2), 147-178.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society, 16*(5), 748-753.

Gunning-Dixon, F. M., Head, D., McQuain, J., Acker, J. D., & Raz, N. (1998). Differential aging of the human striatum: a prospective MR imaging study. *American Journal of Neuroradiology, 19*(8), 1501-1507.

Hedden, T., & Gabrieli, J. D. E. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews Neuroscience, 5*(2), 87-96.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature, 466*(7302), 29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and brain Sciences, 33*(2-3), 61-83.

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia, 42*(9), 1212-1222.

Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2003). Latent Change Models of Adult Cognition: Are Changes in Processing Speed and Working Memory Associated With Changes in Episodic Memory? *Psychology and aging, 18*(4), 755-769.

Jack, C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., . . . Kokmen, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology, 51*(4), 993-999.

Jiang, L., Cao, X., Jiang, J., Li, T., Wang, J., Yang, Z., & Li, C. (2019). Atrophy of hippocampal subfield CA2/3 in healthy elderly men is related to educational attainment. *Neurobiology of Aging, 80*, 21-28.

Kavé, G., & Sapir-Yogev, S. (2020). Associations between memory and verbal fluency tasks. *Journal of Communication Disorders*, 105968.

Kempler, D., Curtiss, S., & Jackson, C. (1987). Syntactic Preservation in Alzheimer's Disease. *Journal of Speech, Language, and Hearing Research, 30*(3), 343-350.

Lai, Y.-H., & Lin, Y.-T. (2013). Factors in action-object semantic disorder for Chinese-speaking persons with or without Alzheimer's disease. *Journal of neurolinguistics, 26*(2), 298-311.

Lee, J. C., Nopoulos, P. C., & Tomblin, J. B. (2020). Procedural and declarative memory brain systems in

developmental language disorder (DLD). *Brain and Language, 205*, 104789.

Li, G., Cheung, R. T. F., Gao, J. H., Lee, T. M. C., Tan, L. H., Fox, P. T., . . . Yang, E. S. (2006). Cognitive processing in Chinese literate and illiterate subjects: An fMRI study. *Human brain mapping, 27*(2), 144-152.

Lum, J. A. G., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex, 48*(9), 1138-1154.

Mok, E. H. L., Lam, L. C. W., & Chiu, H. F. K. (2004). Category Verbal Fluency Test Performance in Chinese Elderly with Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders, 18*(2), 120-124.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., . . . Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society, 53*(4), 695-699.

Piai, V., Anderson, K. L., Lin, J. J., Dewar, C., Parvizi, J., Dronkers, N. F., & Knight, R. T. (2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proceedings of the National Academy of Sciences, 113*(40), 11366.

Pu, Y., Cheyne, D., Sun, Y., & Johnson, B. W. (2020). Theta oscillations support the interface between language and memory. *NeuroImage, 215*, 116782.

Rajah, M. N., & McIntosh, A. R. (2005). Overlap in the Functional Neural Systems Involved in Semantic and Episodic Memory Retrieval. *Journal of Cognitive Neuroscience, 17*(3), 470-482.

Raz, N., Gunning-Dixon, F., Head, D., Rodrigue, K. M., Williamson, A., & Acker, J. D. (2004). Aging, sexual dimorphism, and hemispheric asymmetry of the cerebral cortex: replicability of regional differences in volume. *Neurobiology of Aging, 25*(3), 377-396.

Reifegerste, J., Veríssimo, J., Rugg, M. D., Pullman, M. Y., Babcock, L., Glei, D. A., . . . Ullman, M. T. (2020). Early-life education may help bolster declarative memory in old age, especially for women. *Aging, Neuropsychology, and Cognition*, 1-35.

Rieckmann, A., & Bäckman, L. (2009). Implicit Learning in Aging: Extant Patterns and New Directions. *Neuropsychology Review, 19*(4), 490-503.

Shan 单泽周. (1997). 郴州汉语方言概述. *湘南学院学报*(3), 39-46.

Squire, L. R., & Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus, 8*(3), 205-211.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-404). New York: Academic Press.

Tulving, E., & Schacter, D. L. (1990). Priming and Human Memory Systems. *Science, 247*(4940), 301-307.

Ullman, M. T. (2001a). The Declarative/Procedural Model of Lexicon and Grammar. *Journal of Psycholinguistic Research, 30*(1), 37-69.

Ullman, M. T. (2001b). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience, 2*(10), 717-726.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition, 92*(1), 231-270.

Ullman, M. T. (2013). The role of declarative and procedural memory in disorders of language. *Linguistic Variation, 13*(2), 133-154.

Ullman, M. T. (2014). Language and the brain. In J. Connor-Linton & R. W. Fasold (Eds.), *An Introduction to Language and Linguistics (2nd ed.)* (pp. 249-286): Cambridge University Press.

Ullman, M. T. (2016). Chapter 76 - The Declarative/Procedural Model: A Neurobiological Model of Language Learning, Knowledge, and Use. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 953-968). San Diego: Academic Press.

Ullman, M. T., Earle, F. S., Walenski, M., & Janacsek, K. (2020). The Neurocognition of Developmental Disorders of Language. *Annual Review of Psychology, 71*(1), 389-417.

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex, 41*(3), 399-433.

Ullman, M. T., & Pullman, M. Y. (2015). A compensatory role for declarative memory in neurodevelopmental disorders. *Neuroscience & Biobehavioral Reviews, 51*, 205-222.

Walenski, M., Mostofsky, S. H., & Ullman, M. T. (2014). Inflectional morphology in high-functioning autism: Evidence for speeded grammatical processing. *Research in Autism Spectrum Disorders, 8*(11), 1607-1621.

Wang 王健, 邹义壮, 崔界峰, 范宏振, 陈娆, 陈楠, . . . 晏丽娟. (2015). 韦克斯勒记忆量表第四版中文版(成人版)的修订. *中国心理卫生杂志, 29*(1), 53-59.

Yang, C.-C., Hua, M.-S., Chiu, M.-J., Chen, S.-T., Yip, P.-K., Chen, T.-F., . . . Tu, P.-C. (2006). Semantic Memory Deficits in Low-educated Patients with Alzheimer's Disease. *Journal of the Formosan Medical Association, 105*(11), 926-935.

Zahodne, L. B., Mayeda, E. R., Hohman, T. J., Fletcher, E., Racine, A. M., Gavett, B., . . . Mungas, D. (2019). The Role of Education in a Vascular Pathway to Episodic Memory: Brain Maintenance or Cognitive Reserve? *Neurobiology of Aging*.

Appendices

Appendix A. Five types of information unit

Immediate recall	**Delay recall**	**Type**	**Scores**
一个农村的早上，鸡叫了。[8] One morning in the countryside, the rooster crowed.	一个农村的清晨，鸡叫了，天亮了。 One morning in the countryside, the rooster crowed and it dawned.	Valid information unit	Immediate recall: 1 Delay recall: 1
Two baskets of pears have been collected.	Two baskets of pears have been collected.	Missing information unit	Immediate recall: 0 Delay recall: 0
打了一个口哨。他回过头来，他又……，这个男孩子呢，就把草帽呢，送给他，给他送过去。 Whistled. He(bicycler) looked back, and he…, this boy, uh, send the straw hat to him(bicycler), send him(bicycler) the straw hat.	然后呢就冲他吹了一下口哨。把这个草帽呢给他送过去。 Then he whistled at him(bicycler). Send him(bicycler) this straw hat.	Mismatch information unit	Immediate recall: 1 Delay recall: 0
然后，装好。他，然后就走了，也没说谢谢。 Then, pack it. He then left without saying thank you.	帮他搬在了车子上，自行车上。这个男孩子呢，骑上车呢就往前走了，推着车吧，不是骑，推着车。 Help him move it on the car, on the bike. The boy, uh, got on the bike and walked forward, pushing the bike, not riding, pushing the bike.	Mismatch information unit	Immediate recall: 0 Delay recall: 1
然后，装好。他，然后就走了，也没说谢谢。 Then, pack it. He then left without saying thank you.		Unrelated information unit	Immediate recall: 0 Delay recall: 0
他们一人拿了一个，然后就吃着走过来了。他就骑车走了。 They each took one and came over while eating. He rode away on his bike.		Wrong information unit	Immediate recall: 0 Delay recall: 0

[8] This is the original utterance that the participant produced, and it is followed by its translation.

A corpus-based analysis of Chinese relative clauses produced by Japanese and Thai learners

Yike Yang
Department of Chinese and Bilingual Studies,
The Hong Kong Polytechnic University
Hong Kong SAR
yi-ke.yang@connect.polyu.hk

Abstract

Concerning the acquisition of relative clauses (RCs), studies on head-initial languages consistently reported a preference for subject-gapped RCs, but the issue of subject-object asymmetry is still a controversial one in research on the acquisition of RCs in head-final languages. Using written corpus data, this study investigated the second language production of RCs in Mandarin Chinese (Chinese) by Japanese-speaking and Thai-speaking learners with various proficiency levels. We first extracted the RCs produced by Japanese and Thai learners from the HKS Dynamic Composition Corpus, and coded head type and gap type for further analyses. The learners from the intermediate-level groups produced a significant number of error-free RCs, which suggests that the intermediate learners have already mastered Chinese RCs. Both Japanese and Thai learners exhibited a strong preference for the subject RCs, which is consistent with predictions that follow from the Noun Phrase Accessibility Hierarchy (NPAH) and the results of studies on head-initial languages. Our data also provided partial support for the Subject-Object Hierarchy (SOH). However, the size of the corpus was insufficient to exhaustively investigate the tested theories. More data are needed to examine the applicability of the NPAH and SOH hypotheses in L2 Chinese and in general.

Keywords: corpus; second language acquisition; Chinese relative clauses; subject-object asymmetry

1 Introduction

The acquisition of relative clauses (RCs) in both first language (L1) and second language (L2) has been widely studied for decades. However, there remain controversial issues in previous research, such as that of subject-object asymmetry, especially in the context of Asian languages. Subject-object asymmetry has been the focus of inquiry in RC acquisition since the Noun Phrase Accessibility Hierarchy was proposed in the 1970s (Keenan & Comrie, 1977), and studies on post-nominal Indo-European languages generally accept that there is a subject preference (Izumi, 2003; Keenan & Hawkins, 1987), but the data from pre-nominal languages yield mixed results (Chan et al., 2011; Zhang & Yang, 2010). Using written corpus data, this study examined the production of RCs in Mandarin Chinese (henceforth, Chinese) by Japanese and Thai learners to investigate the subject-object asymmetry. Our data will also shed light on whether and how typological variations affect the acquisition process.

This paper begins with a brief introduction of RCs in Chinese, Japanese and Thai languages, and then reviews issues in the acquisition of RCs, particularly in the acquisition of Chinese RCs. The research questions and methods of the study will be described in Section 2. The results of the study will be presented in Section 3. General discussion and concluding remarks will appear in the final section.

1.1 Relative constructions

A relative construction comprises a nominal (i.e. head) and a subordinate clause (i.e. relative clause) that attributively modifies the nominal (Lehmann, 1986). An 'under-represented' element of the RC is co-indexed with the head, and this element picks up its interpretation from the head (O'Grady, 2011), as shown in (1), where the green apple has the same referent as the apple eaten by the addressee:

(1) The apple$_i$ [that you ate t_i] was green.
(O'Grady, 2011: 19)

Conventionally, RCs have been divided into two types: 1) the pre-nominal type, where the head appears to the right of the RC (RC-head), and 2) the post-nominal type, where the head occurs to the left of the RC (head-RC) (Keenan & Comrie, 1977). Dryer (2013) conducted a typological investigation on the relationship between the basic word order and the order of the RC and head. The results are presented in Table 1. He found that almost all the world's VO languages qualify as the second type, the post-nominal type; only five out of 421 VO languages have pre-nominal RCs, and Mandarin Chinese is one of those rare cases. Among the OV languages, there are more languages that have pre-nominal RCs than those that have post-nominal RCs (132 vs. 113). Given the special case of Chinese RCs, it is theoretically interesting to examine the acquisition of Chinese RCs by speakers of more 'typical' languages to determine whether and how typological variations affect the acquisition process.

Basic word order	Order of RC and head	Number	Example
Verb-object	RC-head	5	Mandarin
Verb-object	Head-RC	416	Thai
Object-verb	RC-head	132	Japanese
Object-verb	Head-RC	113	Persian
Languages that do not fall into the four types		213	Kutenai

Table 1: Typology of word order and order of RC and head (adapted from Dryer, 2013)

Chinese uses a relative marker *de* to link the matrix clause and the embedded RC (Li & Thompson, 1981). The RC always precedes the head, and *de* precedes the head to mark the RC, as illustrated in (2). (2a) is a subject-gapped (SU) RC, where the subject of the RC is co-indexed with the relativised head, and (2b) is an example of a direct object-gapped (DO) RC. With the basic word order of SOV, Japanese shares the pre-nominal property with Chinese RCs, as in (3). However, in Japanese, the head is directly modified by the RC, and there lacks any overt relative marker or relative pronoun (Yabuki-Soh, 2007), so the addressee will only notice the RC when the head is heard or read. The structure of Thai RCs resembles that of English RCs. Both Thai and English are VO languages with post-nominal RCs, and their RCs are introduced by a relative marker that follows the head, as illustrations in (1) for English and in (4) for Thai (Sornhiran, 1978). There are three relative markers in Thai (i.e. *thîi*, *sŷŋ* and *an*), among which *thîi* is the most productive one because it can be used under all circumstances (Phoocharoensil, 2014). It is clear that the SU RCs have a longer gap-filler dependency and that the DO RCs share their ordering with a canonical SVO clause in Chinese. Like those in Chinese, the SU RCs in Japanese also have a longer gap-filler dependency. In Thai, the SU RCs show a shorter gap-filler dependency. The contrast between Japanese and Thai RCs may lead to some differences in Japanese and Thai learners' acquisition of Chinese RCs and help us better understand cross-linguistic influence (Yang, 2020).

(2) a. *[t_i mai shu de] nanhai$_i$*
buy book REL boy
'the boy who bought a book'
b. *[ta mai t_i de] shu$_i$*
he buy REL book
'the book that he bought'

(3) a. *[t_i watashi ni hon o kure-ta] hito$_i$*
me DAT book ACC give-PST person
'the person who gave me a book'
b. *[watashi ga kinoo t_i at-ta] hito$_i$*
I NOM yesterday meet-PST person
'the person whom I met yesterday'
(Yabuki-Soh, 2007: 228)

(4) a. *phét$_i$ [thîi t_i mii khâa mahăasăan]*
diamond REL have value tremendous
'the diamond that has tremendous value'
b. *dèki [thîi chăn líaŋ ti maa]*
child REL I bring_up come
'the child that I brought up'
(Sornhiran, 1978: 177)

1.2 Subject-object asymmetry

Several hypotheses have been adopted to account for the subject-object asymmetry in the acquisition and processing of RCs; among these, the Noun Phrase Accessibility Hierarchy (NPAH) and the Subject-Object Hierarchy are of particular relevance to the current study. The NPAH, a typological generalisation proposed by Keenan and Comrie (1977), has been employed to predict the difficulty order in the acquisition of various types of RCs over recent decades (Shirai & Ozeki, 2007; Song, 2002). According to the NPAH, there exists a universal hierarchy of psychological ease of relativisation of the RC head, as suggested in (5). This means that, the subject position in a sentence is always the most accessible to undergo the process of relativisation,

Type of RC	Illustration	No. of discontinuity
SS	*[$_{TP}$ t_i xuyao gongzuo] de nanren$_i$ bangzhu-le ta.* need job REL man help-PST her 'The man who needed a job helped her.'	1
SO	*[$_{TP}$ mama [$_{VP}$ yang t_i]] de tuzi$_i$ chi-le huluobo.* mom feed REL rabbit eat-PST carrot 'The rabbit that mom feeds ate the carrot.'	2
OS	*wo kanjian [$_{DP}$ [$_{TP}$ t_i mai shu] de nanhai$_i$].* I see buy book REL boy 'I saw the boy who bought a book.'	2
OO	*ta mai-le [$_{DP}$ [$_{TP}$ nver [$_{VP}$ xiangyao t_i]] de hua$_i$].* he buy-PST daughter want REL flower 'He bought the flower that his daughter wanted to have.'	3

Table 2: Illustration of discontinuity in Chinese RC according to SOH

whereas the object of comparative position is the least accessible to be relativised. SU and DO RCs have been the focus of subsequent studies testing the NPAH, and, consistent with the prediction of the NPAH, studies on post-nominal Indo-European languages have discovered that SU RCs are easier to acquire and process than DO RCs in both L1 (Diessel & Tomasello, 2001; Keenan & Hawkins, 1987) and L2 (Gass, 1979; Izumi, 2003) contexts. Inquiries into pre-nominal languages, however, have not yet reached an agreement, because some results favour the SU type (Lau, 2016; Lee, 1992), whereas others support the DO type (Chen & Shirai, 2015; Hsiao & Gibson, 2003).

(5) A universal hierarchy of ease of relativisation: Subject > Direct Object > Indirect Object > Oblique > Genitive > Object of comparative

Furthermore, Hamilton (1994) postulated the Subject-Object Hierarchy (SOH) to rank four types of RCs according to the position of the head in the matrix clause and the role of the relativised head in the subordinate RC: OS > OO/SS > SO, where the first code refers to the position of the head in the matrix clause and the second indicates the role of the relativised head in the RC. The SOH followed the idea of processing discontinuity (O'Grady, 1987) and proposed two levels of processing discontinuity: 1) if an RC is embedded in the middle of the main clause, it will create a processing discontinuity in the main clause, and 2) an SU RC establishes one discontinuity within the RC, whereas a DO RC sets up two discontinuities (as illustrated in Table 2 with examples in Chinese). The Chinese RCs are typologically different from their English counterparts in the order of the head and the RC, so we adapted the SOH to accommodate Chinese RCs in Table 3, which also lists the rankings and the number of discontinuity for Japanese and Thai RCs. The information from the updated SOH will help us determine whether there is any transfer from the learners' L1s to their L2 Chinese.

Languages	Updated SOH
Chinese	SS (1) > SO/OS (2) > OO (3)
Japanese	SS (1) > SO/OS (2) > OO (3)
Thai	OS (1) > OO/SS (2) > SO (3)

Table 3: An updated SOH

Note: The parentheses after each RC type indicate the number of discontinuity.

1.3 Acquisition of Chinese relative clauses

Although a consensus on the subject preference in Indo-European languages has been reached, previous studies on the acquisition and processing of Chinese RCs exhibited complex results on subject-object asymmetry. For example, a preference for subject RCs in Chinese has been supported from L1 child comprehension data (Hu et al., 2016), L1 adult comprehension data (Lin & Bever, 2006), L2 comprehension data (Li et al., 2016) and L2 production data (Xu, 2014). In contrast, a preference for object

RCs in Chinese has also been suggested in L1 child production data (Chen & Shirai, 2015), L1 adult comprehension data (Chen et al., 2008) and L2 comprehension data (Packard, 2008). Moreover, Lam (2017) showed a slight subject preference in the production but an object preference in the comprehension of Chinese RCs by typical Cantonese-speaking children. However, consistency has been found in corpus-based studies on the spoken and written Chinese of native speakers, which report a preference for the subject RCs over the object RCs (e.g. 73.8% vs 26.2% in Pu (2007) and 60.8% vs 39.2% in Wu et al. (2011)).

Conflicting results were also found in Chinese for the hierarchy generated from the SOH. Only one comprehension study by Cheng (1995) exactly followed the hierarchy in Table 3. A corpus study of native written data (Wu et al., 2011) and an act out task performed by Chinese children (Lee, 1992) roughly supported the postulate of SOH, as shown by the following ranking: SS > OS > SO > OO. Pu (2007) analysed oral and written data gathered from native speakers, and both types of data followed the same pattern: SS > OS > OO > SO. Chang (1984) tested the comprehension of RCs by Chinese children and found the following hierarchy: SS/SO > OO > OS.

As reviewed above, although data from native Chinese speakers corroborated the generally accepted subject preference, acquisition studies on Chinese RCs presented more inconsistency. Besides, the results concerning the SOH hierarchy did not converge with each other, which needs to be further examined in this study.

2 The current study

2.1 Research questions

To fill gaps in the field, this study addresses the following research questions:

- Is there subject-object asymmetry in the production of RCs by Japanese and Thai learners with different levels of proficiency?

- Is the updated version of the SOH consistent with the written data of Japanese and Thai learners?

2.2 Research methods

A corpus-based approach was adopted for the current study. Correspondence between ease of processing and frequency of occurrence has been suggested in literature (Hawkins, 2004; Wu et al., 2011). This is why distributional data from a learner corpus can inform us of the possible ease of processing for L2 learners and help us answer our research questions.

The data were extracted from the Version 1.1 of the HSK Dynamic Composition Corpus[1] developed by the Beijing Language and Culture University. The Version 1.1 of the corpus contains more than 11,500 essays (more than four million Chinese characters) written by L2 learners of Chinese who took the standardised Chinese proficiency test for non-native speakers (Hanyu Shuiping Kaoshi, HSK). The L2 learners varied in their language background, so it is possible to make cross-linguistic comparisons. We chose Japanese and Thai learners because of the properties of Japanese and Thai RCs and also because of the number of Japanese and Thai learners in the corpus. In addition, there were a variety of topics, based on which the test takers were required to write essays during the test.

In this study, data from learners at the upper intermediate (henceforth, intermediate) level and advanced level were collected. Because there were much more intermediate learners than advanced learners, we first exhausted all the essays of advanced Japanese and Thai learners and manually identified typical SU and DO RCs[2]. Then, we reviewed the essays of intermediate learners and randomly selected the essays to match the number of typical RCs produced by intermediate and advanced learners who shared the same language background. We also considered the topic of the selected essays when choosing the essays. In total, there were 80 essays collected and analysed in this study, with 47

[1] The Version 1.1 of the HSK Dynamic Composition Corpus is not available. The current version of the corpus is Version 2.0, which can be accessed at http://hsk.blcu.edu.cn/.

[2] By 'typical SU and DO RCs', we are referring to the real Chinese RCs where the head noun is co-indexed with the subject or object of the RC. We did not include adjunct or possessive RCs (Lin, 2018).

An adjunct RC:
you kunnan de shihou, women yinggai huxiang bangzhu.
have trouble REL time we should each_other help
'We should help each other when we are in trouble.'

A possessive RC:
xuesheng najiang de laoshi jintian mei lai.
student win_prize REL teacher today NEG come
'The teacher whose student won the prize did not come today.'

essays produced by Japanese learners and 33 essays by Thai learners.

The role of the head in the main clause, the role of the gap in the RC, the type of the verb in the RC and the animacy of the head noun were manually coded. The analyses of the first two properties are reported in this paper. Since our data concern the frequency of occurrence of RCs, we adopted the chi-square test and the binomial test in the statistical analysis. When the sample size was insufficient for a chi-square test (as reported in Section 3.3), we used the Fisher's exact test instead, which calculates the deviation from a null hypothesis directly and is more appropriate for small-scale data.

3 Results

3.1 An overview of the corpus data

An overview of the corpus data can be found in Table 4, which shows the number and percentage of typical RCs produced by each group. There was a tendency that Japanese and Thai learners produced a larger proportion of RCs as their Chinese proficiency increased. When analysing the data, we not only examined the intermediate and advanced groups separately but also considered all the learners with the same language background as one group, so there were six groups in the table. Each group produced a considerable number of RCs, and no errors in RC usage were spotted throughout our careful investigation, which suggested that the intermediate learners had already mastered Chinese RCs fairly well. Sentences in (6) and (7) were produced by Japanese and Thai learners, respectively. Sentences (6a) and (7a) are examples of SU RCs, and (6b) and (7b) represent DO RCs.

Group	No. of essays	No. of sentences	No. of typical RCs	
JP_IN	28	430	48	(11.16%)
JP_AD	19	270	47	(17.41%)
JP_Sum	47	700	95	(13.57%)
TH_IN	21	266	26	(9.77%)
TH_AD	11	174	25	(14.37%)
TH_Sum	33	440	51	(11.59%)

Table 4: An overview of the corpus data

Note: JP_IN = intermediate Japanese learner group; JP_AD = advanced Japanese learner group; JP_Sum = all Japanese learners as a group; TH_IN = intermediate Thai learner group; TH_AD = advanced Japanese learner group; TH_Sum = all Thai learners as a group.

(6) a. *wo jiushi yi ge [t_i bu xiyan de] ren_i*
I be one CL NEG smoke REL person
'I am a person who does not smoke.'
b. *women yao zuo yixie [women yingdang zuo t_i de] $shiqing_i$*
we need do some we should do REL thing
'We need to do the things that we should do.'

(7) a. *[t_i cizhi huijia de] $funv_i$ ye yinggai zhuyi yixie wenti*
resign go_home REL woman also should care some issue
'The women who resigned and went back home should also pay attention to some issues.'
b. *wo kaishi ganxie [muqin dangchu suo zuo t_i de] $jueding_i$*
I start thank mother then ACC make REL decision
'I started to feel grateful for the decision that my mother made at that time.'

Chi-square tests of independence were employed to examine the relationship between occurrence of RCs or non-RCs and language background as well as between occurrence of RCs or non-RCs and proficiency level. The proportion in the production of RCs or non-RCs did not differ by language background ($\chi^2(1) = .780$, $p = .377$), indicating that the two language groups produced similar proportions of RCs. There was significant association between the occurrence of RCs or non-RCs and language proficiency in Japanese learners ($\chi^2(1) = 4.994$, $p = .025$) but not in Thai learners ($\chi^2(1) = 1.741$, $p = .187$), although an increase of the proportion of RC production is seen in both Japanese and Thai learners.

3.2 Subject-object asymmetry

To answer our first research question, we compared the SU RCs and DO RCs produced by each group. As shown in Figure 1, a clear subject preference was found in both Japanese and Thai learners at both proficiency levels, and the distribution pattern was very similar across the four groups. The intermediate Japanese learners produced 33 SU RCs and 15 DO RCs (68.75% vs 31.25%), and the advanced Japanese learners produced 35 SU RCs and 12 DO RCs (74.47% vs 25.53%). Similarly, the intermediate Thai learners produced 20 SU RCs and 6 DO RCs (76.92% vs 23.08%), and the advanced Thai learners produced 18 SU RCs and 7 DO RCs (72.00% vs 28.00%).

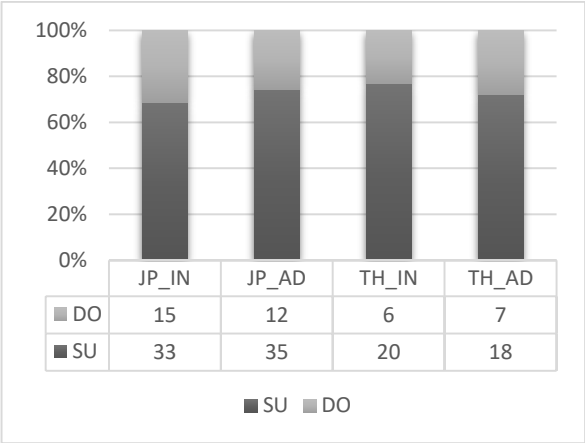


Figure 1: Production of two types of RCs by Japanese and Thai learners

Binomial tests with the type of RC as the test variable showed that there was a significant difference between the production of the two types of RCs within each group: p = .013 for the Japanese intermediate group, p = .010 for the Japanese advanced group, p < .001 for all Japanese learners, p = .009 for the Thai intermediate group, p = .043 for the Thai advanced group, and p = .001 for all Thai learners, revealing an obvious subject RC preference in all the groups. Chi-square tests of independence were then employed to test whether there was any association between RC type and language background and between RC type and proficiency level. The occurrence of SU or DO RC was unaffected by language background ($\chi^2(1)$ = .034, p = .854) or proficiency level (Japanese learners: $\chi^2(1)$ =.152, p = .696; Thai speakers: $\chi^2(1)$ = .007, p = .935), which further confirmed our findings from the binomial tests.

3.3 The Subject-Object Hierarchy

To test the ranking of the four types of RCs compared in the SOH, we further filtered out some RCs from our data and kept only the RCs in which the heads are in the subject or object positions[3]. As Figure 2 and Table 5 suggest, the six groups did not conform to the same ranking of hierarchy, and none of them followed the hierarchies as predicted by the SOH in Table 3. However, a bias toward the OS and SS types and a dispreference for the OO and SO types can be observed among the six groups, which required further interpretations.

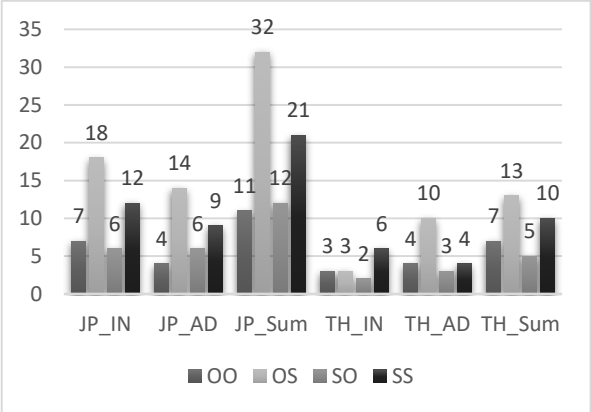


Figure 2: Production of four types of RCs by Japanese and Thai learners

Group	Hierarchy
JP_IN	OS > SS > OO > SO
JP_AD	OS > SS > SO > OO
JP_Sum	OS > SS > SO > OO
TH_IN	SS > OO/OS > SO
TH_AD	OS > SS/OO > SO
TH_Sum	OS > SS > OO > SO

Table 5: Hierarchy of the four types of RCs

Because the sample size was too small for a chi-square test, we used the Fisher's exact test to examine the relationship between the language background and the RC type as well as the relationship between the proficiency level and the RC type. No significant association between language background and type of RC was observed (p = .893). Nor did we find any statistically significant association between proficiency level and the RC type (p = .937 for the Japanese groups; p = .365 for the Thai groups). The results suggested that, in general, the hierarchies of the language and proficiency groups in our data did not distinguish from each other.

[3] Headless RCs and RCs in which the RC head is not in the subject or direct object position of the main clause were removed from our analysis.

A headless RC:
xie-wan zuoye de dou hen nuli.
write-PST assignment REL all very diligent
'Those having done the assignment are diligent.'

A case of a RC head modifying the subject noun of the main clause:
jiajing bu fuyu de ren de yiliaofei dui jiaren fudan tai da
situation not rich REL people POSS medical_cost for family burden too big
'The medical cost of the poor is a heavy burden for the family.'

4 Discussion and conclusions

This paper provided an analysis of Chinese RCs produced by Japanese and Thai learners using written corpus data. As suggested in Table 4, the Japanese and Thai learners at different proficiency levels produced a significant number of error-free Chinese RCs (ranging from 9.77% to 17.41%), which is even higher than the proportion of Chinese RCs produced by native speakers reported in Pu (2007)[4]. The high frequency of RCs in L2 learners' production indicates that Chinese RCs are acquirable by L2 learners at the intermediate level and that the L2 learners have the mental representations of RCs similar to those of native speakers, although RCs are complex in nature and are not very frequently used by native speakers. The productivity of RCs in our data can be explained by the modality of the data. RCs are embedded in a main clause, and a Chinese RC always interrupts the processing of the main clause because the RC precedes the head noun in the main clause. For example, in Sentence (9b), the readers will take 'mother' as the object of the sentence when they encounter the first four words '*wo kaishi ganxie muqin* (I start thank mother)', but as they read on, they will recognise that they should re-analyse this sentence with 'mother' as the subject of the RC (and retain the incomplete information of the main clause at the same time). The processing of RCs requires a heavy cognitive load and is thus not preferred in oral communication. Also, our data were from L2 learners of Chinese attending the HSK test, during which the learners must have been striving for higher scores by writing sentences with diverse and complex structures. It is plausible that they tried to use RCs as much as possible so that their essays may look more professional to the examiners.

This study explored the issue of subject-object asymmetry in the L2 Chinese RCs of Japanese and Thai learners. Despite the fact that Japanese and Thai are typologically different in terms of basic word order (SOV vs SVO) and the order of RC and head (RC-head vs head-RC), the distribution of the SU and DO RCs in the Japanese and Thai learners' production aligns with that of native Chinese speakers (Pu, 2007). Specifically, a clear preference for the subject RCs over the object RCs has been identified across all the groups in the production data. We thus provided some support for the NPAH with evidence from L2 Chinese. Due to the limitation of the corpus size, however, only SU and DO RCs were extracted and compared in our analysis. A detailed analysis of all the possible RC types in Chinese is needed to yield a fuller picture of the difficulty order predicted by NPAH in L2 Chinese, which requires either a more large-scale L2 corpus or a carefully-designed elicited production test.

Since SU RCs have a longer filler-gap dependency than DO RCs in Chinese, and the DO RCs also mirror the word order of Chinese sentences, it is logical that DO RCs should be more preferred than SU RCs in Chinese. But corpus data suggest that SU RCs occurs much more frequently than DO RCs, and the pattern is consistent both in L1 (Pu, 2007) and L2 speakers of Chinese. Pu (2007) borrowed the notion of 'markedness' from Givón (1993) to provide some explanations. According to Givón (1993: 178), there are three criteria for markedness: structual complexity, discourse ditribution and cognitive complexity. The marked categories are more complex in structure, less fequent in distribution and harder to process than the unmarked categories. As a pro-drop language, Chinese allows a null subject and a null object, and a null subject is much more likely to occur than a null object (Pu, 1997; Xu, 1986), which should have resulted from the subject usually being the topic of a sentence in Chinese and is thus allowed to be an empty category (Tsao, 1990). If the null subject is unmarked, it is reasonable to assume that SU RC is also unmarked. Then the higher frequency of SU RC than DO RC can be accounted for.

As for the SOH, our results in Table 5 did not exactly follow the predicted hierarchies in Table 3. This may be attributed to the relatively small number of RCs collected from the corpus, especially after we further removed the headless RCs and the RCs whose head was not in the subject or object position of the main clause. But there is a tendency of development from the intermediate groups to the advanced groups that fits approximately with the predicted hierarchies. For example, it has been proposed that the SO type should be easier to process and thus should appear more frequently than the OO type in Japanese (in Table 3). The intermediate Japanese learners produced more OO RCs than SO RCs,

[4] Pu (2007) collected both written and oral data from native Chinese speakers. The proportion of RCs is 4.66% for the written data and 0.77% for the oral data.

but a reversed hierarchy is observed for the advanced Japanese learners. The same case is also found for the Thai learners, which indicates a developmental pattern in the learners' L2 Chinese. However, this pattern is surprising, because the data of advanced learners showed more similarities to the learners' native languages than the data of intermediate learners. Further studies are needed to provide satisfactory explanations to this phenomenon.

Also, the difference in the hierarchy of the Japanese and Thai learners might be a case of cross-linguistic influence where participants' L1s are playing a role (Kidd et al., 2015), and we can hypothesise that the L1s are affecting the learners' L2 acquisition. However, no statistical difference was found between the Japanese and Thai learners, suggesting that the Japanese group and the Thai group generally resembled each other in the production data. The question then arises of why there was little cross-linguistic influence. One possible explanation comes from the typological differences documented by Dryer (2013). Chinese RCs are typologically scarce, and they diverge from Japanese and Thai RCs. Consequently, the learners may treat RCs in Chinese as a new structure that is not equivalent to the RCs in their L1s (Comrie, 2007). If this is the case, not finding any obvious transfer from the L1s would be reasonable. And this may also explain the phenomenon we discussed in the previous paragraph.

In conclusion, both Japanese and Thai learners produced native-like Chinese RCs at the intermediate level, and our findings partially supported the hypotheses of the NPAH and the SOH. Moreover, no obvious cross-linguistic influence was found in our data. However, the size of the corpus was insufficient to exhaustively investigate the tested theories. More data are needed to the examine the applicability of the NPAH and SOH hypotheses in L2 Chinese and in general.

Acknowledgements

Part of the data from the Japanese-speaking learners has been presented at the Pacific Second Language Research Forum 2016 in Chuo University, Tokyo, Japan. The author would like to thank the audience for their suggestions. We also acknowledge the useful comments from the PACLIC reviewers that have improved the quality of this paper.

References

Chan, Angel, Stephen Matthews, and Virginia Yip, 'The Acquisition of Relative Clauses in Cantonese and Mandarin', in *The Acquisition of Relative Clauses: Processing, Typology and Function*, ed. by Evan Kidd (Amsterdam: John Benjamins Publishing Company, 2011), pp. 197–225 <https://doi.org/10.1075/tilar.8.10cha>

Chang, Hsing-Wu, 'The Comprehension of Complex Chinese Sentences by Children: Relative Clause', *Chinese Journal of Psychology*, 26 (1984), 57–66

Chen, Baoguo, Aihua Ning, Hongyan Bi, and Susan Dunlap, 'Chinese Subject-Relative Clauses Are More Difficult to Process than the Object-Relative Clauses', *Acta Psychologica*, 129 (2008), 61–65 <https://doi.org/10.1016/j.actpsy.2008.04.005>

Chen, Jidong, and Yasuhiro Shirai, 'The Acquisition of Relative Clauses in Spontaneous Child Speech in Mandarin Chinese', *Journal of Child Language*, 42 (2015), 394–422 <https://doi.org/10.1017/S0305000914000051>

Cheng, Sherry Ya-Yin, 'The Acquisition of Relative Clauses in Chinese' (National Taiwan Normal University, 1995)

Comrie, Bernard, 'The Acquisition of Relative Clauses in Relation to Language Typology', *Studies in Second Language Acquisition*, 29 (2007), 301–9 <https://doi.org/10.1017/S0272263107070155>

Diessel, Holger, and Michael Tomasello, 'The Development of Relative Clauses in Spontaneous Child Speech', *Cognitive Linguistics*, 11 (2001), 131–51 <https://doi.org/10.1515/cogl.2001.006>

Dryer, Matthew S, 'Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun', in *World Atlas of Language Structures Online*, ed. by Matthew S. Dryer and Martin Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013)

Gass, Susan M., 'Language Transfer and Universal Grammatical Relations', *Language Learning*, 29 (1979), 327–44 <https://doi.org/10.1111/j.1467-1770.1979.tb01073.x>

Givón, Talmy, *English Grammar: A Function-Based Introduction* (Amsterdam/Philadelphia: John Benjamins Publishing Company, 1993)

Hamilton, Robert L., 'Is Implicational Generalization Unidirectional and Maximal? Evidence from Relativization Instruction in a Second Language', *Language Learning*, 44 (1994), 123–57 <https://doi.org/10.1111/j.1467-1770.1994.tb01451.x>

Hawkins, John A., *Efficiency and Complexity in Grammars* (Oxford: Oxford University Press, 2004)

Hsiao, Franny, and Edward Gibson, 'Processing Relative Clauses in Chinese', *Cognition*, 90 (2003), 3–27

<https://doi.org/10.1016/S0010-0277(03)00124-0>
Hu, Shenai, Mirta Vernice, Maria Teresa Guasti, and Anna Gavarró, 'The Acquisition of Chinese Relative Clauses:Contrasting Two Theoretical Approaches', *Journal of Child Language*, 43 (2016), 1–21 <https://doi.org/10.1017/S0305000914000865>
Izumi, Shinichi, 'Processing Difficulty in Comprehension and Production of Relative Clauses by Learners of English as a Second Language', *Language Learning*, 53 (2003), 285–323 <https://doi.org/10.1111/1467-9922.00218>
Keenan, Edward L., and Bernard Comrie, 'Noun Phrase Accessibility and Universal Grammar', *Linguistic Inquiry*, 8 (1977), 63–99 <https://doi.org/10.2307/4177973>
Keenan, Edward L., and Sarah Hawkins, 'The Psychological Validity of the Accessibility Hierarchy', in *Universal Grammar: 15 Essays*, ed. by Edward L. Keenan (London: Croom Helm, 1987), pp. 60–85
Kidd, Evan, Angel Chan, and Joie Chiu, 'Cross-Linguistic Influence in Simultaneous Cantonese–English Bilingual Children's Comprehension of Relative Clauses', *Bilingualism: Language and Cognition*, 18 (2015), 438–52 <https://doi.org/10.1017/S1366728914000649>
Lam, Scholastica Wai Sze, 'Acquisition of Chinese Relative Clauses by Deaf Children in Hong Kong', *Language and Linguistics*, 18 (2017), 72–115 <https://doi.org/10.1075/lali.18.1.03lam>
Lau, Elaine, 'The Role of Resumptive Pronouns in Cantonese Relative Clause Acquisition', *First Language*, 36 (2016), 355–82 <https://doi.org/10.1177/0142723716648840>
Lee, Thomas Hun-tak, 'The Inadequacy of Processing Heuristics: Evidence from Relative Clause Acquisition in Mandarin Chinese', in *Research on Chinese Linguistics in Hong Kong*, ed. by Thomas Hun-tak Lee (Hong Kong: Linguistic Society of Hong Kong, 1992), pp. 47–85
Lehmann, Christian, 'On the Typology of Relative Clauses', *Linguistics*, 24 (1986), 663–80 <https://doi.org/0024-3949/86/0024-0663>
Li, Charles N., and Sandra A. Thompson, *Mandarin Chinese: A Functional Reference Grammar* (Oakland: University of California Press, 1981)
Li, Qiang, Xiaoyu Guo, Yiru Yao, and Müller Nicole, 'Relative Clause Preference in Learners of Chinese as a Second Language', *Chinese Journal of Applied Linguistics*, 39 (2016), 199–215 <https://doi.org/10.1515/cjal-2016-0013>
Lin, Chien-Jer Charles, 'Subject Prominence and Processing Dependencies in Prenominal Relative Clauses: The Comprehension of Possessive Relative Clauses and Adjunct Relative Clauses in Mandarin Chinese', *Language*, 94 (2018), 758–97 <https://doi.org/10.1353/lan.2018.0053>
Lin, Chien-Jer Charles, and Thomas G. Bever, 'Subject Preference in the Processing of Relative Clauses in Chinese', in *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. by Donald Baumer, David Montero, and Michael Scanlon (Somerville, MA: Cascadilla Proceedings Project, 2006), pp. 254–60 <http://www.iub.edu/~lacl/att/paper1456.pdf>
O'Grady, William, *Principles of Grammar and Learning* (Chicago: University of Chicago Press, 1987)
———, 'Relative Clause: Processing and Acquisition', in *The Acquisition of Relative Clause: Processing, Typology and Function*, ed. by Evan Kidd (Amsterdam: John Benjamins Publishing Company, 2011), pp. 13–38
Packard, Jerome L., 'Relative Clause Processing in L2 Speakers of Mandarin and English', *Journal of the Chinese Language Teachers Association*, 43 (2008), 107–46
Phoocharoensil, Supakorn, 'Errors on the Relative Marker WHERE: Evidence from an EFL Learner Corpus', *3L: Language, Linguistics, Literature*, 20 (2014), 1–20 <https://doi.org//10.17576/3L-2014-2001-01>
Pu, Ming-Ming, 'Zero Anaphora and Grammatical Relations in Mandarin', in *Grammatical Relations: A Functionalist Perspective*, ed. by Talmy Givón (Amsterdam/Philadelphia: John Benjamins Publishing Company, 1997), pp. 283–322
———, 'The Distribution of Relative Clauses in Chinese Discourse', *Discourse Processes*, 43 (2007), 25–53 <https://doi.org/10.1207/s15326950dp4301_2>
Shirai, Yasuhiro, and Hiromi Ozeki, 'Introduction', *Studies in Second Language Acquisition*, 29 (2007), 155–67 <https://doi.org/10+10170S027226310707009X>
Song, Jae Jung, 'Linguistic Typology and Language Acquisition: The Accessibility Hierarchy and Relative Clauses', *Language Research*, 38 (2002), 729–56
Sornhiran, Pasinee, 'A Transformational Study of Relative Clauses in Thai' (University of Texas at Austin, 1978)
Tsao, Feng-Fu, *Sentence and Clause Structure in Chinese: A Functional Perspective* (Taipei: Student Book, 1990)
Wu, Fuyun, Elsi Kaiser, and Elaine Andersen, 'Subject Preference, Head Animacy and Lexical Cues: A Corpus Study of Relative Clauses in Chinese', in *Processing and Producing Head-Final Structures*, ed. by Hiroko Yamashita, Yuki Hirose, and Jerome Packard (Dordrecht: Springer Netherlands, 2011), pp. 173–93 <https://doi.org/10.1007/978-90-481-9213-7>

Xu, Liejiong, 'Free Empty Category', *Linguistic Inquiry*, 17 (1986), 75–93
Xu, Yi, 'Evidence of the Accessibility Hierarchy in Relative Clauses in Chinese as a Second Language', *Language and Linguistics*, 15 (2014), 435–64 <https://doi.org/10.1177/1606822X14520666>
Yabuki-Soh, Noriko, 'Teaching Relative Clauses in Japanese: Exploring Alternative Types of Instruction and the Projection Effect', *Studies in Second Language Acquisition*, 29 (2007), 219–52 <https://doi.org/10.1017/S027226310707012X>
Yang, Yike, 'Acquisition of the Mandarin Ba-Construction by Cantonese Learners', *Macrolinguistics*, 8 (2020), 88–104 <https://doi.org/10.26478/ja2020.8.12.6>
Zhang, Qiang, and Yiming Yang, 'Object Preference in the Processing of Relative Clause in Chinese: ERP Evidence', *Linguistic Sciences*, 9 (2010), 337–53

Appendix A. List of abbreviations

ACC	Accusative case
DAT	Dative case
DO	Direct object-gapped
JP_AD	Advanced Japanese learner group
JP_IN	Intermediate Japanese learner group
JP_Sum	All Japanese learners as a group
L1	First language
L2	Second language
NEG	Negative
NOM	Nominative case
NPAH	Noun Phrase Accessibility Hierarchy
OO	DO RC appearing in object position
OS	SU RC appearing in object position
OV	Object-verb
POSS	Possessive marker
PST	Past
RC	Relative clause
REL	Relative marker
SO	DO RC appearing in subject position
SOH	Subject-Object Hierarchy
SOV	Subject-object-verb
SS	SU RC appearing in subject position
SU	Subject-gapped
SVO	Subject-verb-object
TH_AD	Advanced Japanese learner group
TH_IN	Intermediate Thai learner group
TH_Sum	All Thai learners as a group
VO	Verb-object

Poster Papers

Aspect-based Sentiment Analysis on Indonesia's Tourism Destinations Based on Google Maps User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples)

Dian Arianto
Faculty of Computer Science
Universitas Indonesia
Indonesia
dian.arianto@ui.ac.id

Indra Budi
Faculty of Computer Science
Universitas Indonesia
Indonesia
indra@cs.ui.ac.id

Abstract

In this paper, we conducted an aspect-based sentiment analysis using Google Maps user reviews of Indonesia's tourism destinations which are Borobudur and Prambanan Temple. The aspects we used are *Attractions*, *Amenities*, *Accessibility*, *Image*, *Price* and *Human Resources*. The dataset obtained is in code-mixed language. We applied five machine learning algorithms which are Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET). The evaluation performed by making eight scenarios which are the combination of stopwords removal (SR), stemming (SM), emoji processing (EP), our own stopwords dictionary (OSD), and Suciati and Budi stopwords dictionary (SSD). The model performance was measured by ten folds cross-validation. The results suggest that SM without SR, and with or without EP, SSD, and OSD did not result in a significant difference for the F1-scores. However, the combination of SM and EP, and the combination of SR, EP, and SSD did improve the performance of models for classifying sentiments.

1 Introduction

Indonesia has many tourism destinations that can be visited by both domestic tourists and foreign tourists. There is an increasing number of tourists visiting Indonesia each year. Domestic tourists and foreign tourists continue to arrive every year to visit tourist attractions in Indonesia. Foreign tourists experienced a significant increase to visit Indonesia from 2009 to 2018. The average growth of foreign tourists in the 2009-2013 period was 9% per year and rose to 14% per year in the 2014-2018 period (Widowati, 2019).

Indonesia in 2015 set a program of 10 Priority Tourism Destinations or called "10 New Bali" to promote Indonesian tourism and increase foreign tourist visits. One of the 10 New Bali is Borobudur Temple, Central Java. The tourism destination of temples in Indonesia besides Borobudur Temple which is quite famous for tourists is Prambanan Temple.

Every year, reviews on online platforms experience significant competition. Since 2015, Google has experienced a significant increase in the number of reviews shared by internet users compared to other platforms such as Facebook, Yelp[1], TripAdvisor or Foursquare (Murphy, 2018). The increase in the number of reviews on the Google platform is supported by Google's program called Google Local Guides[2]. This program was originally launched in 2015 as a way to deal with Yelp Elites (Yelp contributors), which allows the most active Google Maps contributors to get rewards. In 2016, this program had 5 million contributors globally and increased every year to 120 million globally in 2019 (Sterling, 2019).

In Indonesia, research on text mining in the field of tourism had been done before to obtain

[1] https://www.yelp.com/
[2] https://maps.google.com/localguides

important information from tourism destinations. Prameswari et al. (2017) conducted a sentiment analysis for hotels in Bali and Labuan Bajo by using five aspects of the hotel domain, those are: *Accessibility*, *Activities & Entertainment*, *Food & Beverage Operations*, *Human Resources*, and *Physical Environment*. Herry et al. (2019) conducted sentiment analysis to obtain important information from user reviews on TripAdvisor on 10 Indonesian tourist destinations.

While research in the field of tourism using Google Maps reviews is still very limited both in Indonesia and globally. Munawir et al. (2019) conducted a study using Google Maps reviews to get a visitor's perspective on parks in the city of Bandung in Indonesia. They used TF-IDF to examine the term of reviews that had important value for the visitor.

In this paper, we conducted an aspect-based sentiment analysis using Google Maps user reviews of Indonesia's tourism destinations which are Borobudur Temple and Prambanan Temple. The aspects we used are *Attractions*, *Amenities*, *Accessibility*, *Image*, *Price* and *Human Resources* (World Tourism Organization, 2007). The dataset obtained from Google Maps is in code-mixed language (Indonesian and English).

The rest of this paper is arranged as follows: in Section 2, we review the related works with our study. Then we describe the research methodology applied in this work in section 3. In section 4, we discuss the results and analysis of the experiment. Finally, in section 5, we conclude our results and define future work.

2 Related Work

In the field of Natural Language Processing (NLP), research on sentiment analysis has been carried out to extract information from the opinions of internet users in several domains. In the restaurant domain, Suciati and Budi (2019) conducted a sentiment analysis of internet user reviews for several restaurants in Indonesia. They used several machine learning methods such as Random Forest (RT), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET) to classify aspects of internet user reviews in code-mixed languages (Indonesian-English). In addition, they used a combination of stemming and removing stopwords to discover what was the best scenario that was able to increase the performance of the models. They also used their own stopwords instead of using common Indonesian and English stopwords to gain better performance of the models.

In the tourism domain, Prameswari et al. (2017) conducted an aspect-based sentiment analysis for online reviews of TripAdvisor users in hotels in Bali using the Recursive Neural Tensor Network (RNTN) algorithm at the sentence level. They used eight aspects in their research. With an average accuracy of 85%, the proposed algorithm managed to predict well in classifying sentiments from words or aspects. While the average F1-score was 77% with the highest F1-score of positive sentiment was 90%.

Other research related to sentiment analysis in tourism is carried out by Kuhamanee et al. (2017). They conducted research to obtain information on sentiments from foreign tourists who aim to improve and develop the tourism industry in Bangkok. The dataset used was 10,000 tweets from the Twitter platform in 2017. The methods used were Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) using Rapidminer Studio 7.4. They gained insight that most tourists visiting Bangkok were for the purpose of nightlife activities, Thai culture, and shopping with a percentage of 65.54%, 16.07%, and 16.07%.

Kurniawan et al. (2019) conducted hierarchical sentiment analysis on hotel reviews taken from the Traveloka[3] website using Naïve Bayes. The data used were 1,720 reviews consisting of 430 positive reviews, 430 negative reviews, and 860 neutral reviews. The results of their study indicated that the use of hierarchical classifications for sentiment analysis was able to increase the average performance of the classification model by 2.3%.

Souza et al. (2018) performed a sentiment analysis on hotel reviews using CNN and compared it with other methods such as Lemmatization, Polarity inversion, and Laplace smoothing that had previously been done with the same dataset. The dataset used is 69,075 reviews from TripAdvisor about hotels located in the city of Rio De Janeiro, written in Brazilian Portuguese. The results showed that using only positive and negative reviews as in previous studies containing

[3] https://www.traveloka.com/

the same number of reviews in each class, the resulting accuracy of the model was 95.74%.

Khine and Aung (2019) conducted a sentiment analysis using the SenticNet MA-LSTM deep learning approach for restaurant reviews. The dataset consists of 20,000 sentences from TripAdvisor, which were categorised as positive, negative, and neutral sentences. The results of the experiments show that SenticNet MA-LSTM achieved the best results with an accuracy of 87.2% compared to the ordinary LSTM, which is equal to 82%.

From these studies, several approaches can be used to conduct sentiment analysis using either machine learning or deep learning. Machine learning is generally used for medium-sized data, and deep learning can be used for data with fairly large size.

3 Research Methodology

In this section, the research methodology that applied in this study will be described. This work consists of five steps, as shown in Figure 1.

Firstly, we collected the data from the website a using crawling tool. The second step was applying few text preprocessing techniques, the third one was doing feature extraction, the fourth was doing an experiment with machine learning models, and the last one, we evaluated the models.

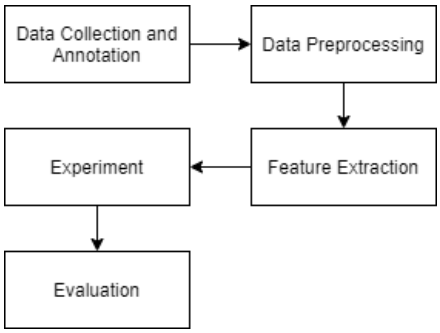


Figure 1. Research Methodology

3.1 Data Collection and Annotation

In the data collection step, we collected 5,592 reviews from Borobudur and Prambanan Temple by crawling them using BotSol's Google Maps Review Crawlers[4]. We collected 2,796 reviews from both Borobudur and Prambanan Temple. The reviews are in Indonesian, English, and mixed (Indonesian and English). The example of the data that we retrieved are as follows:

- Review in Indonesian: "*Disini adalah salah satu wisata di 7 keajaiban dunia, bahkan taman wisata borobudur ini sudah termasuk ke kawasan wisata 10 new bali, jadi semakin terkenal di kancah international, dan tiket masuknya ini 50.000 jika ingin sekaligus ke candi Prambanan bisa dengan harga tiketnya 75.000.*" (Here is one of the 7 wonders of the world tourist attractions, even Borobudur tourism park is already included in the 10 new Bali tourist area, so it is increasingly famous in the international arena, and the entry ticket is 50,000 if you want to go to Prambanan temple at the same time the ticket price is 75,000.)
- Review in English: "*The temple itself is beautiful but everything else about this place is terrible. The prices are extremely high, especially for sunset and sunrise, absolute rip off."*
- Review in mixed Indonesian-English: "*A very nice temple. Sangat edukatif, penjelasan mengenai sejarah tiap candi dan kejadian terekam dengan baik. Harga yang memadai untuk pemandangan yang indah dan menyenangkan. Highly recommended for local and international tourist!*" (A very nice temple. Very educative, an explanation of the history of each temple and events are well recorded. Adequate prices for beautiful, pleasant scenery. Highly recommended for local and international tourists!)

After retrieving the data, we then annotated them manually with the help of volunteers. Each data annotated by three annotators and we used majority voting to decide the final label of each review.

The aspects that we used in this research are *attractions*, *amenities*, *accessibility*, *image*, *price*, and *human resources* as the important aspects of tourism recommended by World Tourism Organization (2007). Then we divided the sentiment polarities into "positive", "negative", and "neutral" based on Suciati and Budi (2019). In addition, we added the "none" sentiment polarity to those reviews that did not contain six aspects of tourism. Aspects are classified as "positive" if the review mentioned positive words or phrases such as "*beautiful*", "*nice park*", "*cukup bersih*" (clean enough), "*amazingly cheap*", "*bagus*" (good), etc. We classified the negative aspects if there are negatives words or phrases such as "*unattractive*

[4] https://www.botsol.com/Products/GmapsReviewCrawler

spot", "*nothing special*", "*expensive*", "*toilet gak ada air*" (toilet no water), "*kurang petugas*" (inadequate staff), etc. For the "neutral" aspects, we classified the reviews that are not both positive and negative. For example, the reviews that contain phrase like "*standard entry ticket*", "*part of the seven wonders of the world*", "*toilets are fairly up to standard*", "*one of unesco world heritage*", etc. For the "none" aspects, we classified the reviews that do not contain the six aspects of tourism.

The following example shows how we annotated the data:

Review: "Borobodur is also known as Temple on the Hill. Breakfast and torch light included. Ticket was IDR 475,000 for the sunrise package. They'll give you a scarf as a souvenir when you return the torchlight. There were lots of people during the sunrise and even more after 6am, so it's still best to come at sunrise as there are lesser people. Nice place to see!" {*"attractions": "positive"; amenities: "positive"; "accessibility": "none"; "image": "positive"; "price": "neutral"; "human resources": "none"*}

After we annotated the data, we filtered the reviews and we only used 4,395 reviews of the data we obtained in this research.

3.2 Data Preprocessing

After we collected and annotated the data, we then performed several data preprocessing techniques in text mining to clean the data. These are the techniques we used:

1) *Emoji Processing*: at the first step, we processed the emojis that appeared in the text to string that represents its meaning which is in Indonesian. We used the top 10 positive, negative, and neutral representation of emojis based on Novak et al. (2015). We changed 10 positives emojis which are: "😂, ❤️, ♥, 😍, 😘, 😊, 👌, 💞, 👏, and 😁" to "*positif*" (positive); 10 negatives emojis which are: "😭, 😩, 😒, 😔, 😡, 😅, 🔫, 😞, 😪, and 😫" to "*negatif*" (negative); and 10 neutral emojis which are: "☯, ✨, ★, ▉, 🔥, ♫, 🔫, ©, 👀, and 🐓" to "*netral*" (neutral).

2) *Case Folding*: Next, we performed case folding to make the words in the text become lower case. For example, "Very beautyfulll" becomes "very beautyfulll".

3) *Remove Username, Numbers, and Punctuation*: the next step, we removed the usernames, numbers, and punctuation occurred in the data. For instance, "Amazing place open from 06.00 am until 17.00pm #visualine my instagram @antoniusandryano if you like please follow" converted to "Amazing place open from am until pm visualine my instagram if you like please follow".

4) *Text Normalization (part 1):* in this step, we normalised the spelling and abbreviation of the words into the formal words in Indonesian and English by using the modified dictionary from Suciati and Budi (2019). For example, "*wisata budaya yg sgt bagus*" changed to "*wisata budaya yang sangat bagus*" (the cultural tourist attraction which is very good).

5) *Stopwords Removal*: for this step, we used two dictionaries which are the stopwords dictionary used in Suciati and Budi (2019) and stopwords that we built and combined based on Tala (2003) for Indonesian and Countwordsfree Tools[5] for English. We used two dictionaries to investigate how the stopwords used in Suciati and Budi (2019) affect the performance of the models. Since they did not remove words such as "not" or "tidak" (not) in their dictionary in order to avoid missing information about the negation of positive words.

6) *Removing Duplicate Characters and Whitespace*: the next step, we removed the duplicate character occurred in text such as "beautifullll" changed to "beautiful". We also removed the whitespace in the text.

7) *Text Normalization (part 2)*: next, we performed normalization again to correct the spelling after we did duplicate character removal. We did this because there are some words result of removing duplicate characters like "peningalan" should be "peninggalan" (heritage) that can affect the stemming step after this step.

8) *Stemming*: in the last step, we performed stemming functions from libraries. We used two libraries, which are Snowball Stemmer from NLTK[6] library for English and Sastrawi library[7]

[5] https://countwordsfree.com/stopwords
[6] https://www.nltk.org/
[7] https://github.com/har07/PySastrawi

for Indonesian. For instance, the Indonesian review is "*tamannya tertata rapi*" (the park is neat) converted to "*taman tata rapi*" (neat park layout). For English review "*it's definitely an human patrimony, but it could have some more explanation (like small cards)*" converted to "*it definit an human patrimony, but it could have some more explan (like small cards)*"

3.3 Feature Extraction

In this step, we extracted the feature that would be used in the classification models. We used bigram term for the feature and its vector were extracted by vectorizing the words in the reviews. In addition, we also used the combination of stemming, stopwords removal, the use of our mixed stopwords dictionary, Suciati and Budi (2019) stopwords dictionary, and emoji processing steps to see whether they can increase the performance of the models.

3.4 Experiment

In our experiment, we performed eight scenarios and applied five machine learning algorithms that were used in Suciati and Budi (2019). Then, we measured and compared their performance using their F1-scores.

1) *Experiment Scenarios*: we examined eight scenarios for our experiments in this research. The objective is to see how the stopwords removal, stemming, the use of stopwords dictionaries, and emoji processing can affect the performances of the machine learning models when they applied to our dataset. In the first scenario, we built machine learning models by applying stopwords removal, emoji processing, and the use our stopwords dictionary, but we did not use stemming and the use of Suciati and Budi (2019) stopwords dictionary. For the second scenario, we applied stopwords removal, emoji processing, and the use of Suciati and Budi (2019) stopwords dictionary, but we did not use stemming and the use of our stopwords dictionary. The rest of scenarios can be seen in Table 1 with *SR = Stopwords Removal; SM = Stemming; EP = Emoji Processing; OSD = Our Stopwords Dictionary; SSD = Suciati and Budi Stopwords Dictionary*.

2) Dataset: we used all annotated data (4,395 reviews) for all scenarios. From Figure 2, we can see that "none" polarity had the highest number in *Amenities*, *Accessibility*, *Price*, and *Human Resources* (HR) aspects, while "positive" polarity had the highest number in *Attractions* and *Image* aspects. Positive reviews for *Attractions* aspect appeared almost in all reviews and there were only 855 reviews that were not positive. In *Image* aspect, the "positive" and "none" polarities had slightly different number of reviews. In contrast, *Amenities*, *Accessibility*, *Price*, and *HR* aspects had more than 3,500 reviews that had no polarity or "none".

Scenarios	SR	SM	EP	OSD	SSD
Scenario 1	✓	×	✓	✓	×
Scenario 2	✓	×	✓	×	✓
Scenario 3	✓	×	×	✓	×
Scenario 4	✓	×	×	×	✓
Scenario 5	×	✓	✓	✓	×
Scenario 6	×	✓	✓	×	✓
Scenario 7	×	✓	×	✓	×
Scenario 8	×	✓	×	×	✓

Table 1. Experiment Scenarios

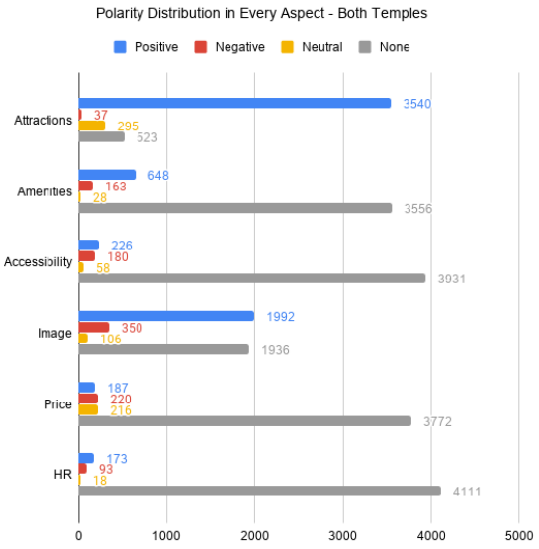


Figure 2. Polarity Distribution in Every Aspect

3) Classification Algorithms: for the experiments, we used five algorithms that were proposed by Suciati and Budi (2019) with different scenarios from the scenarios they used in their work. The machine learning algorithms we used are Decision Tree (DT), Random Forest (RF),

Logistic Regression (LR), Extra Tree (ET) or extremely Randomized Tree, and Multinomial Naïve Bayes (NB) classifiers. After we selected the machine learning algorithms and conducted the classification experiments, we then compared the obtained F1-scores of the models. Then, we can see the best machine learning algorithms for our dataset.

3.5 Evaluation

In this research, we used five classifiers which are NB, LR, RF, ET, and DT with cross-validation as the validation technique. After we did the experiments, we evaluated and compared the F1-scores of the machine learning models in all scenarios. In this experiment, we used ten folds for cross-validation technique. The performance of the models with the eight scenarios can be seen in Discussion section.

4 Discussion

In this section, we discussed the performance of the models in every aspect. For every aspect in Table 2 until Table 9 we use abbreviation in the table which are: *Att = Attractions; Amn = Amenities; Acc = Accessibility; Img = Image; Prc = Price; and HR= Human Resources*.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.655	0.744	0.796	0.479	0.773	0.818
LR	**0.733**	0.772	0.855	**0.516**	0.834	0.906
RF	0.729	0.763	0.855	0.455	0.841	0.905
ET	0.727	0.787	0.871	0.498	0.866	0.906
DT	0.717	**0.796**	**0.874**	0.500	**0.871**	**0.908**

Table 2. Result of First Scenario

Table 2 shows the F1-scores of the models for the first scenario which was applying SR, EP, and OSD, but without SM and SSD. From the table, we can see that LR had the highest F1-scores for Att and Img aspects which were 73.3% and 51.6% respectively. For the other aspects, it was led by DT by obtaining 79.6%, 87.4%, 87.1%, and 90.8% for Amn, Acc, Prc, and HR respectively.

Table 3 shows the F1-scores of the models for the second scenario which was applying SR, EP, and SSD, but without SM and OSD. From the table, we can see that LR had the highest F1-scores for Att aspects which were 73.4%. For the other aspects, it was led by DT excluding the Prc aspect that led by ET. Compared to Table 2, only NB had higher result in every aspect than other algorithms.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.672	0.748	0.801	0.490	0.782	0.826
LR	**0.734**	0.776	0.855	0.527	0.835	0.906
RF	0.733	0.747	0.860	0.477	0.839	0.906
ET	0.727	0.780	0.875	0.520	**0.867**	0.907
DT	0.720	**0.789**	**0.881**	**0.530**	0.866	**0.911**

Table 3. Result of Second Scenario

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.654	0.744	0.797	0.474	0.773	0.816
LR	**0.733**	0.772	0.854	**0.517**	0.834	0.906
RF	0.731	0.762	0.857	0.458	0.841	0.904
ET	0.724	0.785	0.871	0.502	0.865	0.906
DT	0.719	**0.794**	**0.876**	0.496	**0.872**	**0.908**

Table 4. Result of Third Scenario

Table 4 shows the F1-scores of the models for the third scenario which was applying SR and OSD, but without SM, EP, and SSD. From the table, we can see that LR had the highest F1-scores for Att and Prc aspects which were 73.3% and 51.7% respectively. For the other aspects, it was led by DT. Compared to Table 2 and 3, it can be concluded that result of all models were lower.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.670	0.746	0.803	0.486	0.782	0.827
LR	**0.734**	0.775	0.855	0.525	0.836	0.906
RF	0.728	0.755	0.857	0.481	0.845	0.906
ET	0.725	0.780	0.875	0.519	0.865	0.908
DT	0.719	**0.791**	**0.883**	**0.527**	**0.867**	**0.912**

Table 5. Result of Fourth Scenario

Table 5 depicts F1-scores of the models for the fourth scenario which was applying SR and SSD, but without SM, EP, and OSD. From the table, we can see that LR had the highest F1-score again for Att aspect which was 73.4%. For the other aspects, it surprisingly was all led by DT.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.827	0.520	0.810	0.868
LR	**0.744**	**0.810**	0.869	0.561	0.856	0.906
RF	0.732	0.778	0.863	0.525	0.840	0.906
ET	0.731	0.805	0.881	**0.571**	**0.879**	0.910
DT	0.713	0.799	**0.885**	0.539	0.871	**0.915**

Table 6. Result of Fifth Scenario

Table 6 depicts F1-scores of the models for the fifth scenario which was applying SM, EP, and OSD, but without SR and SSD. From the table, we can see that LR had the highest F1-scores again for Att aspect which were 74.4% and for the first time for Amn was led by LR which were 81%. For the Img and Prc aspects it was led by ET, and for Acc and HR it was led by DT.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.827	0.520	0.810	0.868
LR	**0.744**	**0.810**	0.869	0.561	0.856	0.906
RF	0.732	0.778	0.863	0.525	0.840	0.906
ET	0.730	0.807	0.878	**0.571**	**0.881**	0.912
DT	0.713	0.799	**0.885**	0.539	0.871	**0.915**

Table 7. Result of Sixth Scenario

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.828	0.519	0.810	0.868
LR	**0.742**	**0.810**	0.869	0.560	0.856	0.906
RF	0.734	0.773	0.860	0.524	0.846	0.905
ET	0.731	0.803	0.879	**0.572**	**0.880**	0.913
DT	0.711	0.794	**0.882**	0.542	0.876	**0.914**

Table 8. Result of Seventh Scenario

Table 7 depicts F1-scores of the models for the sixth scenario which was applying SM, EP, and SSD, but without SR and OSD. From the table, we can see that LR had the highest same F1-scores as the fifth scenario for Att and Amn. For the other aspects, it can be seen that every aspects had the same score as the fifth scenario except for the Prc aspect that had 0.02% difference.

Table 8 depicts F1-scores of the models for the seventh scenario which was applying SM and OSD, but without SR, EP, and SSD. From the table, we can see that LR had the highest F1-scores Att and Amn which were 74.2% and 81% respectively. For the other aspects, it can be seen that ET had the highest scores for Img and Prc aspects, and DT for Acc and HR aspects.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.828	0.519	0.810	0.868
LR	**0.742**	**0.810**	0.869	0.560	0.856	0.906
RF	0.734	0.773	0.860	0.524	0.846	0.905
ET	0.728	0.805	0.883	**0.571**	**0.877**	0.912
DT	0.711	0.794	**0.882**	0.542	0.876	**0.914**

Table 9. Result of Eighth Scenario

Table 9 shows F1-scores of the models for the last scenario which was applying SM and SSD, but without SR, EP, and OSD. From the table, we can see that every aspect had the highest score with the models same as seventh scenario which the result is in Table 8. However, for ET, the models had the lower scores compared to Table 8 which had 0.01% difference for Img aspect and 0.03% for Prc aspect.

In summary, by seeing Figure 3 and Figure 4, DT was the best algorithm to predict the sentiment in almost six aspects for the first, second, third, and fourth scenarios, and the rest were LR and ET that obtained the highest F1-scores in the fifth, sixth, seventh, and eighth scenarios. It can be seen that the combination of SR, SM, EP, OSD, and SSD can affect the performance of models. The Att and Amn aspects obtained the highest scores (74.4% and 81% respectively) by LR in the fifth and sixth scenario which were applying SM, EP, and OSD without SR and SSD for the fifth scenario and applying SM, EP, and SSD without SR and OSD for the sixth scenario. Besides, Acc and HR aspects achieved the highest score by DT in the fifth and sixth scenarios as well, which were 88.5% and 91.5% respectively. It seems that SM and EP affect the models for Att, Amn, Acc, and HR aspects more than SR and EP because their scores were higher every time SM and EP used in the models. OSD and SSD did not affect the performance of models because SR was not used in the scenarios. While Img aspect obtained the highest score, which was 57.2% in the seventh scenario by ET,

and Prc achieved the highest score, which was 88.1% in the sixth scenario by ET as well. It can be seen that the application of SM without EP in the seventh scenario did not result in a significant difference compared to the application of SM and EP in the fifth or sixth scenario since in the fifth and sixth scenarios Img aspect achieved 57.1% by ET.

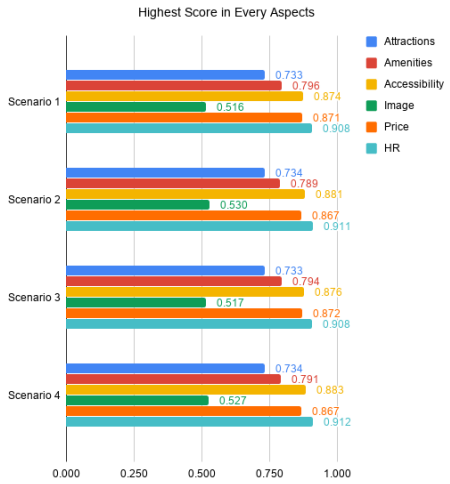


Figure 3. Highest Scores in Every Aspect

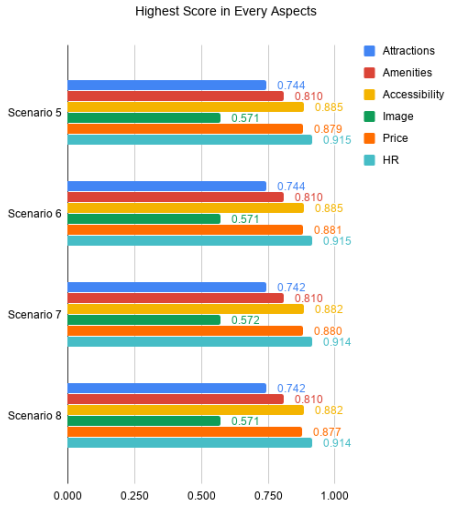


Figure 4. Highest Scores in Every Aspect

Furthermore, the F1-scores achieved by the models were various from 45.5% to 91.5%. However, the highest score obtained by Img aspect was only 57.2% while other aspects were above 70%. This was highly likely caused by the annotated data for the Img aspect had quite a significant difference annotation results between one annotator and another, leading to deletion of some data in this study and causing the models unable to predict sentiment well on the Img aspect. For example, the review was: "*Everything is good, but sometimes it gets very crowded. Avoid to go here when holidays or weekends*," and the results for the Img aspect for three annotators were "*negative, positive, and none*". So the data was removed and not included in the training dataset. This was also happened in other data that used for training dataset for Img aspect. For other aspects, the results of the data annotation were good enough so that the model managed to predict sentiment quite well by generating an F1-scores above 70%.

5 Conclusion

In this work, we have examined the performances of five machine learning algorithms and eight scenarios to classify the sentiment of Google Maps user reviews in Borobudur and Prambanan temples which is in code-mixed. The machine learning algorithms that we used are Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET). The aspects are *Attractions* (Att), *Amenities* (Amn), *Accessibility* (Acc), *Image* (Img), *Price* (Prc), and *Human Resources* (HR). The evaluation performed by making eight scenarios which are the combination of stopwords removal (SR), stemming (SM), emoji processing (EP), our own stopwords dictionary (OSD), and Suciati and Budi stopwords dictionary (SSD). The model performance was measured by ten folds cross-validation, and the results show that LR achieved the highest score for Att (74.4%) and Amn (81%), DT achieved the highest score for Acc (88.5%) and HR (91.5%). While ET achieved the highest scores for Img (57.2%) and Prc (88.1%) aspects. By seeing the results, it can be concluded that SM without SR, and with or without EP, SSD, and OSD did not result in a significant difference for the F1-scores. However, the combination of SM and EP, and the combination of SR, EP, and SSD did improve the performance of models for classifying sentiments.

In this experiment, we only applied aspect-based sentiment analysis for the reviews, in the future we will conduct topic modelling to see what topics that people frequently talked in the reviews and will use deep learning for the larger dataset.

Acknowledgments

The authors would like to thank Universitas Indonesia for funding this work through International Indexed Publication (PUTI Proceedings) research grant No. NKB-870/UN2.RST/HKP.05.00/2020.

References

Herry Irawan, Gina Akmalia, R. A. M. (2019). *Mining Tourist's Perception toward Indonesia Tourism Destination Using Sentiment Analysis and Topic Modelling*. (1).

Joana Gabriela Ribeiro de Souza, A. de P. O., & Guidson Coelho de Andrade, A. M. (2018). A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews. In *NLDB 2018*. https://doi.org/10.1007/978-3-319-91947-8

Khine, W. L. K., & Aung, N. T. T. (2019). Applying Deep Learning Approach to Targeted Aspect-based Sentiment Analysis for Restaurant Domain. *2019 International Conference on Advanced Information Technologies, ICAIT 2019*, 206–211. https://doi.org/10.1109/AITC.2019.8920880

Kuhamanee, T., Talmongkol, N., Chaisuriyakul, K., San-Um, W., Pongpisuttinun, N., & Pongyupinpanich, S. (2017). Sentiment analysis of foreign tourists to Bangkok using data mining through online social network. *Proceedings - 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017*, 1068–1073. https://doi.org/10.1109/INDIN.2017.8104921

Kurniawan, S., Kusumaningrum, R., & Timu, M. E. (2019). Hierarchical Sentence Sentiment Analysis of Hotel Reviews Using the Naïve Bayes Classifier. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 104–108. https://doi.org/10.1109/ICICOS.2018.8621748

Munawir, Koerniawan, M. D., & Dewancker, B. J. (2019). Visitor perceptions and effectiveness of place branding strategies in thematic parks in Bandung City using text mining based on google maps user reviews. *Sustainability (Switzerland)*, *11*(7). https://doi.org/10.3390/SU11072123

Murphy, R. (2018). Comparison of Local Review Sites: Which Platform is Growing the Fastest? Retrieved February 3, 2020, from https://www.brightlocal.com/research/comparison-of-local-review-sites/

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, *10*(12), 1–22. https://doi.org/10.1371/journal.pone.0144296

Prameswari, P., Surjandari, I., & Laoh, E. (2017). Opinion mining from online reviews in Bali tourist area. *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, *2018-Janua*, 226–230. https://doi.org/10.1109/ICSITech.2017.8257115

Prameswari, P., Zulkarnain, Surjandari, I., & Laoh, E. (2017). Mining online reviews in Indonesia's priority tourist destinations using sentiment analysis and text summarization approach. *Proceedings - 2017 IEEE 8th International Conference on Awareness Science and Technology, ICAST 2017*, *2018-Janua*(iCAST), 121–126. https://doi.org/10.1109/ICAwST.2017.8256429

Sterling, G. (2019). Google Maps becomes more 'social' with Local Guides follow feature. Retrieved June 27, 2020, from https://searchengineland.com/google-maps-becomes-more-social-with-local-guides-follow-feature-325322

Suciati, A., & Budi, I. (2019). Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia. *Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019*, 59–64. https://doi.org/10.1109/IALP48816.2019.9037689

Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. *M.Sc. Thesis, Appendix D*, *pp*, 39–46.

Widowati, H. (2019). 5 Tahun Terakhir, Rerata Pertumbuhan Kunjungan Wisatawan Mancanegara 14%. Retrieved February 3, 2020, from Https://Databoks.Katadata.Co.Id website: https://databoks.katadata.co.id/datapublish/2019/07/17/5-tahun-terakhir-rerata-pertumbuhan-kunjungan-wisawatan-mancanegara-14

World Tourism Organization. (2007). A Practical Guide to Tourism Destination Management. In *A Practical Guide to Tourism Destination Management*. https://doi.org/10.18111/9789284412433

Imbalanced Chinese Multi-label Text Classification Based on Alternating Attention

Hongliang Bi, Han Hu, Pengyuan Liu*
Beijing Language and Culture University, Beijing 100083, China
Language Resources Monitoring & Research Center, Beijing 100083, China
{201821198617, 201821198609}@stu.blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

In this work, we construct an imbalanced Chinese multi-label text classification dataset, IMCM. The imbalance is mainly reflected in: (1) The degree of discrimination among labels is different. (2) The distribution of labels is moderately imbalanced. Then, we adopt several methods for multi-label classification and conduct thorough evaluation of them, which show that even the most competitive models struggle on this dataset. Therefore, to tackle these imbalanced problems, we proposed an alternating attention model, AltXML. Two attention heads which alternately reading sequence enable the model capture different parts of the document rather than one point. Experimental results show that our proposed model significantly outperforms the state-of-the-art baselines in our IMCM dataset, and also achieves quite good results in several public datasets.

1 Introduction

Multi-label classification (MLC) is an important task in natural language processing (NLP) due to the increasing number of fields where it can be applied, such as text classification, tag suggestion, information retrieval, and so on. Compared to single-label classification task, multi-label classification task aims to assign a set of labels to a single instance simultaneously. However, the number of label sets grows exponentially as the number of class labels increases and the uncertainty in the number of labels per instance inevitably makes the MLC task much more difficult to solve. Therefore, the key challenge of this task lies in the overwhelming and uncontrollable size of output space.

Large amount of efforts have been done towards MLC task, including Binary Relevance (BR) (Boutell et al., 2004), Classifier Chains (CC) (Read et al., 2011), Label Powerset (LP) (Tsoumakas and Vlahavas, 2007), PD-Spare (Yen et al., 2016), SLEEC (Bhatia et al., 2015), AnnexML (Tagami, 2017), PfastreXML (Jain et al., 2016), Parabel (Prabhu et al., 2018).. In addition to the above methods, neural networks provide some new approaches: CNN (Kim, 2014), CNN-RNN (Chen et al., 2017), SGM (Yang et al., 2018), etc. These methods have made great progress in capturing label correlations to cope with the exponential-sized output space, but still face the problem of high computational complexity and poor scalability.

While utilizing correlations among labels is essential for MLC task, in real-word scenarios, there are no obvious semantic boundaries among some labels and some seemingly distinct labels may appear together, especially for text. Moreover, the distribution of labels may be imbalanced. On the one hand, the number of instance belonging to a certain label may outnumber other labels. On the other hand, there may be a relatively high number of examples associated with the most common labels or infrequent labels (Gibaja and Ventura, 2015). These may affect the performance of models utilizing correlations of labels. Therefore, it is important to explore the balance between using correlation to reduce output space and improving the ability to refine labels.

* Corresponding Author.

We inspect the commonly used multi-label text classification datasets consist of Rcv1v2 (Lewis et al., 2004), AAPD (Yang et al., 2018), etc. Some of them has been used as benchmarks, but still can not meet the actual demand. The numbers of class labels or labels per instance is small, and the semantic boundaries among the labels are obvious to some extend. Therefore, to further explore this field, we propose an imbalanced Chinese multi-label text classification dataset, IMCM [1].

Furthermore, we conduct a detailed evaluation for diverse MLC models on our dataset and two public datasets. Experimental results show that several models that perform well on other datasets struggle on our dataset. Our point of view is that, different from single label classification models which need to focus on the most important part of the document, multi-label classification models need to be aware of different parts. That means that models can't be bound by a certainly associated label.

Therefore, inspired by the idea of dilated convolution which has become popular in semantic segmentation (Yu and Koltun, 2016), we propose our alternating attention model, AltXML. Two attention heads which alternate reading sequence enable the model capture different parts of the document rather than one point. We evaluate our model on different datasets. Comparison with other models indicates that the trade-off between using correlation to reduce output space and improving the ability to refine labels needs further research. In summary, our contribution is three-fold:

- We construct an imbalanced Chinese multi-label text classification dataset, IMCM.
- We implement diverse MLC models and propose our alternating attention model.
- We conduct a detailed evaluation for these models on three datasets with different imbalance ratios, by comparing on them, our model achieves promising performance.

2 Related Work

Multi-label classification studies the problem where each example is represented by a single instance while associated with a set of labels simultaneously. There are two main types of methods for MLC task: problem transformation methods and algorithm adaptation methods.

Binary Relevance (BR) transforms the task of multi-label classification into the task of binary classification, which is simple and reversible but ignores potential correlations among labels and may lead to the issue of sample imbalance. Label powerset (LP) generates a new class for each possible combination of labels and then solves the problem as a single-label multi-class one. Classifier Chains (CC) treats this task as a sequence labeling problem and overcomes the label independence assumption of BR due to classifiers are built upon the previous predictions. In addition to traditional machine learning methods, Neural networks provide some new approaches to MLC task. These methods have made great progress in multi-label classification task, but still face the problem of high computational complexity and poor scalability to meet high-order label correlations.

CNN uses multiple convolution kernels to extract text feature, which is then input to the linear transformation layer followed by a sigmoid function to output the probability distribution over the label space. CNN-RNN incorporated CNN and RNN so as to capture both global and local semantic information and model high-order label correlations.

Nam et al. (2017) also treat the multi-label classification task as a sequence labeling problem but replace classifier chains with RNN. It allows to focus on the prediction of the positive labels only, a much smaller set than the full set of possible labels. Yang et al. (2018) propose to view the MLC task as a sequence generation problem to take the correlations between labels into account.

Typically, there are two main available multi-label text classification datasets, which all stem from English reading materials. Rcv1v2 (Lewis et al., 2004) is widely used in multi-Label classification task. It consists more than 80,000 manually classified English newswire stories, which divided by Lin et al. (2018). The total number of topic labels is 103.

AAPD (Yang et al., 2018) is a large English multi-label text classification dataset. It contains abstract and corresponding topics of 55,840 papers in the computer science field on the Arxiv. The total number of subjects is 54.

[1] https://github.com/NLPBLCU/imcm-dataset

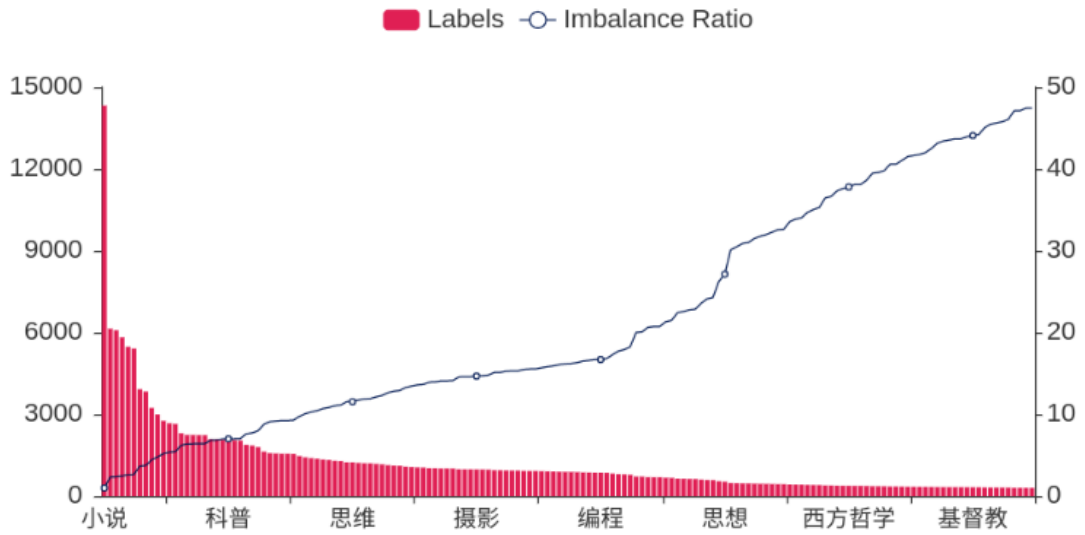


Figure 1: Distribution and Imbalanced Ratio of labels on IMCM dataet. Imbalanced Ratio is the ratio of the frequency of the label to the highest frequency.

Datasets	Inst.	Lab	Card.	Dens.	Len.	IR.	Train Set	Valid Set	Test Set
Rcv1v2	804,414	103	3.24	0.031	123.94	17.44	802,414	1,000	1,000
AAPD	55,840	54	2.4	0.044	163.43	6.58	53,840	1,000	1,000
IMCM	52,052	158	3.7	0.023	348.91	10.35	41,642	5,205	5,205

Table 1: Comparison of IMCM dataset with existing MLC datasets. Inst and Lab denote the total number of instances and labels, respectively. Card means the average number of labels per instance. DENS normalizes Card by the Lab. Len refers to the average length of the instance. IR indicates how imbalanced the top 50 percentage of labels are.

3 IMCM Dataset

For the purpose of constructing highly reliable multi-label text classification dataset, we have collected nearly 60,000 books' information from Douban [2], which consists of content summary and author introduction. Labels of each book are manually marked by members of Douban. Unlike the above described datasets, the difference among some labels in the IMCM is very subtle, such as Humanistic and Human nature. And distribution of labels is very imbalanced, which can be seen in figure 1. These characteristics make it not feasible for labels to be classified in an extensive way. Therefore, we limited the number of words per instance no less than 50 to provide adequate information. Finally, we got 52,286 documents.

In order to evaluate the data effectively, we carry

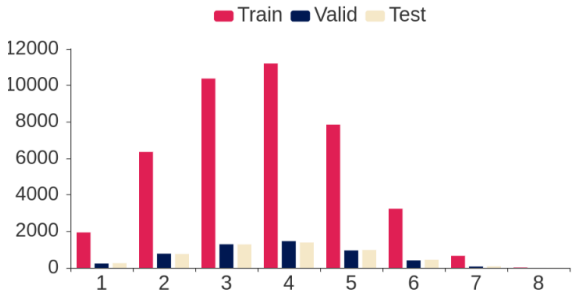


Figure 2: Distribution of the number of labels per instance.

on the same distribution sampling to the data. In the end, we got 41,829 training data, 5,228 validation data and 5,229 test data. The total number of labels is 158, the average number of labels per instance is 3.7 (can be seen in figure 2), the average length of the instance is 348.91 and the imbalanced ratio

[2] https://book.douban.com

of labels is 10.35. Comparison of IMCM dataset with existing MLC datasets can be seen in Table 1. We can see that our dataset is longer than the other two. Besides, neither like the extreme imbalance of the labels of the Rcv1v2 dataset nor like the small-scale imbalance of the labels of the AAPD dataset, our dataset makes a trade-off. This avoids the overwhelming interference caused by the extreme imbalance of data, and allows us to make some explorations on this basis.

4 Alternating Attention Model

We introduce our proposed model in detail in this section. First, we give an overview of the model in Figure 3. It consists of four layers: Word Representation Layer, Bidirectional LSTM Layer, Alternating Attention Layer, and Classification Layer.

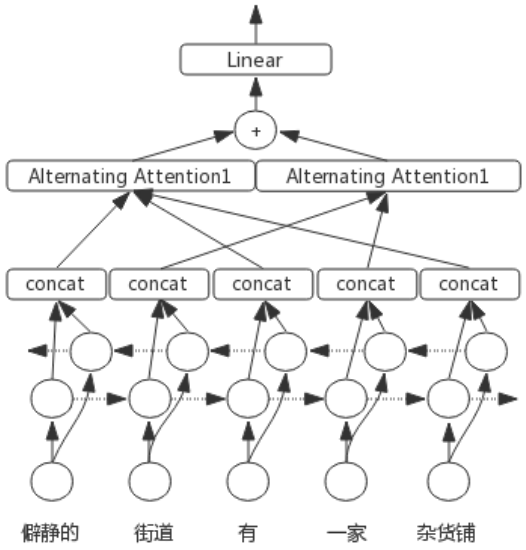


Figure 3: Overview of the AltXML model

4.1 Word Representation Layer

The input of AltXML is raw tokenized text, each word is represented by word embedding. Let T and d respectively represent the length of the input text and the dimension of word representation. The output of word representation as follows:

$$X = (x_1, x_2, ..., x_T)$$

where x_t is a dense vector for each word.

4.2 Bidirectional LSTM Layer

We use a Bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to capture both the left-sides and right-sides context at each time step, the output of BiLSTM can be obtained as follows:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_t}, C_{t-1})$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_t}, C_{t-1})$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

where h_t is obtained by concatenating forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$.

4.3 Alternating Attention Layer

We alternately send the output of the BiLSTM to the two attention layers, reduce the coupling between information, so that it is able to remove the negative effects such as information loss caused by general attention mechanism, such as focus on one key point. The output of alternating attention can be obtained as follows:

$$m_{2i} = \frac{e^{h_{2i} w_m^T}}{\sum_{t=1}^{T} e^{h_{2i} w_m^T}} ; m_{2i+1} = 0$$

$$n_{2i+1} = \frac{e^{h_{2i+1} w_n^T}}{\sum_{t=1}^{T} e^{h_{2i+1} w_n^T}} ; n_{2i} = 0$$

$$a = \sum_{i=1}^{T} Relu(m + n) * h_i$$

where m_i and n_i is the normalized coefficient of h_i.

Besides, it is able to expand the attention at the polynomial level without increasing the number of parameters. Thus, it becomes possible for alternating attention to capture longer-term dependency and avoid gridding effects caused by dilation.

4.4 Classification Layer

AltXML has one fully connected layers as output layer. Then, predicted probability $\hat{y}$ for the label can be obtained as follows:

$$\hat{y} = f(aw^T + b)$$

where, function f is sigmoid activation function.

4.5 Loss Function

We uses the binary cross-entropy loss function, which was used in XML-CNN (Liu et al., 2017) as the loss function. The loss function is given as follow:

$$L(\theta) = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})$$

where N is the number of samples, L is the number of labels, $\hat{y}_{ij} \in [0, 1]$ and $y_{ij} \in \{0, 1\}$ are the predicted probability and true values, respectively, for the i-th sample and the j-th label.

5 Experiments

5.1 Setting

Training details of neural network models are illustrated as follows.

- Vocabulary: For training efficiency and generalization, in all datasets, we truncate the full vocabulary and set a shortlist of 60,000. Note that, for Chinese, we use Jieba [3] to cut words and not use domain dictionary.
- Embedding layer: We set word embedding dimension to 256 and use randomly initialized embedding matrix with the normal distribution $\mathcal{N}(0, 1)$. Note that, no pre-trained word embeddings are used in our experiments.
- BiLSTM layer: We use single-layered bidirectional LSTM that output dimension in each direction is 100, and randomly initialized it with uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where $k = \frac{1}{\text{hidden\_size}}$. As LSTM still suffers from the gradient exploding problem, we set gradient clipping threshold to 10 in our experiments.
- Dropout: We used Dropout after embedding layer and set dropout ratio to 0.5.
- Optimization: We used the AdamW optimizer (Loshchilov and Hutter, 2018) with an initial lr = 0.001 and wd=0.01. The batch size is set to 64.
- Training: We trained model for 20 epochs and choose the best model according to the performance of validation set.

[3] https://github.com/fxsjy/jieba

Note that, the hyperparameters are consistent across all datasets.

5.2 Evaluation Metrics

We used the micro-F1 score as our main evaluation metrics. micro-F1 (Mi-F1) can be interpreted as a weighted average of the precision and recall. It is calculated globally by counting the total true positives, false positives, and false negatives.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$micro\text{-}F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.3 Baselines

- **Binary Relevance (BR)** (Boutell et al., 2004) transforms the task of multi-label classification into the task of binary classification, which is simple and reversible but ignores potential correlations among labels and may lead to the issue of sample imbalance.
- **Label powerset (LP)** (Tsoumakas and Vlahavas, 2007) generates a new class for each possible combination of labels and then solves the problem as a single-label multi-class one.
- **Classifier Chains (CC)** (Read et al., 2011) treats this task as a sequence labeling problem and overcomes the label independence assumption due to classifiers are built upon the previous predictions.
- **CNN** (Kim, 2014) uses multiple convolution kernels to extract text feature, which is then input to the linear transformation layer followed by a sigmoid function to output the probability distribution over the label space.
- **CNN-RNN** (Chen et al., 2017) incorporated CNN and RNN so as to capture both global and local semantic information and model high-order label correlations.
- **SGM** (Yang et al., 2018) **(state-of-the-art)** views the multi-label classification task as a

sequence generation problem, and apply a sequence generation model with a novel decoder structure to solve it.

- **RNN+att** is our implementation of the RNN-based model with the normal attention mechanism.

6 Results

The results of AltXML and baseline models on our IMCM dataset are presented in Table 2. From the results of the conventional baselines, it can be found that the machine-learning-based methods for multi-label text classification still own competitiveness compared with the deep-learning-based methods.

For the generating model, the SGM+GE achieve significant improvements on the IMCM dataset, compared with the machine-learning-based models. However, there is still a certain gap compared with the classification model. By contrast, our proposed model can capture more key features at the same time and achieve the best performance in the evaluation of micro-F1 score, which improves 6.1% of micro-F1 score compared with the SGM+GE.

Model	Mi-P	Mi-R	Mi-F1
BR	76.8	36.8	49.8
CC	70.5	39.9	51.0
LP	50.7	44.9	47.6
SGM+GE	60.6	54.3	57.3
RNN+Att	69.2	57.2	62.6
AltXML	70.0	57.8	**63.3**

Table 2: Results on IMCM Dataset.

We also implement our experiments on public datasets. On the AAPD dataset, similar to the models' performance on the IMCM dataset, our AltXML model achieved good performance, with a 0.8% increase in micro-F1 scores compared to the best, as shown in Table 3.

On the Rcv1v2 dataset, our AltXML model still achieves similar performance on micro-F1 on this dataset comparded with Seq2Seq model (SGM+GE), which illustrates the robustness of our model. Because we have not adjusted the hyperparameters, there is still a lot of space for improvement. The results can be seen in Table 4.

Model	Mi-P	Mi-R	Mi-F1
BR	64.4	64.8	64.6
CC	65.7	65.1	65.4
LP	66.2	60.8	63.4
SGM+GE	74.6	67.5	71.0
RNN+Att	72.0	69.7	70.8
AltXML	71.8	71.9	**71.8**

Table 3: Results on AAPD Dataset.

Model	Mi-P	Mi-R	Mi-F1
BR	90.4	81.6	85.8
CC	88.7	82.8	85.7
LP	89.6	82.4	85.8
CNN	92.2	79.8	85.5
CNN-RNN	88.9	82.5	85.6
SGM+GE	89.7	86.0	**87.8**
RNN+Att	89.1	85.2	87.1
AltXML	90.1	84.6	87.2

Table 4: Results on Rcv1v2 Dataset.

An interesting finding is that, by comparing on three datasets, although the Seq2Seq models achieves the state-of-the-art performance on the Rcv1v2 English dataset, the generalization on our IMCM dataset is insufficient. We think there are two reasons: (1) Compared to the other two datasets, the number of labels for each instance in our dataset is more and there are no obvious semantic boundaries among some labels. (2) Due to the attention mechanism cannot improve the performance of the Seq2Seq model in this task (Lin et al., 2018), Seq2Seq model cannot capture some useful information.

By comparing on the three datasets, our model achieves promising performance.

7 Conclusions

In this paper, we introduce the first Chinese multi-label text classification dataset, IMCM. This dataset focuses on imbalanced multi-label classification. Among many datasets, our model could also give significant improvements over various state-of-the-art baselines. Furthermore, we propose an alternat-

ing attention model to handle the imbalanced problems, and further analysis of experimental results demonstrates that our proposed model not only capture the correlations between labels, but also capture the more features when predicting different labels.

Acknowledgements

This work was supported by Beijing Natural Science Foundation(4192057). We thank anonymous reviewers for their helpful feedback and suggestions.

References

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 730–738. Curran Associates, Inc.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757 – 1771.

G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria. 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2377–2383, May.

Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Comput. Surv.*, 47(3):52:1–52:38, April.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *empirical methods in natural language processing*, pages 1746–1751.

David Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Junyang Lin, Xu Sun, Pengcheng Yang, Shuming Ma, and Qi Su. 2018. Semantic-unit-based dilated convolution for multi-label text classification. *empirical methods in natural language processing*, pages 4554–4564.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. 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. ACM.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. *arXiv: Learning*.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5413–5423. Curran Associates, Inc.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 993–1002, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, Jun.

Yukihiro Tagami. 2017. Annexml: Approximate nearest neighbor search for extreme multi-label classification. pages 455–464. the 23rd ACM SIGKDD International Conference, 08.

Grigorios Tsoumakas and Ioannis Vlahavas. 2007. Random k-labelsets: An ensemble method for multilabel classification. In Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 406–417, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926. Association for Computational Linguistics.

Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. 2016. Pd-sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 3069–3077, New York, New York, USA, 20–22 Jun. PMLR.

Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *ICLR*.

How State-Of-The-Art Models Can Deal With Long-Form Question Answering

Minh-Quan Bui, Vu Tran, Ha-Thanh Nguyen, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
{quanbui,vu.tran, nguyenhathanh, nguyenml}@jaist.ac.jp

Abstract

Question answering is an essential task in natural language processing. According to our review, the datasets for this task often contain short answers. We raise a research question of whether state-of-the-art models can perform well when a longer answer is needed. We propose a dataset that contains much longer answers called FitQA[1] and conduct a brief performance analysis among current state-of-the-art models on this dataset. The robust transformer architecture like ALBERT achieved 90.9% F1 on SQuAD 2.0 but only got 47.3% F1 score on FitQA. Our hypothesis is that for longer context, the model needs to be guided to focus on longer dependent words. We conduct a curriculum-learning-based framework. Experimental results show that our approach could improve the performance with the appropriate answer length up to 55.3% on F1.

Acknowledgments: This work was supported by JST CREST Grant Number JPMJCR1513 and the Asian Office of Aerospace R&D (AOARD), Air Force Office of Scientific Research (Grant no. FA2386-19-1-4041)

1 Introduction

Machine reading comprehension (MRC), or the ability to read and understand the unstructured text and answer questions about it remains a challenging task in natural language understanding. This challenge spurs the development of large datasets and deep learning architectures. The current state-of-the-art models can overwhelm the human performance. RoBERTa (Liu et al., 2019) performs well with 89.4%F1 score on SQuAD. Besides, A Lite BERT(ALBERT) (Lan et al., 2019) is even better when getting 91.365% F1 score and 88.716% exact match(EM) on SQuAD 2.0. In contrast, the human performance only gets 89.5% F1 score and 86.8% EM. Also, Microsoft has already built a high-quality dataset called NewsQA (Trischler et al., 2017), a challenging dataset with more than 100.000 question-answer pairs, But a new improving method

Article: 5 Fat Loss Myths You Still Believe
Question: How to Limit Metabolic Compensation?
Context: How to Limit Metabolic Compensation The good news is there are some ways to reduce metabolic compensation. Here are some things to do: **Do your best to maintain as much muscle as you can. The metabolic rate will not slow as much and be more resistance to fat regain. This means to make weight lifting the dominant part of your fitness regime during fat loss. Cardio becomes a little more important after weight loss, when the metabolic rate has lessened. You may want to save your cardio for after, rather than during the competition diet. Eat more protein, see the first point above about maintaining muscle mass. And probably increase the amount of protein as a percent of total calories. Do this during, but perhaps more importantly, after fat loss**. Cycle the calorie gap, having times where you're in a strong deficit and other times where you're in no deficit at all. The recent MATADOR study (minimizing adaptive thermogenesis and deactivating obesity rebound) showed this strategy got better results, had less metabolic adaptation, and much longer lasting results. Don't eat like an asshole when it all ends. Focus on blander foods and less variety of them. Doing the traditional burger, pizza, and cheesecake binges will trigger the brain's hedonistic response and cause you to want more of that same dopamine hit all this when the metabolism is at its most vulnerable in terms of fat storage. And finally you may want to consider some type of adaptogen like rhodiola or ashwagandha. I have no studies to back this up, but I have very good success clinically with using these herbs along with the recommendations above to keep the command and control center of the metabolism (the brain's hypothalamus) stress-resistant and happy.

Figure 1: Example for FitQA dataset with the answer is in bold

[1]The dataset will be published along with the paper

for BERT also known by the name SpanBERT (Joshi et al., 2020), which masking spans instead of token masks and the performance when applied to NewsQA, has 73.6% F1 score on NewsQA. In Robin Jia's work (Jia and Liang, 2017), his proposed method test whether a model can give a correct answer while paragraphs contain some additional sentences, which are noise for deep learning models. This method worked well by decreasing the accuracy of sixteen models drops from 75% F1 score to 36%. SQuAD-Open is built by Chen (Chen et al., 2017), an open domain question answering dataset and contains only question and answer. The model has to extract the response by the relevant context from Wikipedia articles. This problem seems to be a trend in question answering datasets. While deep learning models are becoming more powerful and reaching human performance, more complex datasets are needed to accelerate method and model development.

We build the FitQA dataset to contribute to not only computer science but also for the entire society. FitQA is collected by crawling more than 200 articles from `bodybuilding.com` (bod, 1999) and `t-nation.com`(LLC, 1998). The main purpose of this dataset is for the health of society. On the internet, we have a lot of fakes and lack of information news about nutrition and training like 30 days sit up for sick-pack or deltoid drinking for losing weight, etc. We cannot explain knowledge with a short sentence, that is why we want FitQA to be very detailed and diverse in answer. Figure 1 shows an example of the phenomena of FitQA. We experiment with different models and find that FitQA creates a significant challenge to current comprehension models. In this paper, we also describe curriculum learning (Bengio et al., 2009), a training approach for state-of-the-art models with the maximum of answer length (MAL) is 30 and 60 to handle low resource limitation.

2 Related Datasets

FitQA is built following the format of some traditional comprehension datasets. These vary in length, size, problem, collection, and each has its distinctive feature.

2.1 NewsQA

NewsQA (Trischler et al., 2017), a machine comprehension dataset with more than 100K question-answer pairs. These pairs are created by human and based on a set of over 12K news articles from CNN. The answers consist of spans of text in the articles. In NewsQA, they created some conditions for answers to make the dataset more complex:

1. Word Matching: The similarities between questions and answers are low; this condition makes deep learning architecture harder to extract the answer from the article.

2. Paraphrasing: In each article, it must contain one sentence that can answer the question. This sentence requires synonym and global knowledge.

3. Inference: Their answers must be found from a piece of information in the article or by overlap.

4. Synthesis: The answers are only found by the assumption of information through several sentences.

5. Ambiguous/Insufficient: Some questions do not have answers in the article.

Because of the complexity, this dataset is challenging for transformer architectures. To overcome this challenge, SpanBERTJoshi et al. (2020) was developed with a new training approach by masking random spans instead of random tokens and training the span boundary representation for predicting the whole content of the masked span, without depending on the token representations within it. SpandBERT achieved 73.6% F1 on NewsQA, that 83.6% F1 on TriviaQA, 84.8% F1 on Search QA and more significant results on other datasets.

2.2 TriviaQA

TriviaQA (Joshi et al., 2017) was collected into 650K question-answer-evidence triples. TriviaQA has over 95K question-answer pairs and at least six evidence documents for each question-answer pair. In this dataset, up to 92% of the answers are the article titles in Wikipedia, about 4% are numerical answers, and the rest are free texts. The challenge in TriviaQA is the overlap of each example in sev-

Length(word)	Proportion
1-10	17.7%
11-20	22.7%
21-30	12%
31-40	10%
41-50	12.8%
51-60	10%
61-70	4.9%
> 70	9.9%

Table 1: Length statistics.

eral categories. It means each question-answer pair can be found in multiple evidence documents. Right now, the first place on the TriviaQA leaderboard is 83.99% on F1 score.

2.3 SQuAD 2.0

SQuAD 2.0, also called SQuADRUN, is created base on SQuAD, but higher difficulty which contains more than 130K examples in over 442 articles. Beside answerable questions, it has more than 50K unanswerable questions that look similar to answerable questions. To perform well on SQuAD 2.0, the models also need to decide to answer the question or not when the context does not support the answer. Despite the fact that many challenges in SQuAD 2.0 dataset, the state-of-the-art model can surpass 90% on F1 score.

3 FitQA

For this work, we analyzed the data collected from `bodybuilding.com` (bod, 1999) and `t-nation.com`(LLC, 1998). These contain a varied topic that includes nutrition, training, diet, fat loss, etc. FitQA focuses on topics that people are often interested in like nutrition or training. By observing the habit of the people asking questions from some popular forums, we can conclude that long answers usually satisfy the questioner better because it contains more relevant and useful information. However, sometimes, a short answer is all they need. Based on that observation, we build FitQA as a data set of variable length answers. FitQA has almost 700 question-answer pairs, the length of an answer is from 1 to 139 words. FitQA follows the format of SQuAD 2.0 dataset, and answers are extracted from spans of text in the article. The different and challenges that make it different from SQuAD 2.0 are as below:

1. The average length of articles in FitQA is double of that in SQuAD 2.0.

2. The average length of answers in FitQA is ten times higher than SQuAD 2.0.

The state-of-the-art models are overwhelming human performance, and the dataset must be harder and more challenges to be able to achieve some important achievements in machine comprehension task. There are some participants in this case, NewsQA (Trischler et al., 2017) have more challenges than SQuAD (Rajpurkar et al., 2016) by having less word matching examples(7.1%), more paraphrasing example(7.3%) and more synthesis and inference examples(13.4%). On the other hand, TriviaQA has 69% questions that have different syntactic structure and 41 % of them have lexically different. Moreover, the information needed to answer the question is scattered over multiple sentences. Based on these ideals, we increase the complexity of FitQA by the diversity in answer length. The length statistics is showed as Table 1. To test the performance of state-of-the-art models, we create the test set by picking 100 examples with varied length, and the rest is for the training set.

4 Curriculum Learning

We examine some previous work and propose a useful method for handling the long articles and answers. Curriculum learning (Bengio et al., 2009) is a learning strategy in machine learning, we let the deep learning model learn with easy examples first and then gradually handles harder cases. Several works have shown that this problem can be overcome by using this learning strategy. As a result of Cao Liu (Liu et al., 2018) in his natural answer generation task, curriculum learning can increase his model performance by 6.8% and 8.7% in the accuracy for easy and hard questions.

In our question-answering task, we defined the complexity by the length of the answer. We assume that an example containing a short answer is easy, and an example having a long answer is difficult. We want models to learn from easy to difficult sample

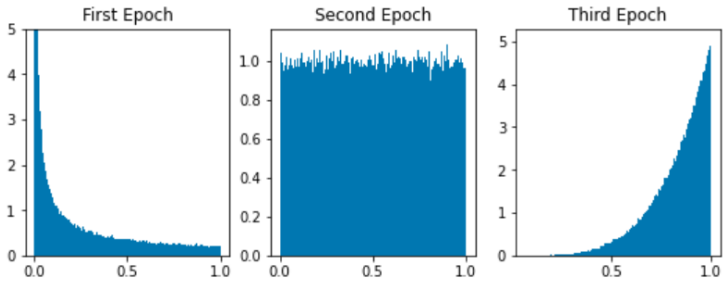


Figure 2: Illustration of the probability of picking example by curriculum learning with 3 epochs and temperature base $\gamma = 5$

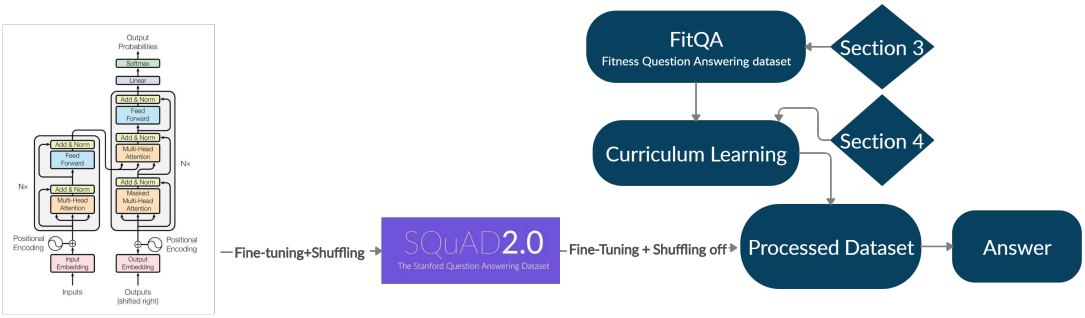


Figure 3: The overall diagram of question-answering task bases on transformer model and curriculum learning

and from the difficult to easy sample. We give every single example a score equal to its answer length. After that, we sort the list of examples in ascending order by scores to pop the sample by its index easily. The formula and pseudo-code of curriculum learning (Algorithm 1) to calculate the probability of picking an example $Index$ from n examples as below:

$$index_{ij} = \lfloor nx_{ij}^{t_i} \rfloor \tag{1}$$

$$x_{ij} \sim U[0, 1) \tag{2}$$

where t_i is the temperature of epoch i^{th} obtained by:

$$t_i = \gamma^{\alpha_i} \tag{3}$$

$$\alpha_i = \frac{2(1-i)}{N_{epochs} - 1} + 1 \tag{4}$$

where γ is a temperature base, N_{epochs} is the total number of training epoch. The temperature base γ reflects how high the probability of picking a long or short example through the epoch.

We illustrate the probability of picking example curriculum learning as Figure 2. In the first epoch,

Algorithm 1: Curriculum learning pseudo-code

Result: Sample by length
Input:
The list of n_{sample} samples is sorted in ascending order:SL;
γ is temperature base;
N_{epochs} is total number of epochs;
Output:;
The list of n_{sample} samples is arranged by curriculum learning:OL;
Function: curriculum_learning(SL,γ,N_{epochs}):
OL $\leftarrow$ {}
for *i = 1 to* N_{epochs} **do**
  $temp_{Li}$ = SL
  $\alpha_i = \frac{2(1-i)}{N_{epochs}-1} + 1$
  $t_i = \gamma^{\alpha_i}$
  for *j = 0 to* n_{sample} **do**
    x_{ij} = random(0,1)
    $index_{ij}$ = length($temp_{Li}$)
    OL.push($temp_{Li}[index_{ij}]$)
    $temp_{Li}$.pop($index_{ij}$)
  end
end

Table 2: *F1 and Exact Match(EM) scores on FitQA with MAL=30*

Model	FitQA uniform		SQuAD 2.0+FitQA uniform		SQuAD 2.0+FitQA γ = 2		SQuAD 2.0+FitQA γ = 3		SQuAD 2.0+FitQA γ = 5	
	EM(%)	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***
albert-base v1	11.1	34.9	16.9	42.5	15.3	43.1	16.7	44.0	15.0	42.1
albert-base v2	12.1	37.8	15.3	42.5	14.3	43.5	14.4	43.2	17.0	**45.6**
bert-base-uncase	6.9	31.1	17.8	45.0	17.8	45.0	17.0	44.2	16.7	**46.1**
roberta-base	5.6	25.1	14.1	42.7	14.3	**45.5**	14.7	42.3	14.4	43.8

Table 3: *F1 and Exact Match(EM) scores on FitQA with MAL=60*

Model	FitQA uniform		SQuAD 2.0+FitQA uniform		SQuAD 2.0+FitQA γ = 2		SQuAD 2.0+FitQA γ = 3		SQuAD 2.0+FitQA γ = 5	
	EM(%)	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***
albert-base v1	8.8	38.6	18.3	51.3	15.7	51.3	16.7	50.8	17.0	51.4
albert-base v2	17.0	47.3	21.2	54.2	16.7	51.5	19.3	52.0	18.3	53.0
bert-base-uncase	6.9	31.1	17.0	52.6	15.4	51.4	20.33	53.9	17.7	52.1
roberta-base	4.9	30.9	19.0	52.7	18.0	53.3	20.3	53.7	19.0	**55.3**

we can easily see that the probability of picking the examples with short answer is high and it is pretty low with examples with long answer. In the second epoch, the probability of picking example is uniform distribution. In the last epoch, the probability of picking the examples with long answer is extremely higher than the short answer examples.

5 Experiment

5.1 Experiment Settings

From SQuAD and NewsQA leaderboard, there are some approaches perform better performance, but all of them are built base on the pre-trained model. To test FitQA for the machine comprehension task, we compare the performance of four common pre-trained deep learning models: bidirectional encoder representations from transformers (BERT), two versions of a Lite BERT (ALBERT), and RoBERTa. We describe details of all the pre-trained models as below:

1. bert-base-uncased: 12 layer, 768 hidden, 12 heads, 110M parameters, and trained on low-cased english text.

2. albert-base-v1: 12 layer, 768 hidden, 128 embedding, 12 heads, 11M parameter.

3. albert-base-v2: 12 layer, 4096-hidden, 128 embedding, 64-heads, 223M parameters.

4. roberta-base:12-layer, 768 hidden, 12-heads, 125M parameters RoBERTa using the BERT-base architecture.

We conduct experiments on SQuAD, FitQA. Performance on these datasets is measured by exact match (EM) and per answer token-based F1 score, which was published by Rajpurkar et al(2016) (Rajpurkar et al., 2016). The detailed settings are described as below:

1. **FitQA uniform:** Using 4 pre-trained models to test the performance on FitQA with uniform probability of picking examples.

2. **SQuAD 2.0 + FitQA uniform:** Using 4 pre-trained models, we first fine-tune the models on SQuAD 2.0, then fine-tune the models again on FitQA with uniform probability of picking examples.

3. **SQuAD 2.0 + FitQA (Curriculum Learning):** Using 4 pre-trained models, we first fine-tune the models on SQuAD 2.0 with uniform probability of picking examples, then fine-tune the models again on FitQA with curriculum learning with different γ (2, 3 and 5).

Figure 3 can illustrate the whole process of setting 2 and 3.

5.2 Main Result

According to information from SQuAD 2.0 leaderboard, the best performance that albert single can archive is 88.592% EM score and 91.286% F1 score. However, in section III, we showed that the length of some examples in FitQA are extremely long and diverse. This is the reason makes 4 state-of-the-art models cannot work well on FitQA. In our experiments, albert-base-v2 has the best result but only 37.8% F1 score and 12.1% on EM. SQuAD 2.0 is the most similar dataset to FitQA. To maximize the performance, we firstly train all models on SQuAD 2.0 and fine-tune FitQA. After training on SQuAD and fine-tune FitQA the performance increase 5.68% on EM and 8.9%F1 score on average. Next, we mute the shuffling feature, then apply curriculum learning to the training set. We start with temperature base γ = 2 and MAL=30. As the results in Table 2, curriculum learning made average F1 score from 43.2% to 44.3%. Especially, it can increase the performance of roberta-base by 2.8%. With γ = 3, there are no significant improvement. We increase γ base γ to 5, and we can get the best results with 46.1% and 45.6% on bert-base-uncase and albert-base-v2. Next, we want model to face the harder challenge by increasing MAL to 60, and it leads to good results, roberta-base gets the best result overall with 55.3% on F1 even if the improvement is negligible.

5.3 Result Analysis

With MAL=30,Bert-base-uncased and albert-base v2 with temperature base γ = 5 seems to be the best for FitQA, so we analyze the results and compare them to bert-base-uncase without curriculum learning. As a result of Table 3, we show the accuracy of the answer group was mentioned in Table 1. By comparing the results from these settings, we expect to determine that curriculum learning is useful for extracting more text or capture more related information to answer the question. In lengths from 0 to 10 words, we can see there is no significant change between all settings. Starting from 11 words, we can see that these results go beyond the uniform distribution setting. With temperature base γ = 5 from Table 5, we can see bert-base-uncase works well in lengths from 11 to 60 words, and the performance in this range increase 2.08% on average compare to uniform distribution bert-base-uncase. Albert-base-v2 can also perform well in this range with 3.44% increase in total. TABLE 6 summarizes overall statistics of 3 best settings on FitQA with MAL=60. It is worth discussing these interesting facts revealed by the results of bert-base-uncased . The test in range 41 to more than 70 words found differences from bert-base-uncase compare to albert-base-v2 with 2.8% improvement on F1 score.

One limitation is found in these experiments. From TABLE 4, the most extended answer can be extracted is 50 words. It means for the examples have answer more than 50 words, the EM score will be zero. Not only that, but it is also hard to extract long answer correctly from the context, and some answers are a subset of gold answers. This may be the reason why EM score equal to 0 in some evaluations. We show several examples to demonstrate for this limitation in Table 7. The answers are extracted by models is not wrong, but not enough in these cases.

6 Conclusion

As the results are shown in Table 5 and Table 6, we have succeeded in improving the length that the model can extract by applying curriculum learning on the training set. This success leads to an increase in F1 score. The problem is all the state-of-the-art models perform poorly under long answer form dataset. The best result that these models can get is just 21.2% on EM and 55.3% on F1 score. We believe that the long-form answer dataset is the big challenge for machine comprehension task. Further, we want to apply curriculum learning not only base on the length of the answer but also other features to solve the low performance of state-of-the-art models on FitQA.

References

Bodybuilding. 1999. URL `https://www.bodybuilding.com`. Accessed:2020-1-25.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Model	Total Tokens	Longest Answer	Shortest Answer
bert-base-uncase γ = 5,MAL=30	1462	28	1
albert-base-v2 γ = 5,MAL=30	1384	28	1
bert-base-uncase uniform,MAL=30	1271	28	1
albert-base-v2 uniform,MAL=60	2255	46	1
roberta-base γ = 5,MAL=60	1980	42	1
bert-base-uncase γ = 3,MAL=60	2330	50	1

Table 4: *Total number of tokens, longest and shortest answer that models can extract from 100 examples of test set*

Length	bert-base-uncase γ = 5		albert-base-v2 γ = 5		bert-base-uncase uniform	
	EM(%)	*F1(%)*	*EM(%)*	*F1(%)*	*EM(%)*	*F1(%)*
1-10	41.7	62.4	47.9	58.5	50	63.4
11-20	37.0	**59.1**	40.7	**64.0**	38.8	57.0
21-30	20.5	**52.9**	10.3	46.3	20.5	50.9
31-40	0.0	43.4	0.0	**48.3**	0.0	44.8
41-50	0.0	**34.6**	0.0	**35.0**	0.0	30.4
51-60	0.0	**37.7**	0.0	**40.9**	0.0	34.2
61-70	0.0	34.7	0.0	32.4	0.0	37.7
>70	0.0	23.8	0.0	21.4	0.0	24.3

Table 5: *F1 and Exact Match(EM) scores on best settings base on length with MAL=30*

Length	albert-base-v2 uniform		roberta-base γ = 5		bert-base-uncase γ=3	
	EM(%)	*F1(%)*	*EM(%)*	*F1(%)*	*EM(%)*	*F1(%)*
1-10	35.4	49.4	47.9	**67.9**	29.1	47.5
11-20	37.0	63.9	25.9	62.2	44.4	59.9
21-30	46.2	67.2	35.9	61.5	30.8	65.0
31-40	30.3	61.2	15.1	53.4	21.1	57.5
41-50	0.0	44.5	0.0	**46.1**	0.0	**46.9**
51-60	0.0	56.6	0.0	**56.9**	0.0	**56.8**
61-70	0.0	47.2	0.0	41.8	0.0	**50.0**
>70	0.0	35.8	0.0	36.3	0.0	**41.6**

Table 6: *F1 and Exact Match(EM) scores on best settings base on length with MAL=60*

D. Chen, A. Fisch, J. Weston, and A. Bordes. 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 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://www.aclweb.org/anthology/P17-1171.

R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://www.aclweb.org/anthology/D17-1215.

M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://www.aclweb.org/anthology/P17-1147.

M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

C. Liu, S. He, K. Liu, and J. Zhao. Curriculum learn-

Question	Gold Answer	Predicted Answer
Are Testosterone Boosters Safe?	Always read re-views before purchasing, and choose a testosterone booster from a reputable, established supplement company. Only take the recommended dose and keep your doctor in the loop about what you're taking if you have other health concerns or take medications.	Always read reviews before purchasing, and choose a testosterone booster from a reputable, established supplement company.
what is the difference between brown fat and white fat?	Both types store energy, but white fat cells each contain only one droplet of fat, while brown fat contains lots of tiny droplets of fat. Brown fat also contains tons of the brownish cellular organelles known as mitochondria, which use the droplets of fat to create energy and, as a byproduct of creating energy, heat.	Both types store energy, but white fat cells each contain only one droplet of fat, while brown fat contains lots of tiny droplets of fat.

Table 7: *2 Examples for the most limitation of FitQA*

ing for natural answer generation. In *IJCAI*, pages 4223–4229, 2018.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

T. N. LLC. Bodybuilding. 1998. URL `https://www.t-nation.com/`. Accessed: 2020-2-02.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 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, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL `https://www.aclweb.org/anthology/W17-2623`.

Prosody Features of Collaborative Construction in Mandarin Conversation

GUAN Yue
The Hong Kong Polytechnic University – Peking University Research Centre on Chinese Linguistics, Beijing
guanyue92@sina.com

Abstract

This paper examines the projection of prosody and syntactic structure in conversation, while trying to find out which side is more powerful when there is a conflict between prosodic expression and syntactic structures. The central task of this paper is to look at the phonetic performance of the collaborative construction from the general indicators of prosodic features: duration (length), loudness (intensity) and frequency (pitch). It compares phonetic parameters of the collaborative construction with the adjacent turns and describes it in detail.

1 Introduction

In everyday face-to-face conversation, language forms are always accompanied by prosodic features of speech.

In this paper, the author examines the basic prosody phenomenon in collaborative construction. In daily conversation, a syntactically complete sentence can be co-constructed by different speakers in adjacent turns, which is referred to as collaborative construction (CC)[1] in this paper. The utterance produced by interactional participants can project the remaining part of a sentence-in-progress, which is the basis of syntactic cooperation.

It has been sufficiently observed that speakers can express their stance and emotion prosodically (Benjamin and Walker 2013); in fact, prosody also plays a very crucial role in turn-taking and turn organization. For example, Walker (2010) examines the mechanism of rush-through in English as a phonetic design of turn-holding and points out that the articulation rate of the final foot at the end of a TCU (Turn Constructional Unit) roughly doubles as that of the preceding one. Local (2005) discusses some prosodic features of collaborative completion when speakers and listeners are undertaking interactional tasks in everyday talk. These studies focus on how prosodic features can be recognized by listeners as a projection of a possible position of turn-taking (Goodwin 1986; Ford 1993, 2004; Couper-Kuhlen and Selting 1996; Walker 2013).

2 Background

This paper takes an Interactional Linguistic perspective and adopts methodology of Conversation Analysis. Interactional Linguistics emphasizes on studying language in its home environment, manifested as conversations in everyday life. Conversations are “cooperatively achieved

[1] Here is an example quoted from Lerner (1987:16):
David: so if one person said he couldn't invest
(.)
Kerry: then I'd have ta wait
The two speakers co-produced one sentence.

objects", which "need to be adaptable to the emerging and ever-changing trajectory of interaction" (Couper-Kuhlen and Selting 2018). In a conversation, "overwhelmingly, people talk in turns" (Sacks *et al*. 1974). In Conversation Analysis, turn-constructional units (TCUs) are the basic units of conversation that compose turns. TCUs are addressed to their speakers, make relevant the next action and select the next speaker (Lerner 2004). This phenomenon is prevalent in the "compound turn-constructional unit format" (Lerner 1991). Such format may be built with two clauses and sometimes delivered in two separate prosodic units, but is oriented to by participants as one single turn-constructional format (Lerner 1991, 1996). Some researchers have attempted to account for other issues in cross-linguistic perspective based on data from different languages, some of which focus on turns and increments (Ono & Couper-Kuhlen 2007; Luke & Thompson & Ono 2012). Despite the considerable number of research on collaborative TCUs in English (Goodwin and Goodwin 1987; Goodwin 1996; Ford *et al*. 1996), German (Auer 1996; Stivers *et al*. 2009) and Japanese (Hayashi 2005; Iwasaki 2009); studies on Chinese data are still scarce (Song 2019; Guan 2020).

In the field of pragmatics and sociolinguistics, Gumpertz (1982:100) mainly investigated six prosodic features, including intonation, loudness, stress, vowel length, phrasing and overall shifts in speech register. The prosodic features here refer to the pitch, intensity, duration and other qualities that can be recognized by language speakers rather than the acoustic features of the actual voice under the experimental condition. Computers can capture much more accurate acoustic features than human ears, but those are semantically nonsignificant for communication, for the prosodic function of linguistics cannot be directly explained from the experimental data (Gumpertz 1982:108). For example, studies have shown that people's perception of pitch is not the reflection of the actual pitch, but the result of pitch obtrusions as prosodic perception. This kind of consequence separates the utterances that achieves prosodic prominence from that of none saliency, and searches for an approximate preset pattern in the referential mental schemata according to the perceived voice.

With the development of CA, other related factors have been included in the study of prosody. One of them is phonation type. Phonation type refers to the state of glottis when speaking, such as "voiced" and "voiceless". Zhu Xiaonong (2010:66) summarizes 12 kinds of six types of phonation on the basis of previous studies. Phonation studies the change of pronunciation compared to the "default articulatory configuration"[2]. In previous studies, phonation type is mainly used to describe the special pronunciation mechanism of other ethnic minority languages and dialects (Chao 1922, 1929; Zhu Xiaonong 2005, 2010). These studies of phonetics are major and important, but we still lack studies on the ability of human beings to master the sound and to use language. Recently, there are a few relevant research results. For example, Zhu Xiaonong (2004) points out that cracking voice can not only distinguish tones, but also have diminutive meaning; falsetto can also be used as a high-key diminutive tone. In this research on collaborative construction (CC), it is found that, in naturally occurring data, different phonation strategies are widely used in daily conversation, and they have conventionality, which is represented by prosodic projection[3].

Prosodic factors sometimes overwhelm syntactic factors on influencing the utterances of speakers and listeners (Ford 1993; Ford and Thompson 1996; Couper-Kuhlen and Selting 2018). For example, if an utterance

[2] In this paper, the default articulatory configuration refers to that the vocal cord is in a normal state at the beginning of phonating. At this time, the tone value of the voice is "the default pitch". This pitch is the most basic and its value is [32], which is calculated according to Zhu Xiaonong's four-degree system (Zhu Xiaonong 2010).

[3] Projection means that the earlier part of a structure foreshadows its later trajectory and thus makes its completion predictable (Couper-Kuhlen & Selting, 2018:39).

does not end in the form of language, but there is a pause caused by hesitation, the listener will often have the opportunity to take the turn. In collaborative-constructed sentences, turns that the preliminary speaker wants to transfer are differently designed in phonetic factors compared with the turns he wants to keep. When the subsequent speaker continues a sentence, it shows an on-going intonation in the prosodic characteristics of the turn, in comparison of pitch, comprehensive intensity and other phonetic features.

3 Data and Methods

This research uses visible data from face-to-face daily conversations in Beijing Mandarin, which were collected from July 2018 to May 2019. These dyad conversations were all naturally occurring interactions between friends and/or acquaintances. During recordings, researchers were absent and thus non-interactional factors could be diminished to the greatest extent. The data were transcribed and annotated to reflect the features of talk-in-interaction including co-existing movements as faithfully as possible.

We adopt an empirical method with a focus on language use from the perspective of Interactional Linguistics. It is widely believed in usage-based approach that ordinary conversation is a primordial site for language and thus constantly shapes forms and meanings of language (Couper-Kuhlen and Selting 2001, 2018). In addition, this study integrates a conversation analytical method to examine cases from our own data. We analyze in detail with reference to the observable orientation of both speakers and recipients within the moment-by-moment unfolding of interaction. Research has shown that multimodal resources jointly work together to build turns and courses of action (Li 2014), and that the turn-taking process is composed of syntax, prosody, body movements and other pragmatic factors.

4 Findings

Shen Jiong (2003) points out that the hearing system of humans do process phonetic quality, duration, intensity and pitch separately, but they become recognizable and meaningful only after putting together in human brains. This paper is not a special study of phonetics. The discussion in this paper is mainly designed to explore the prosodic performance of CC and its interactional function. This research tries to avoid talking about some blurred situations such as long gap or overlapping turns[4] just because of difficulties of measurement. Even though, there are still inescapable problems such as dealing with the relationship between "big wave and small wave[5]" between tone of words and intonation. In addition, there are also the effects of phonation on the pronunciation and the influence of non-verbal factors.

It should be noted that from the perspective of prosody, two types of CC can be clearly distinguished. One is peacefully co-constructed and the other is competitive "collaborative completion". These two types of prosodic features have systematic differences. However, they intertwined with the syntactic categories.

4.1 Duration

Previous researchers on prosodic in Chinese have done a lot of experiments on the length of sound, and have drawn many important conclusions (see Zhu 2010). They find out that the duration of words has an important impact on the recognition of tones and stresses. Through the analysis of the CC data, the preliminary hypothesis is that: compared with the normal turn-taking model formulated by the speaker, the total duration of turns of the CC is smaller. On

[4] Overlap means more than one speaker talk in a time. This situation will always happen in everyday conversation but it won't keep to long before one party take the turn.

[5] Chao (1922) describes the relationship between tone of words and intonation as 'small wave rides on the big wave'. That is, each Chinese character has its own tone. When it is put in a sentence, the intonation of the sentence influences the tone while it still remains the feature of the contour of the tone.

the one hand, this is due to the fact that the words at the end of CC is faster than that of the adjacent turn; on the other hand, the jointly-constructed utterances tend to be shorter, that is, have fewer words. As mentioned above, there are differences between peaceful co-constructions and competitive co-constructions. We randomly selected 30 examples of CC to measure.

The data we measured include but are not limited to: 1) total duration of conversation, 2) number of turns, 3) duration of each speaker's utterance and 4) number of words/characters of each speaker's utterance in a sequence of sub topics.

See excerpt (1):

Excerpt 1

1	R	这是高晓松说的, zheshi Gao Xiaosong shuo de It's what Gao Xiaosong[6] said
2		就说, jiushuo it is said
3		历史上纪晓岚这个人就是这样的一个人. lishi shang Ji Xiaolan zhege ren jiushi zheyang de yigeren Ji Xiaolan[7] is just this kind of person in history
4	L	**但是我觉得高晓松其实好多话都|有** danshi wo juede Gao Xiaosong qishi haoduohua dou you… but I think Gao Xiaosong's words often…
5	R	**|不能信,** buneng xin can't be trusted
6		是吧? shiba right?
7	L	对对对. duiduidui right
8		比如说他骗你的那个, biru shuo ta pianni de nage such as what he once cheated you
9	R	就, jiu just…
10		骗我的啥来着? pian wo de sha laizhe what did (you refer to by saying) he cheated me?

According to the research needs, we put the audio of the above transcribed corpus into Praat software to get a series of data.

Table 1 Parameter in excerpt（1）

Starting time	Duration（s）	Sylla-bles	Lines	Speaker
00:34:59	1.006	7	1	R
00:35:00	0.506	2	2	R
00:35:00	2.089	17	3	R
00:35:03	**2.660**	15	**4**	**L**
00:35:06	**0.524**	3	**5**	**R**
00:35:07	0.386	2	6	R
00:35:07	0.507	3	7	L
00:35:07	1.173	9	8	L
00:35:09	0.310	1	9	R
00:35:10	1.088	6	10	R

Basic parameters:

Total time: 10.249(s); Total characters: 65; turns: 10; Speakership changes: 4; Voiced time: 10.328(s)

Calculated items:

Average length: 0.158; Length of the preliminary part (line 4): 0.177s; Length of the latter part (line 5): 0.175s

By making a statistical analysis of the examples of cooperation and co-construction we have collected, we get the following table:

Table 2 Duration parameters of random CC excerpts

[6] A Chinese famous talk show host who is good at talking about the history.

[7] Ji Xiaolan is a prime minister in feudal China. He is famous for his loquacious speaking-talent and knowledgeable, which is a positive figure admittedly. However, in Gao's talk show, he is said to be a bad person.

Excerpt[8]	total time duration（s）	total characters	speed (word/s)	speed of former part	speed of latter part	relation of the two part	overlap or not	type of response in third-position	type of CC
O-06	5.438	30	0.187	0.227	0.154	＞	Y	positive	rush-though
S-11	5.639	33	0.171	0.145	0.209	＜	N	neutral	collaborative completion
AE-52	6.603	41	0.161	0.181	0.301	＜	N	positive	collaborative completion
J-03	5.759	31	0.186	0.246	0.201	＞	Y	neutral	rush-though
AA-10	11.343	73	0.155	0.143	0.192	＜	N	neutral	after-thought
B-12	3.566	24	0.149	0.225	0.147	＞	Y	positive	rush-though
A-01	23.765	144	0.165	0.182	0.178	＞	Y	positive	rush-though
……									

Base on the analysis of the data in the table, it is found out tendentious rules at least: 1) the speaking speed of the former part of CC is slower than that of the adjacent turn of the same sequence; 2) when the speaking speed of the latter part of CC is higher than that of the former one, they often overlap. These two conclusions are applicable to all kinds of CC. Moreover, it is found that CC types are partly determined by the time difference between the two parts of CC. More precise results are still being calculated.

4.2 Stress

According to our observations of the data, the collaborative completion of a CC would be produced in a mild way by speaker within the controllable range, i.e., the collaborative speaker often artificially packages his own utterances as the completion of the previous turn of the other party by using a similar pitch as the previous speaker in the sense of hearing. The prosodic operation is mainly manifested in the special design on making the following utterance sound like a continuation rather than a disjunctive one. This strategy of "loudness-matched" has been studied by some scholars (Local 1992, 2005; Szczepek 2000).

One technical problem is that, when collecting data, researchers should pay attention to the locating place of the recording equipment. The distance from the device to the two speakers should be equidistant. After verifying these, the results still show that there is no obvious difference of higher intensity or stress on the sound spectrogram. The speaker's voice intensity is always maintained at a relatively stable level. It is the pitch and duration that determine the stress in listening comprehension (Chao Yuen-Ren 1968; Lǔ Shuxiang 1979). Furthermore, stress is not the only way to highlight foreground information. In Chinese, words are often disyllabic and have their own cadence. In addition, due to the reason of the flow of speech, stress mainly relates to syntactic units rather than interactive units.

However, in the cases of less cooperative type of CC, it tends to use a recognizable stress using for preempting turns. For example, in the following example, the voice intensity of the co-constructed speaker is intentionally higher than that of the previous speaker in line 14:

[8] The numbers are original ones.

Excerpt 2

1 L 挣得太多了钱.
zhengde taiduo le qian
they earn too much money

2 他们挣得太容易了那钱.
tamen zhengde tairongyi le na qian
the way they earn money is too easy

3 他占太多了.
tamen zhan tai duo le
they occupy too much

4 演一部电视剧这么多钱.
yan yibu dianshiju zheme duo qian
playing movies can earn a lot of money

5 老百姓科学家能这么多钱啊,
laobaixing kexuejia neng zheme duo qian a
Can common people or scientists earn so much money?

6 一辈子都拿不来这么多钱来.
yibeizi dou nabulai zheme duo qian lai
they won't get so much with all their life

7 尤其是科学家,
youqi kexuejia
especially those scientists

8 上个学,
shang ge xue
highly educated

9 不如他们那不上学的都.
buru tamen na bushangxue de dou
not as good as those who haven't educated

10 R 没,
mei
without

11 没有这个文艺,
meiyou zhege wenyi
without performing art

12 不能调节老百姓的生活,
buneng tiaojie laobaixing de shenghuo
can't relax people

13 但是,
danshi
but

14 L **[不能给他抬得太高,**
buneng gei ta taide tai gao
inflate their importance so much

15 R [适可而止,
Shikeerzhi
enough is enough

16 适可而止.
Shikeerzhi
enough is enough

17 哎.
Hey

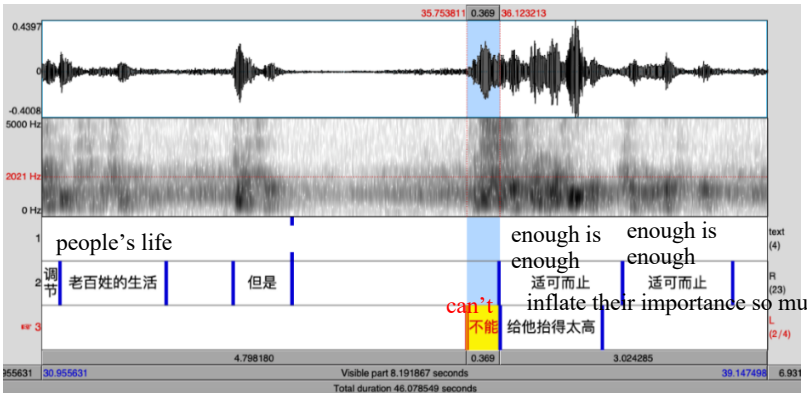


Figure 1 Sound spectrogram in line 14-15

In Excerpt 2, two retired professors are talking about the income among different people. Both of them hold a negative stance to the high-income in the entertainment industries. However, L's negative emotion is much stronger. In line 13, the adversative conjunction *danshi* "but" produced by R indicates his reservation on popular entertainment. However, the CC completion produced by L in line 14 shows that she thinks that R's speech is not fierce enough to express L's mood. Therefore, the sound intensity of L in this turn is obviously enhanced, and the voice of the negative word "no" and degree adverb "too" is significantly enhanced.

4.3 Pitch

In linguistic prosodic analysis, pitch is another important indicator of prosodic features in addition to the above-mentioned duration and loudness. As we all know, for the same individual's pronunciation, relative pitch not only distinguishes tones in words, but also distinguishes the intonation in sentences. Pitch is a core acoustic feature of the perception of stress in phonetics. Somewhat differently, in Mandarin Chinese, the main representation of word stress is duration, which is due to the fact that Chinese has taken tone (presented mainly by pitch) as an important means to distinguish meaning of words (Zhu Xiaonong, 2010). However, the pitch line in the sound spectrogram is still the most obvious evidence.

Shen et al. (1994) point out that the effect on focus stress of duration is not significant, but that of pitch is very important. Wang Yunjia et al. (2016) also point out that the initial pitch, focus pitch and final pitch of intonation are relatively constant. These studies on focus stress and intonation reflect the particular emphasis has been placed on information structure in linguistic research. But these studies are mainly involved in one speaker (monologue). This paper researches everyday conversation in its natural habitat and chooses a perspective of interaction to examine how different speakers negotiate and collaboratively achieve language tasks.

The relationship between Chinese intonations and tones has been fully discussed in the existing Chinese phonetic studies. Chao (1922, 1968) proposes theories like "rubber band effect", "algebraic sum" and "big wave and small wave" to describe this phenomenon. The outline of Chinese intonation concluded by these outstanding previous studies and experiments also has a high consistency in the data of CC, which indicates that the collaboratively constructed design of prosody acts as a part of coding and recognition of CC.

Shen (1985) comes up with the intonation structure of declarative sentence in Mandarin Chinese.

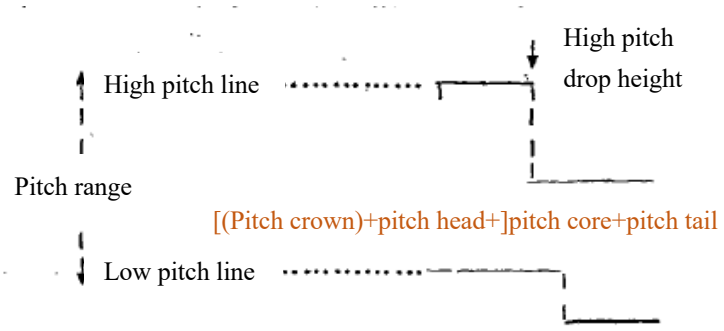


Figure 2 Intonation structure of declarative sentence in Mandarin Chinese (Shen 1985:21)

The following example of CC is a declarative sentence.

Excerpt 3

1 R 我感觉你对这个东西并没有那么深入的热爱(.)和喜欢,
wo ganjue ni dui zhege dongxi bingmeiyou name shenru de re'ai he xihuan
I feel you to this thing not really very deep love and like

2 L 所以,
Suoyi
So

3 **其实我深度热爱和喜欢的东西就是,**
qishi wo shendu re'ai he xihuan de dognxi jiushi
in fact I deep love and like thing COP
In fact, the thing I love most is,

4 R **画画儿.**
Huahuar
Drawing

5 L 嗯
En

Hm
6 (1.0)
7 R p<那你学画画儿可以啊.
na ni xue huahuar keyi a
So you can learn to draw

Speakers collaboratively construct the utterance in line 3-4. The sound spectrogram shows the pitch inline 3-5 as below.

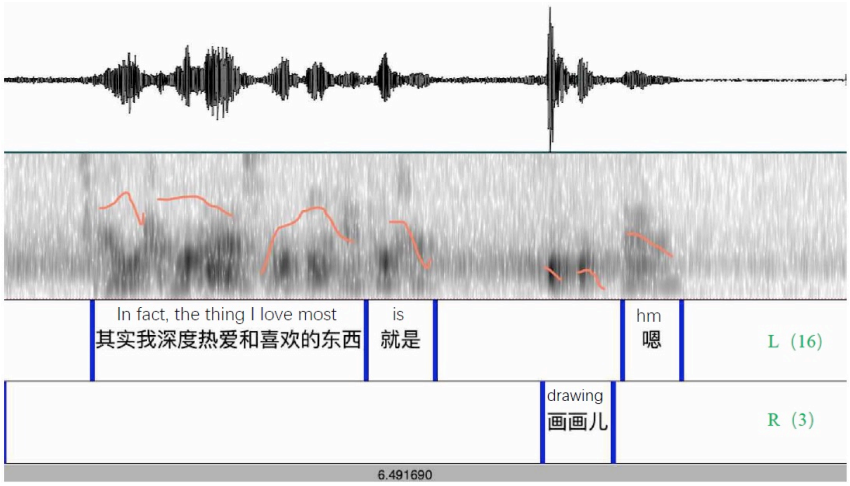


Figure 3 Sound spectrogram in line 3-5

As Figure 3 shows, although there is a gap between the collaboratively completed object produced by R from the previous turn, the intonation of the whole sentence still conforms to the pattern of Mandarin declarative sentence as shown in Figure 2. On the other hand, R's turn does not have the prosodic feature of one-member sentence, showing a phonetic design of collaborative completion.

Another function of pitch design for turn organization is 'recall'. The speaker suggests that the content of the speech is a recall of another past scene or an image of the virtual world via a significant tone change (usually using a high tone). In our data, there are such examples of using prosodic methods to simulate conversational scenes and project subsequent turns.

Excerpt 4

1 R 你还记得我之前说过么,
ni hai jide wo zhiqian shuo guo me
you still remember I before say PAT
Do you still remember what I have said before
2 L 什么?
Shenme
What?
3 R 就是我跟你说过,
jiushi wo genni shuoguo
It is that I have said to you,
4 我说我们有一男老师,
wo shuo women you yi nanlaoshi
I said there was a male teacher in our school
5 我想夸他,'您真年轻'.
wo xiang kua ta 'nin zhen nianqing'
I wanted to flatter him 'you look so young'
6 L @哈哈@
@haha@
7 噢,我[记得啊]
o wo jide a
Yeah, I remembered
8 R [我说]
wo shuo
I said:
9 我说'您看着就跟三十多岁似的'.
wo shuo nin kanzhe jiu gen sanshiduosui shide
I said 'nin look like only thirties.
10 L **['就是三十多']**
jiushi sanshiduo
'it is thirties'
11 R [然后他说,]
ranhou ta shuo
then he said
12 '我就是三十多呀'
wo jiushi sanshiduo ya
'I AM thirties.'
13 我说,我说'没有,'
wo shuo,woshuo meiyou
I said, I said 'no,'
14 我说'您看着像三十多'
wishuo 'nin kanzhe xiang sanshiduo'
I said 'you just look like thirties'

15 ‘对,我就是三十多’
‘dui wo jiushi sanshiduo’
‘Yeah, I am thirties’
16 @就是他@
jiushi ta
it is him

This excerpt is about reconstructing the past scene. R describes an awkward conversation when she first met a colleague and wanted to flatter him by underestimating his age. Unfortunately, what she estimated is actually the colleague’s real age. In line 9, she imitates the tone of "dialogic style" in prosody, which is shown as exaggerated high tone, so as to mark that the narrative behavior is a scene reappearance. The content projection is followed by the reply of the colleague. What L produces in line 10 is the responding utterance that the other person in the story answered.

Such examples show that in conversation, prosodic projection and syntactic projection occupy different channels, but they are closely related and finally work together. Moreover, one prosodic feature may not be projective in one context (speaker, conversation scene, etc.), but in another context, it becomes pretty relevant and gets highlightedly interpreted.

5 Conclusion

The main research method of this paper is to describe the characteristics of CC through Praat spectrogram. But the purpose of showing the spectrograms and parameters is not to extract numbers, but to verify what is consistent with the speaker's listening sense in communication. In fact, the sounds that the conversation participants "hear" are not same as the sounds "heard" by the machine. In addition, auditory discrimination of human ears has its own focuses, and this kind of ‘selective’ hearing has complex social base.

Prosody is a very important factor in CC. According to the data, the former speaker will have some special prosodic performance when projecting the continuing turns to be co-constructed, such as to slower speaking speed or to delay length of the syllable at the end of the turn. Also in the "reappearance" scene, the former speaker would use some special prosodic performances like the exaggerated tones to project the following turns.

6 References

Allie Hope King 2018 Collaborative completions in everyday interaction: A literature review. *Applied Linguistics and TESOL*, 18(2), 1-14.

Auer, Peter 1996 On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* 6(3), 371-381.

Auer, Peter 2005 Projection in interaction and projection in grammar. *Text* 25(1), 7-36.

Benjamin, Trevor and Traci Walker 2013 Other-initiated repair. In Carole A. Chapelle (ed.), The Encyclopedia of Applied Linguistics, Hoboken, HJ: Wiley-Blackwell.Bolden B. Galina 2003 Multiple modalities in collaborative turn sequences. *Gesture* 3(2), 187-212. Amsterdam: Benjamins.

Chao, Yuen-Ren 1968 *A Grammar of Spoken Chinese*, Berkeley: University of California Press.

Chao, Yuen-Ren 2002[1922] Zhongguo Yuyan Zidiao di Shiyan Fangfa [Experimental study of Chinese word tones]. Sciences, 7:9, 871-882.

Couper-Kuhlen, Elizabeth 2001 Interactional prosody: High onsets in reason-for-the-call turns. *Language in Society* 30:29-53.

Couper-Kuhlen, Elizabeth and Margret Selting (eds.) 1996 *Prosody in Conversation: Interactional Studies*. Cambridge: Cambridge University Press.

Couper-Kuhlen, Elizabeth and Tsuyoshi Ono (eds.) 2007 Turn continuation in cross-linguistic perspective. *Pragmatics* 17(4), Special issue.

Couper-Kuhlen, Elizabeth and Selting, M. 2011 *Studies in Interactional Linguistics,* Amsterdam: Benjamins.

Couper-Kuhlen. Elizabeth. and Selting, M. 2018 *Interactional linguistics: studying language in social interaction*. Cambridge: Cambridge University Press.

Du Bois J. W., Schuetze-Coburn, S., Cumming S. and Paolino D. 1993 Outline of discourse transcription. In Jane A. Edwards and Martin D. Lambert (eds.), *Talking Data: Transcription and Coding in Discourse Research*, 45-89. Hillsdale, NJ: Lawrence Erlbaum.

Fang, Di and Yue Guan 2019 The function of final conjunctions in Mandarin Chinese conversation. Paper presented at *the 16th International Pragmatics Conference*, 9-14June, Hong Kong Polytechnic University.

Fang, Mei 2019 Collaborative completion as speech act alignment in Chinese conversation. Paper presented at *the 16th International Pragmatics Conference*, 9-14 June, Hong Kong Polytechnic University.

Ford, Cecilia E. 1993 *Grammar in interaction*. Cambridge: Cambridge University Press.

Ford, Cecilia E. 2004 Contingency and units in interaction. *Discourse Studies* 6:27-52.

Ford, Cecilia E. and Sandra A. Thompson 1996 Interactional units in conversation: syntactic, intonational, and pragmatic resources for the projection of turn completion. In Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson (eds.), *Interaction and grammar*, 134-184. Cambridge: Cambridge University Press.

Fox, Barbara A., Sandra A. Thompson, Cecillia E. Ford and Elizabeth Couper-Kuhlen 2013 Conversation analysis and linguistics. In Jack Sidnell and Tanya Stivers (eds.), *The Handbook of Conversation Analysis*, 726 – 740. Malden: Wiley-Blackwell.

Goodwin, Charles 1981 *Conversational organization: interaction between speakers and hearers.* New York: Academic Press.

Goodwin, Charles 1986 Audience diversity, participation and interpretation. Text 6: 283-316.

Goodwin, Charles and Marjorie Harness Goodwin 1987 Concurrent operations on talk: Notes on the interactive organization of assessments. IPrA Papers in pragmatics 1:1-54.

Guan, Yue 2020 A Study of Syntactic Collaborative Construction in Mandarin Conversations. Doctoral Dissertation. The Graduate School of Chinese Academic of Social Sciences..

Guan, Yue and Mei Fang 2020 Hanyu duihua zhong dejufa hezuogongjian xianxiang chutan [A preliminary study of syntactic collaborative construction in Mandarin conversations], *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistics Studies]*, 203(3):60-69.

Hayashi, Makoto 2005 Joint turn construction through language and the body: Notes on embodiment in coordinated participation in situated activities. *Semiotica* 156:21-53.

Iwasaki, Shinako 2009 Initiating interactional turn spaces in Japanese conversation: Local projection and collaborative action. *Discourse Processes* 46: 226-246.

Lerner, Gene H. 1987 Collaborative turn sequences: entence construction and social action. Unpublished Doctoral Dissertation. University of California, Irvine.

Lerner, Gene H. 1991 On the syntax of sentences-in-progress. *Language in Society* 20, 3: 441-458.

Lerner, Gene H. 1992 Assisted storytelling: Deploying shared knowledge as a practical matter. *Qualitative Sociology,* 15, 247–271.

Lerner, Gene H. 1995 Turn design and the organization of participation in instructional activities. *Discourse Processes* 19:111-131.

Lerner, Gene H.1996. On the "semi-permeable" character of grammatical units in conversation: conditional entry into the turn space of another speaker. In Elinor Ochs, E. A. Schegloff and Sandra A. Thompson (eds.), *Interaction and grammar*, 238-276. Cambridge: Cambridge University Press.

Lerner, Gene H. 2004 On the place of linguistic resources in the organization of talk-in-interaction: Grammar as action in prompting a speaker to elaborate. *Research on Language and Social Interaction*, 37(2),151-184.

Lerner, Gene H. and Takagi, Tomoyo 1999 On the place of linguistic resources in the organization of talk-in-interaction: A co-investigation of English and Japanese grammatical practices. *Journal of Pragmatics* 31, 49-75.

Levinson, Stephen. C. 1983 *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, Stephen. C. 2013 Action formation and ascription. In Jack Sidnell and Tanya Stivers (eds.), *The handbook of conversation analysis,* 103-130. Malden: Wiley-Blackwell.

Li, Xiaoting 2014 *Multimodality, Interaction and turn-taking in Mandarin Conversation.* Amsterdam: Benjamins.

Local, John 2005 On the interactional and phonetic design of collaborative completions. In William J. Hardcastle and Janet Mackenzie Beck (eds.), *A figure of speech: A festschrift for John Laver*, 263-282. Hillside, New Jersey: Lawrence Erlbaum.

Luke, Kang Kwong, Sandra A. Thompson and Tsuyoshi Ono 2012 Turns and increments: a comparative perspective. *Discourse Processes* 49:155-162.

Lǔ Shuxiang 1979 *Hanyu Yufa Fenxi Wenti*, Beijing: Shangwu Yinshuguan.

Ochs, Elinor, Emanuel A. Schegloff and Sandra A. Thompson (eds.) 1996 *Interaction and Grammar*. Cambridge: Cambridge University Press.

Ono, Tsuyoshi and Elizabeth Couper-Kuhlen 2007 Increments in cross-linguistic perspective: Introductory remarks. *Pragmatics* 17(4):505-512.

Ono, Tsuyoshi, Sandra A Thompson and Yumi Sasaki 2012 Japanese negotiation through emerging final particles in everyday talk. *Discourse Processes* 49:243-272.

Sacks, Harvey 1992 *Lectures on Conversation, 2 Volumes* (Fall 1964-Spring 1972). Oxford: Blackwell.

Sacks Harvey, Emanuel A. Schegloff and Gail Jefferson. 1974 A simplest systematics for the organization of turn-taking for conversation. *Language*,50, 696-735.

Schegloff, Emanuel. A. 1968 Sequencing in conversation opening. *Ameracan Anthripologist* 70:1075-1095.

Schegloff, Emanuel. A. 1979 The relevance of repair to syntax-for-conversation. In Talmy Civón (ed.), *Discourse Syntax*, 261-286. New York: Academic Press.

Schegloff, Emanuel A. 1982 Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen (ed.), *Analyzing Discourse: Text and Talk*, 71 -93. Washington, D.C.: George Town University Press.

Schegloff, Emanuel A. 1996 Turn organization: one intersection of grammar and interaction. In Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson (eds.), *Interaction and Grammar,* 52 – 133. Cambridge: Cambridge University Press.

Schegloff, Emanuel A. 2007 *Sequence Organization in Interaction: A Primer in Conversation Analysis, Vol. 1*. Cambridge: Cambridge University Press.

Shen, Jiong 1985 Beijinghua shengdiao de yinyu he yudiao [The range of tones and intonation in Beijing Mandarin], In Tao Lin and Lijia Wang (ed.), *Beijing Yuyin Shiyanlu* [Enperimental Analysis on Beijing Mandarin] 73-125. Beijing: Beijing University Press.

Shen, Jiong 1992 Hanyu yudiao moxing chuyi [On the pattern of Chinese intonation]. *Yuwen Yanjiu* 4:16-24.

Shen, Jiong 2003 Language in running speech [Yunxing zhongde yuyan], *Linguistic Sciences [Yuyan Kexue]* 2:23-28.

Song, Zixuan 2019 Collaborative Construction of Turn Units in Mandarin Conversation. Master Thesis. University of Alberta.

Tao, Hongyin 1996 *Units in Mandarin Conversation: Prosody, Discourse, and Grammar.* Amsterdam: Beniamins.

ten Have, Paul 2007[1999] *Doing Conversation Analysis.* London: Sage Publications Ltd.

Thompson, Sandra A., Barbara A. Fox and Elizabeth Couper-Kuhlen 2015 *Grammar in Everyday Talk*. London: Cambridge University Press.

Walker, Gareth 2010 The phonetic constitution of a turn-hoding practice: Rush-throughs in English talk-in-interaction. In D. Barth-Weingarten, E. Reber and M. Selting (eds.), *Prosody in interaction.* Amsterdam: Benjamins. 51-72.

Walker, Gareth 2012 Coordination and interpretation of vocal and visible resources: "Trail-off" conjunctions. *Language and Speech* 55(1): 141-163.

Walker, Gareth 2013 Phonetics and prosody in conversation. In Jack Sidnell and Tanya Stivers (eds.), *The handbook of conversation analysis.* 55:141-163. Malden: Wiley-Blackwell.

Wang Yunjia, Higashi Takahiro, Dayoung Jeong 2016 Stability and variation of sentence focus and ending pitches and relevant issues [Jiaodian he jumo yingao de hengding, bianyi jiqi xiangguan wenti], *Essays on Linguistics [Yuyanxue luncong]* 2:66-99.

Wu Zongji 1982 Putonghua Yuju zhong de Shengdiao Bianhua [Tonal changes in sentences in standard Chinese]. *Zhongguo yuwen* 6:439-449.

Zhu, Xiaonong 2004 Intimacy and high pitch [Qinmi yu gaodiao], *Contemporary Linguistics [Dangdai Yuyanxue]* 6(3):193-222.

Zhu, Xiaonong 2005 Experimental phonetics and its contributions to Chinese linguistics [Shiyan yuyinxue he hanyu yuyin yanjiu], *Nankai Linguistics [Nankai Yuyanxuekan]* 1:1-17.

Zhu, Xiaonong 2010 *Yuyinxue [Phonetics]*. Beijing: Shangwu Yinshuguan.

Appendix

A. Transcription convention

,	mid-rise final pitch
.	low-fall final pitch
¿	slightly rise question intonation
?	high rise pitch
（0.5）	paused for 0.5 seconds
（.）	micro pause, often less than 0,2 seconds.
=	latching
-	sudden stop
: : :	voice lengthening, the more colons, the longer the voice
↑	pitch step up
↓	pitch step down
hhh	hearable outbreaths, the more “h”s, the longer breathing
<	inhaled sound, like p<, s, ts<, and so on
ʔ	glottal stop
> <	compressed or rushed
< >	slowed or drawn out
[ ]	overlap
（（*movement*））	non-linguistic actions
（*word*）	uncertain transcription
@……@	words with laughter
→	target line which is referred to in the text
bold	target words

Note: This transcription system basically adopts DT (Discourse Transcription) and its revised edition DT2 (Du Bois et al. 1993, 2006), with a few small modifications according to Chinese data.

B. Glossing convention

AUX	auxiliary word (de hua)
COP	copular (*shi*)
NEG	negatives (*bu*)
POSS	possessive (*de*)
PRT	particle
Q	question marker
3SG	singular

ILP-based Opinion Sentence Extraction from User Reviews for Question DB Construction

Masakatsu Hamashita* and **Takashi Inui**
University of Tsukuba
1-1-1 Tenoudai, Tsukuba, Ibaraki 305-8573, JAPAN
{m.hama@mibel.,inui@}cs.tsukuba.ac.jp

Koji Murakami and **Keiji Shinzato**
Rakuten Institute of Technology - Boston, Rakuten USA
2 South Station Suite 400, Boston, MA 02110
{koji.murakami,keiji.shinzato}@rakuten.com

Abstract

Typical systems for analyzing users' opinions from online product reviews have been researched and developed successfully. However, it is still hard to obtain sufficient user opinions when many reviews consist of short messages. This problem can be solved with an active opinion acquisition (AOA) framework that has an interactive interface and can elicit additional opinions from users. In this paper, we propose a method for automatically constructing a question database (QDB) essential for an AOA. In particular, to eliminate noisy sentences, we discuss a model for extracting opinion sentences that is formulated as a maximum coverage problem. Our proposed model has two advantages: (1) excluding redundant questions from a QDB while keeping variations of questions and (2) preferring simple sentence structures suitable for the question generation process. Our experimental results show that the proposed method achieved a precision of 0.88. We also give details on the optimal combination of model parameters.

1 Introduction

Typical systems for analyzing users' opinions from online product reviews have been researched and developed successfully (Liu, 2012; Jo and Oh, 2011; Kouloumpis et al., 2011; Pozzi et al., 2016). However, it is still hard to obtain sufficient user opinions when many reviews consist of short messages. In this situation, it would be practical to elicit additional opinions by actively asking users questions instead of just waiting for user posts. We define this procedure as an active opinion acquisition (**AOA**).

Suppose an example which is a review post consisting of just one sentence below:

u1 *This wine has a really refreshing aroma!*

It is possible to capture the user opinion "*refreshing aroma*" from **u1**. Here, in the case of an AOA-oriented system (AOAS), the system asks a question like **s1** after **u1**.

u1 *This wine had a really refreshing aroma!*

s1 *How was the aftertaste?*

u2 *The aftertaste was bitter.*

Then, it is also possible to obtain the additional opinion "*bitter aftertaste*" from **u2**. This example shows that an AOAS can efficiently collect user opinions by asking users questions.

Here, a question database (QDB), that is, a set of large quantities of question examples, is an essential resource for realizing dialogues between a user and an AOAS (Murao et al., 2003) because it would enable an AOAS to ask users precise questions in various situations. Nio and Murakami (2018) proposed a question-conversion method for constructing QDBs automatically. This method runs through a machine translation-like architecture and then converts an affirmative sentence to an interrogative form such as:

The aroma was a bouquet.
→ *How was the aroma?*

*Currently, Gunosy Inc.

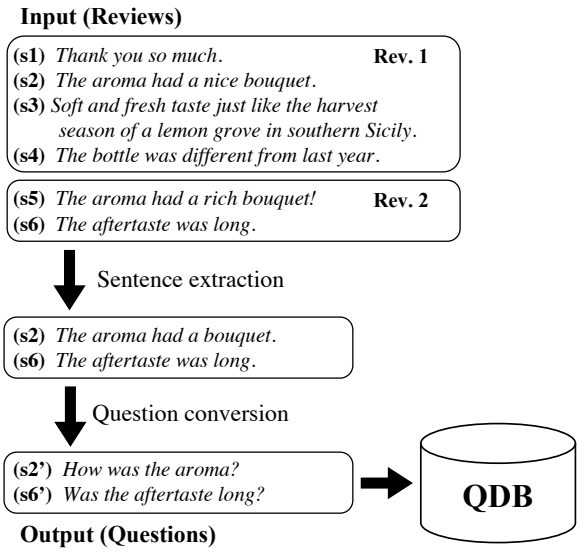


Figure 1: Relationship between sentence extraction and question conversion. Given multiple user reviews, the sentence extraction module is applied for eliminating noisy sentences and then extracted sentences are sent to the question conversion.

Note that a relationship holds that the input opinion sentence is the answer to the output question. Nio and Murakami (2018) reported a method that achieves state-of-the-art performance by using a user-review data set prepared purely for evaluation. Unfortunately, however, real review data is very noisy, so measures against such noisy data are required.

In this paper, we propose a novel sentence extraction model that eliminates noisy sentences and extracts sentences suitable for question conversion. The proposed model works as a preprocessing module for question conversion, as shown in Figure 1. Here, note that each sentence to be extracted needs to include opinion(s) like (**s2**) and (**s6**). Therefore, the proposed model is formulated as a maximum coverage problem of opinions, which makes it possible to exclude sentences including no opinions like (**s1**) and (**s4**). Naturally, the formulation also makes it possible to exclude sentences that have redundant content like (**s5**). Moreover, the basic formulation is extended to exclude sentences having sentence structures that are too complex for question conversion like (**s3**). The extended model enables us to control the number of opinions in each output sentence in order to extract opinion sentences that have simple structures. Details on the proposed model will be given in Section 3.

Through experiments done for evaluation, it is found that the proposed method achieved a precision of 0.880. Furthermore, we revealed the characteristics of the extracted opinion sentences in terms of length and the number of types of opinions. We also give details on the optimal combination of model parameters.

2 Related Work

2.1 QGSTEC

The automatic generation of questions is essential to various applications such as dialog systems and quiz generation in educational E-learning systems. The question generation shared task and evaluation challenge (QGSTEC) is a shared task for automatically generating questions for those applications. In QGSTEC, given a text segment, the goal of a system is to generate questions whose answers are included in the input segment. There have been many successful studies based on QGSTEC (Mannem et al., 2010; Ali et al., 2010; Agarwal et al., 2011). Nevertheless, our final goal is to generate questions that enable an AOAS to elicit user opinions, quite different from QGSTEC.

2.2 Neural Question Generation

Zhang et al. (2018) proposed a question generation model that uses a neural network. On a news web site, if the headline of an article is a question, the click through rate increases; thus, a question headline is generated by using an encoder-decoder model. This model requires correct answer data because it involves supervised learning. Our study differs from this study in that correct answer data is not required because our study involves unsupervised learning with only reviews and question examples are created instead of question headlines.

2.3 ILP-based Sentence Extraction

Sentence extraction has been widely studied as a form of document summarization (Kupiec et al.,

1995; Hirao et al., 2002). Among the methods of extraction proposed so far, integer linear programming (ILP) formulation provides better solutions because of its flexibility and extensibility. Given a set of sentences $D = \{s_1, \ldots, s_N\}$ as an input, ILP-based sentence extraction aims at constructing an appropriate subset $S \subseteq D$. Here, suppose D is represented by an N-dimensional 0/1 vector $\mathbf{y} = \{y_1, \ldots, y_N\}$. When a sentence s_i in D is $s_i \in S$, $\mathbf{y}$ represents the result of sentence extraction as $y_i = 1$; otherwise, $y_i = 0$.

The most fundamental model of ILP-based sentence extraction is formulated as Figure 2.

$$\begin{aligned} \mathbf{y}^* &= \arg\max_{\mathbf{y}} f(\mathbf{y}) \\ \text{s.t.} \quad & \sum_{i=1}^{N} l_i y_i \leq L_{max} \\ & \forall i, \quad y_i \in \{0, 1\} \end{aligned}$$

Figure 2: Fundamental model for ILP-based sentence extraction

Here, L_{max} represents the maximum output length, and l_i represents the length of a sentence s_i. The function $f(\mathbf{y})$ is an objective function that measures the quality of an output candidate $\mathbf{y}$. The model outputs the candidate holding a maximum value of $f(\mathbf{y})$ while satisfying all constraints.

2.4 Maximum Coverage Model

The maximum coverage model (MCM) is an instance of an ILP-based sentence extraction model, that is known to be suitable for multi-document summarization (Yih et al., 2007). MCM prefers to create a summary output that has as many varieties of *concepts*, typically words, as possible. As a result, this model is naturally able to exclude redundant concepts from the output.

Multi-document summarization based on the MCM is formulated as Figure 3. Here, the objective function $f_{mcm}(\mathbf{y})$ is defined as follows:

$$f_{mcm}(\mathbf{y}) = \lambda \sum_i r_i y_i + (1 - \lambda) \sum_k w_k z_k$$

The w_k in $f_{mcm}(\mathbf{y})$ represents the weight of the word k. The r_i represents the similarity score between a sentence s_i and entire input documents. The

$$\begin{aligned} \mathbf{y}^* &= \arg\max_{\mathbf{y}} f_{mcm}(\mathbf{y}) \\ \text{s.t.} \quad & \sum_{i=1}^{N} l_i y_i \leq L_{max} \\ & \forall k, \quad \sum_i o_{ik} y_i \geq z_k \\ & \forall i, \quad y_i \in \{0, 1\} \\ & \forall k, \quad z_k \in \{0, 1\} \end{aligned}$$

Figure 3: Maximum coverage model for multi-document summarization

z_k is a 0/1 variable that is 1 when a word k is included in an output candidate, and 0 otherwise. Also, o_{ik} in Figure 3 is a constant that becomes 1 when s_i contains k, 0 otherwise. The model guarantees consistency between y_i and z_k through the constraint $\sum_i o_{ik} y_i \geq z_k$. Nishikawa et al. (2010) proposed a variation of the MCM for multi-document opinion summarization. This model adopts an opinion as the concept e_k instead of a word to create a summary that has as many varieties of opinions as possible. The objective function $f_{nishikawa}(\mathbf{y})$ is defined as follows:

$$f_{nishikawa}(\mathbf{y}) = \lambda \sum_k w_k z_k + (1 - \lambda) \sum_{i,j} c_{i,j} x_{i,j} \quad (1)$$

The first term is the same as the second term of $f_{mcm}(\mathbf{y})$. In the second term of $f_{nishikawa}(\mathbf{y})$, $x_{i,j}$ is a decision variable that indicates the sentence order, and $c_{i,j}$ is a weight related to the naturalness of the sentence order. This makes it possible to select sentences so that important concepts are included in the summary and arrange those sentences as naturally as possible.

This is similar to our model proposed in the next section. However, its focal point is different from ours. The model of (Nishikawa et al., 2010) does not care how many opinions are included in each sentence in the output, while the proposed model controls the number of opinions in each output sentence in order to extract opinion sentences that have simple structures. The details will be given in the next section.

3 Proposed Method

In this section, we describe our novel sentence-extraction model based on the MCM formulation. Given a set of user review sentences, the model is expected to extract sentences suitable for question conversion, as mentioned in Section 1.

Suppose again that, given the six sentences shown in Figure 1 as input, only (**s2**) and (**s6**) should be extracted and sent to the question conversion process. Sentences (**s1**) and (**s4**) should not be extracted because they include no opinions at all. (**s3**) and (**s5**) are not worth extracting despite both sentences including opinions. (**s5**) is redundant because it has almost the same meaning as (**s2**) [1]. In addition, (**s3**) has too complex of a sentence structure for question conversion.

From these observations, it was found that each sentence output from the proposed model should satisfy the following requirements.

> **Requirement I:** include opinion(s),
>
> **Requirement II:** have a simple sentence structure, and
>
> **Requirement III:** exclude redundant content appearing in other output sentences.

Among these three, the first and third requirements can be achieved by applying a MCM framework, as mentioned in the previous section. In this paper, we propose an extension of the basic MCM to satisfy the second requirement. First, we propose additional constraints to control the number of opinions in each output sentence, and we then describe a novel objective function for estimating how much standard the expression of opinion is.

Figure 4 shows the formulation of the proposed model. Note that an opinion $\langle a_j, e_k \rangle$ is assigned as the *concept* in the MCM framework. Here, $a_j (\in Q_a)$ is an aspect word such as "*aftertaste*," $e_k (\in Q_e)$ is a sentiment word such as "*bitter*," and Q_a and Q_e represent a pre-defined set of aspect words and sentiment words, respectively.

Two constraints, Equation (2) and (3) in Figure 4, are added to control the number of opinions in an output sentence. A_{max} and E_{max} are constants representing the maximum number of aspect and sentiment words included in an output sentence, respectively. The function $c_a(\mathbf{y}_i, a_j)$ in Equation (2) indicates the number of sentences that contain a_j in $\mathbf{y}_i$ and is defined as follows.

$$\sum_{i=1}^{N} h_{ij} y_i$$

The h_{ij} takes 1 if a sentence s_i contains the aspect word a_j and 0 otherwise. Here, $\mathbf{y}_i$ is a vector for which the *i*-th element is the same value as that of $\mathbf{y}$, and the others are 0. As a result, $c_a(\mathbf{y}_i, a_j)$ takes 1 if s_i contains a_j and 0 otherwise, and the function $c_e(\mathbf{y}_i, e_k)$ in Equation (3) is similarly defined as $c_a(\mathbf{y}_i, a_j)$ for sentiment words. The constraint of Equation (4) has the same role as the original MCM in Figure 3. It is modified slightly from the original model due to the *concept* (opinion) structure. Here, z_{jk} is a variable that has 1 when an opinion $\langle a_j, e_k \rangle$ is included in the output and 0 otherwise.

The objective function $f_{prop}(\mathbf{y})$ for the proposed model is defined as follows.

$$f_{prop}(\mathbf{y}) = \sum_{j=1}^{|Q_a|} \sum_{k=1}^{|Q_e|} w_{jk} z_{jk} \quad (5)$$

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} f_{prop}(\mathbf{y})$$

$$\text{s.t.} \quad \sum_{i=1}^{N} l_i y_i \leq L_{max}$$

$$\forall i, \quad \sum_{j=1}^{|Q_a|} c_a(\mathbf{y}_i, a_j) \leq A_{max} \quad (2)$$

$$\forall i, \quad \sum_{k=1}^{|Q_e|} c_e(\mathbf{y}_i, e_k) \leq E_{max} \quad (3)$$

$$\forall j, k, \quad \sum_{i=1}^{N} o_{ijk} y_i \geq z_{jk} \quad (4)$$

$$\forall i, \quad y_i \in \{0, 1\}$$

$$\forall j, k, \quad z_{jk} \in \{0, 1\}$$

Figure 4: Proposed model. It enables control of number of opinions in each output sentence through additional constraints.

[1] On the contrary, (**s2**) is redundant if the model outputs (**s5**).

It forms a simple version of $f_{nishikawa}(\mathbf{y})$. The value of $f_{prop}(\mathbf{y})$ becomes larger when the output includes many different types of opinions. We use half of $f_{nishikawa}(\mathbf{y})$ because our model does not need to consider the order of sentences unlike (Nishikawa et al., 2010).

When asking a user a question, the model prefers standard expressions frequently used. From this fact, the weight w_{jk} of the variable z_{jk} is defined as:

$$w_{jk} = \frac{w_{jk}^{word}}{w_{jk}^{syn}} \tag{6}$$

Here, w_{jk}^{word} represents the co-occurrence probability of an aspect word a_j and a sentiment word e_k in an input document. w_{jk}^{syn} represents the average syntactic distance between a_j and e_k, which increases the weight of syntactically concise opinions in which aspect words and sentiment words appear close to each other. These values are calculated separately from a large review data set.

Now, we explain how to determine which pairs of aspect words and sentiment words are regarded as opinions in a sentence. Given a sentence S, V_a represents a subset of Q_a, whose elements are aspect words in S. Also, V_e represents a subset of Q_e. The opinion $\langle a_j, e_k \rangle$ is determined in S immediately when $(|V_a|, |V_e|) = (1, 1), a_j \in V_a$ and $e_k \in V_e$. However, we need to discover meaningful word pairs when several aspect words and sentiment words are included in S, such as $(|V_a|, |V_e|) = (2, 3)$. We solved this problem by performing maximum weight matching on a weighted complete bipartite graph (Korte et al., 2012), where $G(V_a \cup V_e, E)$ is a complete bipartite graph, in other words, every combination of a_j and e_k in S becomes a candidate of opinions. Each candidate $\langle a_j, e_k \rangle$ is weighted by Equation (5).

Table 1 shows examples of opinions with higher weights that were calculated by using the same data used in the experiments in Section 4.1. Similarly, Table 2 shows the case of lower weights. One can see that plausible opinions are included in Table 1 while meaningless aspect/sentiment word pairs are included in Table 2.

Table 1: Examples of high weight opinions

⟨*balance, good*⟩
⟨*taste, long*⟩
⟨*taste, rich*⟩
⟨*aroma, spread*⟩
⟨*cost-performance, excellent*⟩

Table 2: Examples of low weight opinions

⟨*cork, strong*⟩
⟨*taste, hero*⟩
⟨*label, soft*⟩
⟨*bottle, long*⟩
⟨*price, beautiful*⟩

4 Experiments

4.1 Experimental Settings

The following two experiments were conducted.

Experiment I We conducted a series of experiments where combinations of model parameters (A_{max} and E_{max}) were changed to investigate the relationship between the performance and the parameters of the proposed model. Hereafter, we refer to the proposed model as **ILP+C**$_{(A_{max},E_{max})}$ when showing the parameters of the model clearly.

Experiment II We compared a simple version of the ILP-based sentence extraction model, namely **ILP-only**, with a non-ILP-based method to verify the effectiveness of ILP-based formulation. **ILP-only** is equivalent to the proposed model without the additional constraints [Equations (2) and (3)]. Additionally, the proposed model is compared with ILP-only to evaluate the effectiveness of the additional constraints.

We used a set of Japanese user review sentences posted on Rakuten Japan[2], which is one of the major E-commerce web sites in Japan. First, we crawled the sentences in the wine category and randomly selected 1,000 sentences from 19,160 sentences. Then, two annotators independently judged

[2]https://www.rakuten.co.jp

Table 3: Data set for evaluation

#Sentences(Positive/Negative)	715(367/348)
$\overline{\text{aspect}}$	1.97
$\overline{\text{sentiment}}$	1.61
$\overline{\text{length}}$	53.6

whether sentences satisfied the requirements shown in Section 3. Details on the data set are given in Table 3. Here, the symbol "Positive" indicates that a sentence can be converted into relevant questions, that is, it should be extracted, and "Negative" the opposite. $\overline{\text{Aspect}}$ and $\overline{\text{sentiment}}$ indicate the average number of aspect lexicons and sentiment lexicons per sentence, respectively, and $\overline{\text{length}}$ indicates the average number of characters per sentence. Cohen's Kappa, which means the degree of inter-annotator agreement, was 0.765 (Cohen, 1960).

We handcrafted a set of aspect lexicons Q_a and a set of sentiment lexicons Q_e by collecting opinions that appeared in the data set for evaluation because no Japanese aspect/sentiment lexicons suitable for our data set exist. As a result, we determined that $|Q_a| = 81$ and $|Q_e| = 835$. Here, we collected only sentiment lexicons with a positive polarity according to the findings of (Hamashita et al., 2018); it is suitable that questions used in an AOAS include contents with positive polarity.

In Experiment I, A_{max} and E_{max} in the proposed model are changed from 1 to 5, respectively. The non-ILP-based method used in Experiment II is a weight-based method that extracts sentences with higher weights until the total size of the extracted opinion sentences is over L_{max}. The weight of sentence s_i is calculated by summing up the weights of the opinions w_{jk} defined in Equation (5), included in s_i. We refer to this method as **w/oILP** hereafter. For each run of all experiments, the ILP solution was obtained by using Python's PuLP library (Mitchell et al., 2011), and L_{max} was set to hold a summarization rate of 5%.

We used a precision measure of the extractions, the average length of the extracted sentences (|Sentence|), the number of extracted sentences (#Sentences), and the number of types of opinions included in the extracted sentences (#Opinions) as the evaluation measures.

Table 4: Precision value for each (A_{max}, E_{max})

		E_{max}				
		1	2	3	4	5
A_{max}	1	.666	.701	.735	.735	.735
	2	.821	.810	.794	.774	.782
	3	.864	.826	.794	.833	.819
	4	**.880**	.810	.782	.794	.791
	5	.868	.794	.785	.794	.797

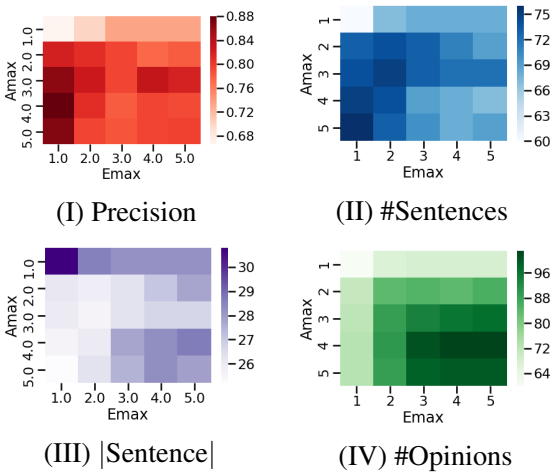


Figure 5: Heat map representations corresponding to results for each evaluation measure. For each map, as metric value becomes larger, cell becomes darker.

4.2 Results

First, Table 4 and Figure 5 show the results of Experiment I. Here, Figure 5 represents heat maps corresponding to the results for each evaluation measure, where the vertical axis indicates A_{max}, and the horizontal axis indicates E_{max}. For each map, the larger the metric value becomes, the darker the color of a cell is.

Table 4 shows the values of the precision measure. It turns out that the precision tended to be large when $E_{max} = 1$. Notably, the best result of 0.880 was achieved for $(A_{max}, E_{max}) = (4, 1)$. We found that almost all opinion sentences extracted by ILP+C$_{(4,1)}$ kept a simple sentence structure. Examples of the extracted sentences are shown in Figure 6(A).

In comparison between Figure 5(I) and

Table 5: Results of Experiments Ⅱ

	w/oILP	ILP-only
Precision	.621	**.803**
|Sentence|	66.2	29.1
#Sentences	29	66
#Opinions	47	102

Table 6: Correlation coefficients

	correlation coefficient
#Sentences	0.85
#Opinions	0.33
$f_{prop}(\mathbf{y}^*)$	0.42

Figure 5(II), precision and #Sentences show similar results. That is to say, both metric values became larger when $E_{max} = 1$.

Figure 5(III) holds the reversed proportion against Figure 5(II). The reason could be that the value of #Sentences multiplied by that of |Sentence| tends to remain constant due to the constraint of L_{max}. Next, it was found in Figure 5(III) that the sentences extracted by ILP+C$_{(1,1)}$ had a large |Sentence| and also found in Figure 5(II) that the precision rapidly decreased when $(A_{max}, E_{max}) = (1, 1)$. Now, we discuss why the precision decreased. Some of the correct and wrong examples extracted by ILP+C$_{(1,1)}$ are shown in Figure 6(B) and Figure 6(C), respectively. From Figure 6(B), we can see that the correct examples had short lengths and simple structures similar to those of ILP+C$_{(4,1)}$, while the wrong examples in Figure 6(C) tended to be long due to their containing useless words. We also observed that the sentences shown in Figure 6(C) were not extracted when A_{max} increased. From the results, it is expected that inappropriate (long) sentences would be over-extracted due to there being a lack of sentences that satisfy the constraints when $(A_{max}, E_{max}) = (1, 1)$.

As shown in Equation (5), the objective function tended to return a larger value when there were a variety of opinions in the output sentences. This relationship immediately lead to the phenomenon that the larger both A_{max} and E_{max} became, the larger #Opinions became. This corresponds to the results shown in Figure 5(IV). Here, we note the results for $(A_{max}, E_{max}) = (5, 5)$. In this case, the precision (0.797) was lower than the best of 0.880 from Table 4. The reason could be that ILP+C$_{(5,5)}$ attempts to extract sentences that include multiple opinions in order to include as many opinions as possible in the output as shown in Figure 6(D).

Next, Table 5 shows the results of Experiment Ⅱ. From the table, we found that (1) ILP-only achieved better precision than w/oILP and that (2) the output obtained by ILP-only included a lot of short sentences with varieties of opinions. Therefore, the ILP-based model was verified to be appropriate for our purpose. The precision of ILP-only was 0.803, confirming that the proposed method had a better extraction precision. ILP-only is an extreme case of the proposed model and strictly equivalent to ILP+C$_{(\infty,\infty)}$. Therefore, ILP-only is considered to be a model similar to ILP+C$_{(5,5)}$. Looking at Table 4 and Table 5, it can be confirmed that the precision of ILP-only and ILP+C$_{(5,5)}$ were similar.

Finally, we discuss how to estimate (A_{max}, E_{max}), which maximizes the precision without seeing it. We mentioned above that the metric #Sentences varies the same as precision. In addition to the findings, we investigated the correlation coefficients between precision and other metrics of each (A_{max}, E_{max}) to find a suitable metric that estimates (A_{max}, E_{max}). The results are shown in Table 6. Since #Sentences and |Sentence| are approximately inversely proportioned, the correlation coefficient with |Sentence| is not included in the table. The function $f_{prop}(\mathbf{y}^*)$ was added to the target metric for the investigation. As a consequence, the correlation coefficient between #Sentences and precision was the largest, while the other correlation coefficients were low. From these results, we can conclude that one can select (A_{max}, E_{max}) with the largest #Sentences. We get $(A_{max}, E_{max}) = (5, 1)$ in the case of our experimental settings if we adopt this strategy. The precision is not optimal but is the second largest when $(A_{max}, E_{max}) = (5, 1)$; thus, we consider that (A_{max}, E_{max}) can be estimated almost exactly by referring to #Sentences.

5 Conclusion

We proposed a novel model for extracting opinion sentences for constructing question DBs. The pro-

(A): sentences extracted by ILP+C$_{(4,1)}$

[c] $_3$果実 $_2$味の $_1$バランスが $_1$素晴らしい． / *The $_1$balance of $_3$fruity $_2$taste is $_1$excellent.*
[c] $_1$インパクトの $_1$強い $_2$味です． / *The $_1$impression of the $_2$taste was $_1$strong.*
[c] $_2$タンニンも $_1$まろやかな $_1$味わいです． / *The $_2$tannin was a $_1$mild $_1$taste.*

(B): sentences extracted by ILP+C$_{(1,1)}$

[c] とても $_1$果実が $_1$豊か． / *It was $_1$very $_1$fruity.*
[c] すべての要素の $_1$バランスが $_1$絶妙です. / *The $_1$balance of all elements was $_1$exquisite.*

(C): sentences extracted by ILP+C$_{(1,1)}$

[w] $_1$ラベルのロゴも $_1$爽やかなライトブルーをあしらってなかなかクールなイメージ. / *The logo of the $_1$bottle label was $_1$refreshing light blue and so cool.*
[w] $_1$程よい$_1$酸味を感じながらスッキリとお飲みいただくことができます. / *You can drink refreshingly with $_1$moderate $_1$acidity.*

(D): sentences extracted by ILP+C$_{(5,5)}$

[c] $_1$柔らかな $_1$タンニンが大きく $_2$広がる $_2$味わい. / *The $_2$taste of $_1$soft $_1$tannin was $_2$widespread.*
[c] $_1$酸味が $_1$甘さを抑えた $_2$バランスの $_2$良い $_3$軽快な $_3$味わい. *The $_3$light $_3$flavor with $_2$best $_2$balance between $_1$acidity and $_1$sweetness.*
[w] $_1$香りは $_1$フルーティーな印象ながら、 $_4$穏やかな泡と $_2$すっきりとした $_2$キレのある $_3$味わいは、$_3$しっかりと冷やして お料理と合わせるのがおすすめです. / *This $_3$cool wine matches your special dinner because it has $_1$fruity $_1$aroma, $_2$clear and $_2$sharp $_3$taste, and $_4$mild foam.*
[w] $_5$タンニンと $_4$酸味の $_1$バランスが $_1$よく、$_3$フルーティーで、$_2$口当たりの $_2$優しいワインが多い. / *A lot of wines which have a good balance of $_5$tannin and $_4$acidity, $_3$fruity, and $_2$pleasant $_2$taste.*

Figure 6: Examples of original Japanese sentences and their literal translations into English. The symbols **[c]** and **[w]** indicate correct and wrong extractions, respectively. Underline parts indicate aspect words, and double underline parts indicate sentiment words. A pair of aspect and sentiment words with the same arabic number means an opinion.

posed model was formulated as a maximum coverage problem of opinions. Our model has additional constraints to control the number of opinions in each output sentence and also has an objective function in order to extract opinion sentences that have simple structures. From the experimental results, we found that ILP+C$_{(4,1)}$ achieved a precision of 0.88. We also found that one can achieve promising results when selecting (A_{max}, E_{max}) with the largest #Sentences.

For future work, it is necessary to improve an opinion detection method suitable for our Japanese data set. While we applied a simple dictionary-based detection method in this work, more sophisticated methods (Brody and Elhadad, 2010; He et al., 2018) could be combined with our model. We also plan to develop an AOAS with a QDB constructed with the proposed model and conduct comprehensive evaluations.

References

[Agarwal et al.2011] Manish Agarwal, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Ed-*

ucational Applications, pages 1–9.

[Ali et al.2010] Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 58–67.

[Brody and Elhadad2010] Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812.

[Cohen1960] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Hamashita et al.2018] Masakatsu Hamashita, Takashi Inui, Koji Murakami, and Keiji Shinzato. 2018. Insertion effect of negative affix in question generation for interactive information collection system. (in Japanese). In *The Association for Natural Language Processing, 25(25)*.

[He et al.2018] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1121–1131.

[Hirao et al.2002] Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002. Extracting important sentences with support vector machines. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7.

[Jo and Oh2011] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824.

[Korte et al.2012] Bernhard Korte, Jens Vygen, B Korte, and J Vygen. 2012. *Combinatorial optimization*, volume 2. Springer.

[Kouloumpis et al.2011] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*.

[Kupiec et al.1995] Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*, pages 68–73.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Mannem et al.2010] Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at upenn: QGSTEC system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91.

[Mitchell et al.2011] Stuart Mitchell, Michael OSullivan, and Iain Dunning. 2011. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand, http://www. optimization-online. org/DB_FILE/2011/09/3178. pdf*.

[Murao et al.2003] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, and Yasuyoshi Inagaki. 2003. Example-based spoken dialogue system using woz system log. In *Proceedings of the Fourth SIGdial Workshop of discourse and dialogue*.

[Nio and Murakami2018] Lasguido Nio and Koji Murakami. 2018. Intelligence is asking the right question: A study on Japanese question generation. In *IEEE Spoken Language Technology conference*.

[Nishikawa et al.2010] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 910–918.

[Pozzi et al.2016] F. Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment analysis in social networks*. Morgan Kaufmann.

[Yih et al.2007] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 7, pages 1776–1782.

[Zhang et al.2018] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 617–626.

Composing Word Vectors for Japanese Compound Words Using Bilingual Word Embeddings

Teruo Hirabayashi Kanako Komiya Masayuki Asahara Hiroyuki Shinnou

Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
{20nd303t, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo, Japan
masayu-a@ninjal.ac.jp

Abstract

This study conducted an experiment to compare the word embeddings of a compound word and a word in Japanese on the same vector space using bilingual word embeddings. Because Japanese does not have word delimiters between words; thus various word definitions exist according to dictionaries and corpora. We divided one corpus into words on the basis of two definitions, namely, shorter and ordinary words and longer compound words, and regarded two word-sequences as a parallel corpus of different languages. We then generated word embeddings from the corpora of these languages and mapped the vectors into the common space using monolingual mapping methods, a linear transformation matrix, and VecMap. We evaluated our methods by synonym ranking using a thesaurus. Furthermore, we conducted experiments of two comparative methods: (1) a method where the compound words were divided into words and the word embeddings were averaged and (2) a method where the word embeddings of the latter words are regarded as those of the compound words. The VecMap results with the supervised option outperformed that with the identical option, linear transformation matrix, and the latter word method, but could not beat the average method.

1 Introduction

Japanese words have many definitions because Japanese does not have word delimiters between words, and word boundaries are unspecific. Therefore, the Japanese dictionary defines words individually. Japanese has different word definitions according to each corpus and dictionary. The long unit for compound words and the short unit for words in UniDic[1] (Maekawa et al., 2010) developed by the National Institute for Japanese Language and Linguistics (NINJAL) are some of them. For example, “いちご狩り, ichigo-gari, strawberry picking” is defined as one word (short unit), whereas “ぶどう狩り, budou-gari, grape picking” is defined as a compound word (long unit) with two words (short unit)[2] in UniDic. Due to the limit of the dictionary’s coverage, a morphological analyzer using UniDic treats “いちご狩り, ichigo-gari, strawberry picking” as one word and “ぶどう狩り, budou-gari, grape picking” as two words, making it impossible to directly compare the word meanings of these two words via word embeddings.

Therefore, to address the word unit discrepancy issue, this study proposes the usage of bilingual word embeddings (BWEs), which is usually used for mapping the word embeddings of two different languages into the same vector space, to map the word embeddings of long and short units into a common vector space. Using the BWE makes it easy to compare the word embeddings of “いちご狩り, ichigo-gari, strawberry picking” and “ぶどう狩り, budou-gari, grape picking” because both are on the same vector space. This situation is more convenient for many application systems like an information retrieval system.

[1] https://unidic.ninjal.ac.jp/ (In Japanese)

[2] いちご means strawberries; ぶどう means grapes; and 狩り means picking or hunting in Japanese.

2 Related Work

According to a survey of cross-lingual word embedding models[3], the BWE is classified into four groups according to how cross-lingual word embeddings are made.

The first approach is monolingual mapping. This approach initially trains monolingual word embeddings and learns a transformation matrix that maps representations in one language to those of the other language. Mikolov et al. (2013) showed that vector spaces can encode meaningful relations between words and that the geometric relations that hold between words are similar across languages. They did not assume the use of specific language; thus their method can be used to extend and refine dictionaries for any language pairs.

The second approach is pseudo-cross-lingual. This approach creates a pseudo-cross-lingual corpus by mixing contexts of different languages. Xiao and Guo (2014) proposed the first pseudo-cross-lingual method that utilized translation pairs. They first translated all words that appeared in the source language corpus into the target language using Wiktionary. They then filtered out the noises of these pairs and trained the model with this corpus, in which the pairs were replaced with placeholders to ensure that the translations of the same word have the same vector representation.

The third approach is cross-lingual training. This approach trains their embeddings on a parallel corpus and optimizes a cross-lingual constraint between the embeddings of different languages that encourages embeddings of similar words to be close to each other in a shared vector space. Hermann and Blunsom (2014) trained two models to output sentence embeddings for input sentences in two different languages. They retrained these models with sentence embeddings using a least squares method.

The final approach is joint optimization, which not only considers a cross-lingual constraint but also jointly optimizes monolingual and cross-lingual objectives. Klementiev et al. (2012) performed the first research using joint optimization. Zou et al. (2013) used a matrix factorization approach to learn cross-lingual word representations for English and Chinese and utilized the representations for a machine translation task. In this study, we used the first approach, monolingual mapping.

The nearest works to this research are those of Komiya et al. (2019) and Kouno and Komiya (2020). Komiya et al. (2019) composed word embeddings for long units from the two word embeddings of short units using a feed-forward neural network system. They classified the dependency relations of two short units into 13 groups and trained a composition model for each dependency relation. Meanwhile, Kouno and Komiya (2020) performed the multitask learning of the composition of word embeddings and the classification of dependency relations.

We utilized the BWE herein for the same purpose. To the best of our knowledge, our study is the first to use the BWE to map the word embeddings of different word delimitation definitions.

3 Methods

The BWE is usually used for cross-lingual applications (e.g., machine translation).

In this study, we mapped the word embeddings of short and long units into the common vector space for a comparison. short units are language units defined from the perspective of morphology (Ogura et al., 2007), whereas long units are those defined based on a Japanese base phrase unit, bunsetsu (Fujiike et al., 2008). A long unit consists of one or more short units. For the BWE, we utilized the linear transformation matrix and the VecMap [4].

3.1 Bilingual Word Embeddings

We used monolingual mapping comprising two steps. First, monolingual word embeddings were trained for each language. We regarded the corpora of different term units as the corpora of two different languages and mapped them to a common vector space such that the word embeddings of the words whose meanings were similar to each other in two languages can be brought closer. The geometrical relations that hold between words are similar across languages; thus a vector space of a language can be transformed into that of another language using a linear projection. We adapted hereikn two methods of the BWE, namely, linear transformation matrix and VecMap. A linear projection matrix W was

[3] https://ruder.io/cross-lingual-embeddings/

[4] https://github.com/artetxem/vecmap#publications

learned when we used a linear transformation matrix. VecMap is an implementation of a framework of Artetxe et al. (2017) to learn cross-lingual word embedding mappings (Artetxe et al., 2018a)(Artetxe et al., 2018b).

3.1.1 Linear Transformation Matrix

We conducted the following experiments when a linear transformation matrix was learned:

1. Generate short and long unit corpora and learn short or long unit embeddings for each corpus from them using word2vec (cf. Figure 1).

2. Learn a linear projection matrix W from the vector space of the short units to that of the long units using pairs of embeddings for common words generated in the last step.

3. Apply matrix W to the short unit embeddings and obtain the projected long unit embeddings for them.

3.1.2 VecMap

VecMap was used as another method of the BWE. We projected the vector space of the short units into that of the long units when we used the linear transformation matrix. However, VecMap projected both the vector spaces of the short and long units into a new vector space. The two options (i.e., supervised and identical) were compared. The supervised VecMap uses the specified words, whereas the identical VecMap uses identical words in two languages as the projection seeds. Therefore, the seed words of the supervised VecMap were the same as the linear transformation matrix but those of the identical VecMap were different.

4 Experiments

We used NWJC2vec (Shinnou et al., 2017) for the word embeddings of the short units and the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa et al., 2014) for the word embeddings of the long units using word2vec.

4.1 Word Embeddings

NWJC2vec is a set of word embeddings generated from the 25 billion word scale NWJC-2014-4Q dataset (Asahara et al., 2014), which is an enormous Japanese corpus, NINJAL Web Japanese Corpus (NWJC), developed using the word2vec tool. The summary statistics for the NWJC-2014-4Q data and the parameters used to generate the word embeddings are respectively presented in Tables 1 and 2. We used continuous bag-of-words (CBOW) as a model architecture to produce the word embeddings.

BCCWJ is the 100 million word scale balanced corpus that contains texts from multiple domains constructed by NINJAL. Each text in this corpus has short ande long unit versions. The summary statistics for BCCWJ are listed in Table 3. The word2vec settings for training the word embeddings with BCCWJ are summarized in Table 4.

NWJC2vec contains morphological information, but the word embeddings generated for the long units using BCCWJ do not contain them. Therefore, the word embeddings for the short units can be differentiated from the words with the same spellings but are different parts of speech, whereas those for the long units cannot. Consequently, for some words, the word embeddings for the short units of some words had multiple vectors, but we still directly used them.

4.2 Bilingual Word Embeddings

The learning parameters of the linear transformation matrix are shown in Table 5. We used a 200-by-200 dimensional linear transformation matrix. We used Adam as the optimizer of loss function and iterated the training for 1,164 epochs. We decided on the number of epochs according to the preliminary experiments using 55,630 words randomly extracted from the training data. We averaged the best number of five trials. The vocabulary size of the word embeddings for BCCWJ and NWJC and the seed words we used for the linear transformation matrix is shown in Table 6.

We used the default settings for the VecMap tool for each option. The default settings of the parameters of each specific option and their general default settings are listed in Table 7. The vocabulary size of the word embeddings for BCCWJ and NWJC and the seed words used for VecMap is presented in Table 8.

The number of long units decreased for VecMap compared with the linear transformation matrix be-

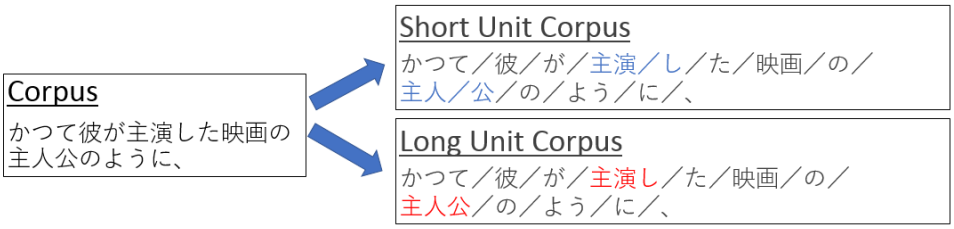


Figure 1: Short and long unit corpora

Number of URLs collected	83,992,556
Number of sentence	1,463,142,939
Number of words (tokens)	25,836,947,421

Table 1: Summary statistics for the NWJC-2014-4Q dataset

Parameters	Options	Settings
CBOW or skip-gram	-cbow	1
Dimensionality	-size	200
Window size	-window	8
Number of negative samples	-negative	25
Hierarchical softmax	-hs	0
Minimum sample threshold	-sample	1e-4
Number of iterations	-iter	15

Table 2: Parameters used to generate NWJC2vec

Number of text samples	172,675
Number of short units (tokens)	104,911,464
Number of long units (tokens)	83,585,665

Table 3: Summary statistics for the Balanced Corpus of Contemporary Japanese (BCCWJ)

Parameters	Options	Settings
CBOW or skip-gram	-cbow	1
Dimensionality	-size	200
Window size	-window	5
Number of iterations	-iter	5
Batch size	-batch$_{words}$	1,000
Minimum count of words	-min-count	1

Table 4: Settings of word2vec

Parameters	Settings
Dimensionality	200×200
Optimization algorithm	Adam
Number of epochs	1,164

Table 5: Learning parameters of the linear transformation matrix

cause of the limitation of the machine power. We used 278,143 seed words and 11,662 compound words annotated with a concept number for the evaluation, which resulted to a total of 289,805 words.

5 Evaluation

We evaluated our methods by the ranking of synonyms using a thesaurus. Using a thesaurus, we can evaluate the similarity of concepts referring knowledge of people. However, if we directly use cosine similarity between concepts, the thresholds are difficult to decide. Therefore, we used the ranking among the nodes of the thesaurus. We used "Word List by Semantic Principles" (WLSP) (National Institute for Japanese Language and Linguistics, 1964) [5] as a thesaurus. The WLSP is a Japanese thesaurus that classifies and orders a word according to its meaning. One record is composed of the following elements: record ID number, lemma number, type of record, class, division, section, article, concept number, paragraph number, small paragraph number, word number, lemma with explanatory note, lemma without explanatory note, reading and reverse reading. The concept number consists of a category, a medium item, and a classification item. The tree structure of the WLSP is shown in Figure 2.

The WLSP has a tree structure; thus, we assumed that the concepts belonging to the same node or synonyms were similar to each other.

[5] https://pj.ninjal.ac.jp/corpus_center/goihyo.html

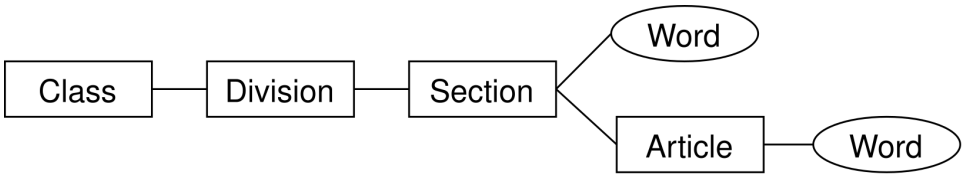


Figure 2: Tree structure of the Word List by Semantic Principles (WLSP)

Corpus	Vocabulary size (Number of word tokens)
BCCWJ (long unit)	2,745,657
NWJC2vec (short unit)	1,534,957
Seed words	278,143

Table 6: Vocabulary size (number of word tokens) of the word embeddings for BCCWJ and NWJC and seed words for the linear transformation matrix

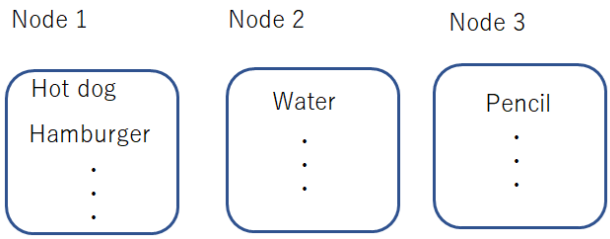


Figure 3: Example of the nodes of the WLSP

An example of the WLSP nodes is presented in Figure 3. In this figure, we assumed that *hot dog* was closer to *hamburger* than *water* or *pencil*. We used *hot dog* instead of long term like "葡萄狩り, grape picking" and *hamburger* instead of short term like "いちご狩り, strawberry picking" for example. We used *water* and *pencil* as short terms in this example.

We evaluated the mapped word embeddings on the basis of this assumption and subsequently defined "long term" and "short term." A compound word that is a long unit and consists of two short units is referred to as "long term," whereas a word that is a short unit with no long unit is referred to as "short term."

5.1 Evaluation Procedure

All the NWJC or BCCWJ words were not listed on the WLSP; thus we had two compound word conditions for evaluation: (1) the compound word should be a long term listed on the WLSP, and (2) its constituents of it should be short terms listed on the WLSP. Hereinafter, wl_i denotes the compound word, and ws_{i_1} and ws_{i_2} denote the constituents. The evaluation procedures are as follows:

1. For each long term wl_i, identify a node $N_i(0)$ to which the long term belongs in WLSP.

   $N_i(0)$ includes synonyms of wl_i and both long and short terms. We assumed that every node has at least two words such that the similarity between them can be calculated. For example, if wl_i is the word *hot dog*, the corresponding node $N_i(0)$ includes synonyms such as *hamburger*. In Figure 3, $N_i(0)$ is Node 1.

2. Calculate $s_i(0)$, which is the average similarity between the word embeddings of wl_i and all the short terms in $N_i(0)$, using the mapped word embeddings.

   For this step, we calculated $s_i(0)$, which is the average similarity between the word embeddings of *hot dog* and those of *hamburger* and other concepts in $N_i(0)$ (Node 1). We used the cosine similarity for the similarity and the arithmetic mean to average the similarity.

3. Obtain sibling nodes $N_i(1)...N_i(n)$ of $N_i(0)$.

   A sibling nodes $N_i(1)...N_i(n)$ include a node that contains a word, such as *water*, and another node that contains a word such as *pencil*. In Figure 3, $N_i(1)...N_i(n)$ includes Nodes 2 and 3.

4. Similarly, calculate $s_i(k)$, which is the average similarity between the word embeddings of wl_i and those of all the short terms in node $N_i(k)$

5. Obtain the ranking of $s_i(0)$ in $s_i(0)...s_i(n)$.

We used 11,459 long terms for the evaluation because 11,662 long terms and their constituent short

Option	Parameter	Default setting of specific option	General sefault setting
Supervised	Batch size	1,000	10,000
Identical	Self-learning	TRUE	FALSE
Identical	Vocabulary_cutoff	200,000	0
Identical	csls_neibourhood	10	0

Table 7: Parameters of VecMap

Corpus	Vocabulary size
BCCWJ (long unit)	289,805
NWJC2vec (short unit)	1,534,957
Seed words	278,143

Table 8: Vocabulary size of word embeddings for BCCWJ and NWJC and seed words for VecMap

terms were annotated with a concept number, but 203 of them had un-annotated synonyms in the node to which the word belongs ($N_i(0)$). The number of nodes we used was 881 after excluding 14 nodes that included a word with no word embeddings.

We performed two comparative methods, namely, average and latter word methods. For the average method, the word embeddings of a long term were calculated as the average of its constituent short terms, that is, the average of the word embeddings of ws_{i_1} and ws_{i_2} was used. For the latter word method, the word embeddings of the latter short term were regarded as the word embeddings of the long term, that is, the word embeddings of ws_{i_2} were used.

5.2 Results and Discussion

The average rankings of the correct node according to each method are shown in Table 9.

Method	Ranking
Linear transformation matrix	187.50
VecMap (supervised)	131.98
VecMap (identical)	330.40
Average	80.41
Latter word	143.16

Table 9: Average rankings of the correct node according to method

Table 9 shows that the best method among the three proposed methods is VecMap with the supervised option. The ranking of the correct node when the method was used was 131.98th. The number of nodes we used was 881; thus, if the node is randomly selected, the ranking would be 440th or 441st. Therefore, VecMap outperformed the random baseline and the latter word method (Table 9). However, the average method known as the strong comparative method was the best among all the methods tested. BWEs could not beat it. This result indicates that the additive compositionality holds for many long units. For future work, Skipgram can be tried instead of CBOW algorithm. Also, other word embeddings such as Glove could be another option. Theoretically, we believe that our methods can be applied even if the dimensionalities of two embeddings are different,but should be tested to know the real results.

6 Conclusion

In this study, we mapped word the embeddings of a compound word and word in Japanese into the same vector space using the BWE. We used the linear transformation matrix and VecMap as the BWE methods. VecMap with the supervised option outperformed one baseline, which was the method where the word embeddings of the latter constituent word are regarded as the word embeddings of the compound word but could not beat another baseline, which was the method where the average of the word embeddings of the constituents was used for the word embeddings of the compound word.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Number 18K11421, 17KK0002, and a project of the Younger Researchers Grants from Ibaraki University.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. *Alexandria*, 25(1-2):129–148.

Yumi Fujiike, Hideki Ogura, Toshinobu Ogiso, Hanae Koiso, Kiyotaka Uchimoto, Satsuki Soma, and Takenori Nakamura. 2008. Short-term unit alanysis of balanced corpus of contemporary japanese. In *Proceedings of the NLP2008, (In Japanese)*, pages 931–934.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

Kanako Komiya, Takumi Seitou, Minoru Sasaki, and Hiroyuki Shinnou. 2019. Composing word vectors for japanese compound words using dependency relations. *CICLING*. no 229.

Shinji Kouno and Kanako Komiya. 2020. Composition of word representation of long-term units from word representations of short-term units using multitask learning. In *Proceedings of the NLP2020, (In Japanese)*, pages 209–212.

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.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

National Institute for Japanese Language and Linguistics. 1964. *Word List by Semantic Principles*. Shuuei Shuppan, In Japanese.

Hideki Ogura, Toshinobu Ogiso, Hanae Koiso, Yumi Fujiike, and Satsuki Soma. 2007. Short-term unit alanysis of balanced corpus of contemporary japanese. In *Proceedings of the NLP2007, (In Japanese)*, pages 720–723.

Hiroyuki Shinnou, Masayuki Asahara, Kanako Komiya, and Minoru Sasaki. 2017. nwjc2vec: Nwjc2vec: Word embedding data constructed from ninjal web japanese corpus. *Journal of Natural Language Processing (In Japanese)*, 24(5):705–720.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

Exploring Discourse on Same-sex Marriage in Taiwan: A Case Study of Near-Synonym of HOMOSEXUAL in Opposing Stances

Han-Tang Hung
Graduate Institute of Linguistics
National Taiwan University
r07142007@ntu.edu.com

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

This research explores the intense conflict of the legalization of same-sex marriage in Taiwan by studying how the near-synonyms denoting homosexual, 同志 *tóngzhì* and 同性戀 *tóngxìngliàn*, are used by the opposing stances. Two research questions related to lexical semantics are addressed, i.e., what the semantic difference of these two lexical items is and how these two are characteristically used by each stance. Collocational analysis with self-compiled corpora is the primary method in this study. For the first question, it is found that the meanings of the pair can be differentiated along an internal-external axis. With regards to the second one, it is discovered that, the opponents of same-sex marriage are inclined to externalize the innateness of homosexual, whereas the supporters tend to reduce the distinctiveness of homosexuality and call for the universality of human rights by the strategies of juxtaposing and compounding.

1 Introduction

Taiwan has become the first in Asia to legalize same-sex marriage. Nevertheless, intense debate on whether same-sex marriage should be legalized between the pro- and against-same-sex marriage groups has not lessened. In the debate, discourses around homosexual has been reproduced and circulated, many of which concern the consequence of same-sex marriage, such as homosexual education, civil law of adopting child and homosexual parenting. This paper aims to address the question of the conflict between opposing stances at the level of lexical semantics. I will use 'homosexual' as a keyword and examine how the word is used in each stance on the issue with help of distributional corpus analysis. However, this issue is complicated by the fact that there are two synonymous words in Mandarin Chinese, that is *tóngxìngliàn* and *tóngzhì*, both of which are used frequently and can be used under similar register and genre. This paper addresses two research questions:

1. From their collocational behavior, what are the semantic difference between *tóngzhì* and *tóngxìngliàn*?

2. What are the characteristic usages of these two words in the opposing stances? What is the relationship between the stances and choices of words?

Below I will first introduce these two words and discuss in what sense they are near synonymous.

1.1 Origin of the two words: *tóngxìngliàn* and *tóngzhì*

In English world, the word homosexual, according to Tamagne (2004), came into existence at the end of the 19th century and was allegedly first used by the Hungarian journalist Karoly Maria Kertbeny in 1869. Before the introduction of the homosexual-heterosexual distinction, "being homosexual was not seen as a quality

of the individual but as a quality of a single act, which was equated with sodomy" (Tamagne, 2004: 7).

Concerning the word denoting homosexual in Chinese, the word *tóngxìngliàn* came earlier than *tóngzhì*. *tóngxìngliàn* came from Japanese 同性愛 *douseiai*, which was the translation of the English word *homosexual* around 1900 (Sang, 2014: 111-4). Other morphological related lexical items are *yìxìngliàn* 'heterosexual' and *shuāngxìngliàn* 'bisexual'.

On the other hand, *tóngzhì* originally means 'comrade'. This sense of *tóngzhì* is still in use, especially in the context of political parties. It is then intriguing to answer the question that how the word with political connotation obtained its meaning as the label for homosexual at that time. One commonly accepted answer is that, the meaning of homosexual was first given to *tóngzhì* by Mai-Ke Lin (林邁克) and then carried over through the influence of Hong kong directress Yi-Hua Lin (林奕華) (Ji, 2015). This standardized answer is not without problem. For example, Ji (2015) gave a more detailed historical view on how the word *tóngzhì* was appropriated and adapted into Taiwan society.

1.2 Synonymity: *tóngxìngliàn* and *tóngzhì*

Synonymy is one of paradigmatic semantic relations of two lexical items whose similarity in meaning is more striking than their difference. The issue of synonym concerns lexical choice (Glynn, 2010), meaning that given two words with synonymous meaning, our task as linguists is to ask what features determine not one word but the other one should be used in a given context.

At the first glance, the senses of the two words seem not to be synonymous but only similar. *Tóngxìngliàn* denotes a sexual orientation, whereas *tóngzhì* refers to the person who has such sexual orientation. If we use the classification of synonymy given by Cruse (2010: 142-5), synonym is said to have three sub-types: absolute-synonym, propositional-synonym and near-synonym. The absolute-synonym is primarily used as a theoretical endpoint whose existence in real language is rare, because difference in form implies difference in meaning. Therefore *tóngzhì* and *tóngxìngliàn* are not of this type. Second, two words are said to be propositional-synonym if the truth value remains the same when they are used a sentence. The examples below show that *tóngxìngliàn* and *tóngxìngliàn* are not of propositional-synonym.

同性戀/* 同志在以前被視作一種心理疾病。

*tóngxìngliàn/*tóngzhì zàiyǐqián bèi shìzuò yīzhǒng xīnlǐjíbìng*

'Homosexual/Gay is used to be considered as mental illness.'

The abnormality of using *tóngzhì* in this sentence lies in the fact that it does not denote the concept of a kind of sexual orientation as *tóngxìngliàn* but the person who has such sexual orientation, which suggests that these two words are far from being synonymous. Just as what Cruse has claimed, "The borderline between propositional synonymy and near synonymy is at least in principle clear ... (while) the borderline between near synonymy and non-synonymy is much less straightforward." Although *tóngzhì* and *tóngxìngliàn* are more closed to the side of non-synonym at the first glance, I am now going to argue that they are near synonym in a specific context, that is, when they are used as a compound modifier in N-N compounds such as *tóngzhì/tóngxìngliàn-hūnyīn* 'homosexual marriage', *tóngzhì/tóngxìngliàn-yùndòng* 'homosexual movement', *tóngzhì/tóngxìngliàn-jiātíng* 'homosexual family' and *tóngxìngliàn/tóngzhì-yìtí* 'homosexual issue'. In these compounds, the difference of homosexual as people and homosexual as a sexual orientation between *tóngzhì* and *tóngxìngliàn* no longer exists.

2 Literature review

In English world, Baker (2004) used corpus to discover that those who were in favor of the reform, which sought to lower the age of consent for homosexual intercourse, used language that framed homosexuality as an identity, whereas those who opposed the reform chose a wording that framed homosexuality as behavior. Engelhart (2012) studies three frequent adjectives de-

noting homosexuality, that is, homosexual, gay, and lesbian, in the COCA corpus with statistical methods, and discovered two results: first, gay and lesbian have similar usage, whereas homosexual reveals a distinct usage; second, homosexual and gay show a higher tendency for negative usage than lesbian.

In Chinese context, Wong and Zhang's (2000) explored the word *tóngzhì* in G&L Magazine, a magazine focuses on sexual minorities in Hong Kong, Taiwan and overseas. They found that the word *tóngzhì* was frequently used with words such as *battle*, *war*, *fight back* and so on, to create an imagined community that shaped the ideologies of equal rights for gays and lesbians. Wong (2005) focuses on the reappropriation of *tóngzhì* in 126 articles of *Oriental Daily News*, a mainstream newspaper in Hong Kong. Wong argues that *tóngzhì* is not a positive term, and the term *tóngxìngliàn* often appears in the news of medical and legal content. Moreover, *tóngzhì* often appears in highly sensationalized news stories, such as murder, fights and domestic disputes of gay couples. He concluded that, *tóngzhì* does not represent sexual minorities in general, but lesbian and gay people who participate in improper behavior. Wong (2008) directly interviewed and collected data on activists'use of *tóngzhì* as well as similar labels such as gay and *tóngxìngliàn*. Focusing on the semantic change of *tóngzhì* from comrade to sexual minorities, Wong argues that the semantic change of social labels is motivated by speakers'desire to take different stances and to project different personae.

In Taiwan context, Lin (2014) examined the collocational profile of the lexical item *same-sex marriage* in newspaper media in Taiwan from a diachronic perspective, and found that the discourse around *tóngzhì* has transformed from private domain to public domain. Chen (2018) explored the linguistic representation of homosexuality, and analyzed discourses to explore public perception toward homosexuality. She found that The collocational profile of *tóngxìngliàn* overlaps with that of *tóngzhì*.

3 Methodology

3.1 Corpus collection and processing

In order to explore the language used by the opposing groups on the issue of same-sex marriage, I compiled two corpora by collecting the electronic texts from the official websites of *Happiness for the Next Generation*[1] (hereinafter referred to as OPP) and *LGBT Families Info*[2] (hereinafter reffered to as PRO), a blog maintained by *Taiwan Alliance to Promote Civil Partnership Rights*. In terms of representativeness, these two websites were chosen for the pro- and against-same-sex marriage issue because the former was the initiator of the referendum, commonly called as anti-gay referendum, held in November, 2018, while the latter was one of the main supporting organizations of gay right.

When doing contrastive study in corpus linguistics, the genre of two corpora should also be taking into consideration. The two corpora chosen are in the form of weblog, where texts are organized into articles of accessible length for general online readers.

The data were crawled from the Internet using Python 3. After preprocessed through the pipeline of removing non-Chinese characters including foreign names, URLs and passages written in English, the texts were put into CkipTagger (Li et al. 2019), the word segmenter and POS tagger developed by CKIP (Chinese Knowledge and Information Processing) Lab, Academia Sinica. Having transformed the unstructured blog posts into segmented words and tagged POS, the texts were then put into Corpus Workbench (Evert and Hardie 2011), a corpus backend engine aimed at linguistic research, for corpus compilation, and CQPweb (Hardie 2012), the frontend UI capable of doing concordance, collocation and other frequency-based analysis for CWB.

Table 1 below shows the statistics of the two corpora.

[1] https://taiwanfamily.com/
[2] https://lgbtfamiliesinfo.tw/

	OPP	PRO
article	309	514
word token	372884	374321
words per article	1206.74	728.25
word type	20020	20480
type-token ration	0.054	0.055

Table 1: Statistics of the corpora of OPP and PRO.

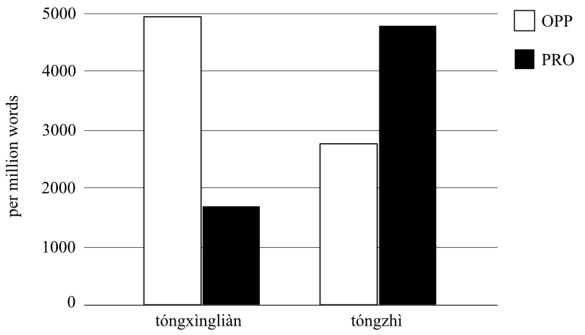


Figure 1: frequency of *tóngxìngliàn* and *tóngzhì* in each corpus

3.2 Frequency and collocational analysis

All corpus-based studies are based on the concept of frequency, from comparison of raw frequency to extracting significant collocation using statistical methods. This study follows the principle that the meaning of words is constructed and maintained by patterns of collocations. As Hunston (2002) indicated, repeated collocations reflect social meanings which can even be in direct opposition to what is claimed. Baker (2008) also mentioned that collocation is an important vehicle for the discursive presentation of social groups. Likewise, Romaine (2001: 153-176) claims that "connotations of words do not raise from words themselves but from how they are used in context".

Figure 1 is the frequency comparison of the occurrence of *tóngzhì* and *tóngxìngliàn* in each corpus. Note that the unit of frequency has been normalized to per million word.

As a preliminary impression, *tóngxìngliàn* is preferred by OPP while *tóngzhì* by PRO. The reason of this discovery will be given later when we look into how these words are used.

For the collocational analysis, Pointwise Mutual Information (PMI) (Church and Hank 1990) is used as the association measure for significant collocations.

The way I approach my first research question is that if there are similar collocates in terms of semantic field in both corpora, then those words within the same semantic field will be regarded as a semantic component of the node word. For example, see Appendix B for the collocate *zuì* 'crime' in OPP and the collocate *chúzuìhuà* 'decriminalization' in PRO. Though having slightly different word form, their belonging in the semantic field of CRIME is rather self-evident. Note that here I am not searching for the semantic prosody of *tóngxìngliàn*.[3] Instead, what I am concerned with is the fact that when the aspect of homosexual as a crime is being discussed, which term, *tóngzhì* or *tóngxìngliàn*, is preferred. To approach the first research question, I use the window size of *5L* to *5R* with the threshold of PMI value set at 3.

To approach the second research question, the method is basically similar with the one mentioned above except a slight difference where window size is now limited to +1. The reason for this is related to the fact that *tóngzhì* and *tóngxìngliàn* are near synonyms only when they are used as compound modifier, as what has already been mentioned in Section 1.2.

4 Result and discussion

This section is divided into two subsections, 4.1 discusses the result for the first research question and 4.2 and 4.3 for the second one.

4.1 Similar usages of *tóngzhì* and *tóngxìngliàn* in both stances

This section deals with the result concerning my first research question. Below I will first discuss the case of *tóngxìngliàn* and then the case

[3]That is, I am not arguing that for a given stance, *tóngxìngliàn* has a negative semantic prosody because one of its significant collocate is "crime". Such argument is difficult to justify in that even if crime turns out to be one of homosexual's collocate, the statement that "homosexual is a word with negative semantic prosody" is still undecidable because one needs to further examine how the words, i.e. the crime of homosexual, are used in the actual context.

of *tóngzhì*. Please refer to Appendix A and Appendix B for the complete collocation table ranked by PMI value.

The first set of collocations of *tóngxìngliàn* within the same semantic field in both corpora is related to homosexual as crime, for example, *zuì* 'crime' in OPP and *dìngzuì* 'convict', *chúzuìhuà* 'decriminalization', *fēifǎ* 'illegal' and *jìnlìng* 'prohibition' in PRO. The second set of significant collocates of *tóngxìngliàn* used by both sides are homosexual as disease, for example, *bìnghuà* 'become disease' in OPP and *juānxiě* 'blood donation' and *jíbìng* 'disease' in PRO.

On the other hand, in the case of *tóngzhì*, the first set of collocations in both corpora is related to gay rights movement, for example, *yóuxíng* 'demonstration', *yùndòng* 'movement' and *rèxiàn* 'hotline'[4]. In Chinese, while the compound *tóngxìngliàn-dàyóuxíng* 'gay pride' is also acceptable, both stances, when mentioning the gay pride held every year, prefer using *tóngzhì-dàyóuxíng* 'gay pride'. The second set of words that collocate with *tóngzhì* in both corpora are words relating to education: *jiàoyù* 'education', *guózhōngxiǎo* 'elementary and junior high school', *zhōngxiǎoxué* 'elementary and junior high school' and *shíshī* 'to put into effect'[5]. These words appear in the context of discussing the pros and cons of the gender equality education.

From the observation given above, the difference between the near synonym pair *tóngxìngliàn* and *tóngzhì* can be summarized in Figure 2.

tóngxìngliàn can be differentiated with *tóngzhì* along an internal-external axis. From the data above we can observe that *tóngxìngliàn* is used by both stances when its aspect of disease and crime is under discussion. Moreover, its meaning as a kind of sexual orientation is also manifested by the collocate *yìxìngliàn* 'heterosexual'. All of these collocates can be seen as related to a rather individual and internal aspect. On the other hand, *tóngzhì* has a considerably different collocational behavior in both corpora. While being synonymous to *tóngxìngliàn*, *tóngzhì* is commonly used in the external and social aspect of homosexual such as education, movement and demonstration.

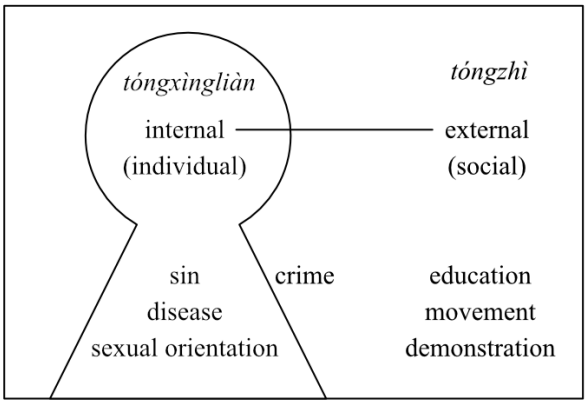


Figure 2: the difference between the near synonym pair *tóngxìngliàn* and *tóngzhì*

4.2 Characteristic usages of *tóngxìngliàn* in each stance

Having discussed the similarity of the use of *tóngxìngliàn* and *tóngzhì* in both stances, in this section I will address the second research question concerning characteristic usages of these two terms. As what I have mentioned already, for this question to be answered, the window size of calculating collocation has been confined to *1R* because *tóngxìngliàn* and *tóngzhì* are near synonym when serving as compound modifiers. That is to say, limiting the window size to *1R* means that the characteristic usages to be discovered are actually characteristic compound heads modified by *tóngzhì* and *tóngxìngliàn* in each corpus. Please refer to Appendix C and Appendix D for the complete collocation table ranked by PMI value. The grey background of a given collocate indicates that the collocate is considered a characteristic one, which only appears in just one corpus. I will first discuss the usage of *tóngxìngliàn*. The case of *tóngzhì* will come afterwards.

In the corpus of OPP, there are six groups

[4]This collocate comes from frequent mention of the name of an organization promoting gay rights called *Taiwan Tongzhi (LGBTQ+) Hotline Association*

[5]This word is mostly used in the context of putting the gender equality education into effect.

of characteristic collocates of *tóngxìngliàn*. The first is *tóngxìngliàn-xíngwéi* 'homosexual behavior'. Although the word *behavior* is a general term for nominalization of an action, when modified by the word *homosexual*, its negative connotation frequently used in the context of illegal behavior is evident.

The second is *tóngxìngliàn-wénhuà* 'homosexual culture'. In using such compound, homosexual is framed as a habit, trend and fashion. Therefore, homosexual is considered to not to be an innate sexual orientation but a conventionalized practice. Examining further into the verbs that are used with the term *homosexual culture*, verbs of higher emotional loading such as *qiǎngpò* 'to force', *yòudǎo* 'to seduce', *huóyuè* 'to be proactive', *gǔlì* 'to encourage' and *chōngmǎn* 'to be filled with' are found.

The third is *tóngxìngliàn-shēnghuó* 'homosexual life'. The use of this term implies that being homosexual is a kind of life style one can consciously choose to follow or to quit. Therefore, the sense of homosexual is no longer an inner state of a person. Instead, it is converted to visible behavior. Put it another way, for the opponent of same-sex marriage, homosexual is not something to be but something to be seen.

The fourth is *tóngxìngliàn-yùndòng* 'gay rights movement'. In the previous section, I have already shown that *yùndòng* 'movement' is a significant collocate of *tóngzhì* in both corpora. And here we found that in OPP, *tóngxìngliàn-yùndòng* is also a significant compound. This observation partly explained the fact that *tóngxìngliàn* is preferred by the opponent of same-sex marriage as mentioned in Section 3.2.

The fifth is a group of words related to labelling and categorization: *zúqún* 'group', *shèqún* 'community' and *quānzi* 'circle'. Although it is not uncommon in everyday life to refer to people with the same attribute as a whole individual group, using these terms often involves over-generalization and simplification. While there exists theoretically "a group" of homosexual people, each of the member has undoubtedly different life experiences and personalities. Therefore, when using terms, the diversity inside a group is ignored and the homogeneity is emphasized and assumed.

On the other hand, the eight significant (PMI > 3) collocates of *tóngxìngliàn* in the corpus of pro-same-sex marriage as shown in Appendix C, compared to 23 significant ones in the opposite stance, are very few. Among the five characteristic collocates, the enumeration comma ("、")[6] has the most frequent occurrence. As we dive into the context where *tóngxìngliàn* and the enumeration comma collocate with each other, we found that in the corpus of the supporter of same-sex marriage, *tóngxìngliàn* is significantly juxtaposed with different sexual orientations such as heterosexual, bisexual and transgender. In these cases, homosexual is seen as but one kind of sexual orientation, corresponding to its original meaning.

4.3 Characteristic usages of *tóngzhì* in each stance

In the corpus of against-same-sex marriage, there are two groups of characteristic collocates of *tóngzhì*. The first group is words related to revelry such as *hōngpā* 'home party', *sānwēnnuǎn* 'sauna' and *wǔ huì* 'dancing club'.

From these concordance lines we can clearly see that these terms are not merely mentioned for the entertainment place for the homosexual, but they are mostly linked negatively to sexual behavior.

The second set of collocation is *quānnèi* 'inside circle', which also corresponds to the result in section 4.2. This term also shows the tendency of marking the homosexual as a homogeneous group.

In the corpus of pro-same-sex marriage, there are three characteristic collocates that worth discussing, that is, *jiātíng* 'family', *hūnyīn* 'marriage' and *rénquán* 'human right'. In using the compound *tóngzhì-jiātíng* 'homosexual family', the supporter of same-sex marriage is discussing whether the children raised by homosexual parents will have different consequence compared to those raised by heterosexual parents.

[6]This punctuation is called 頓號 *dùnhào* in Chinese. It is mainly used for enumeration.

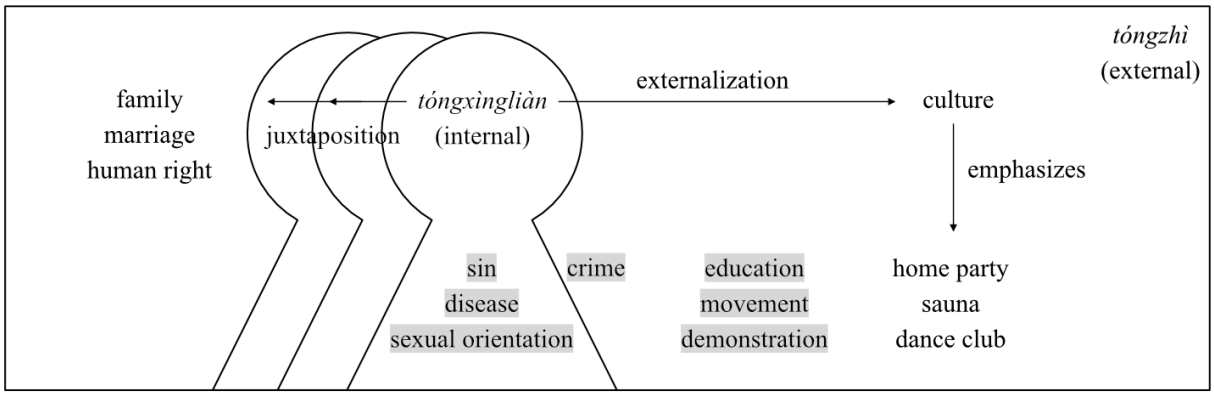


Figure 3: Summary of the finding in terms of the image of semantic field

The second one is *tóngzhì-hūnyīn* 'homosexual marriage', which is unexpected when this compound appears. At first I assumed that both sides were all discussing the issue of same-sex marriage, therefore this term should have appeared frequently in both sides and should not have significantly showed up in only one side. But it turned out that the supporters explicitly discussing the issue of same-sex marriage by using the compound, while the opponents prefer the compound *tóngxìng-hūnyīn* 'same-sex marriage' and not *tóngzhì-hūnyīn* 'homosexual marriage'. The third compound is *tóngzhì-rénquán* 'human right of the homosexual'. It shows that for the supporter, homosexual's human right is often used as a justification for same-sex marriage. The compound itself not only stands for the combination of two nouns, it also makes such concept, which is blended by the two components, to become "more real", in the sense that the status of its existence is just like a normal noun that is durable and persistent in time-space. Such effect should not be ignored. On the other hand, we see no significant mentioning of the same compound in the corpus of the opponent, because for them, although human right is a common value that should be pursued, the marriage right of the homosexual does not belong to the scope of their definition of human right. It is by this reason that the compound *human right of the homosexual* did not appear in the corpus of against-same-sex marriage.

4.4 The full picture

After examining the characteristic usages of *tóngzhì* and *tóngxìngliàn* in each stance in 4.2 and 4.3, now we are finally able to think about how these distinctive collocational behavior can be mapped onto the image of semantic field concluded from my first research question (see 4.1 and Figure 2). Figure 3 is the complete image that is used to demarcate and summarize the observation.

The words in grey background are those from the first research question, i.e. the common collocates of *tóngzhì* and *tóngxìngliàn* in both corpora. The left part of Figure 3 demonstrates the usage of pro-same-sex marriage side, and the right part shows the usage of against-same-sex marriage stance.

Let's first discuss the stance against same-sex marriage. While *tóngxìngliàn* is said to denote the internal and individual aspect of homosexual, we found that in the usage of the opponent of same-sex marriage, the characteristic collocates mentioned above show the tendency of externalization. *tóngxìngliàn* is no longer just a sexual orientation that requires no visible feature. On the contrary, it becomes a kind of individual behavior that is under examination by others, a life style open to individual choice and a culture that is excessively encouraged by the activists. In particular, when using *tóngzhì*, whose usage is the emphasis on the external and social aspect of the homosexual, the so-called homosexual culture is focused on its part of rev-

elry. Such culture is chiefly depicted as indulgence in sex and drug. Furthermore, these two terms, *tóngxìngliàn* and *tóngzhì*, are both used by the opponent of same-sex marriage to label the homosexual as a whole homogeneous group.

On the other hand, in the language use of the stance of pro-same-sex marriage, we first see that the word homosexual is often juxtaposed with other sexual orientation by the enumeration comma, implying that homosexual is actually equal to the others. Moreover, in terms of the social aspect of the word *tóngzhì*, we found that words related to family and marriage are significantly used by the supporter. The human right of homosexual is also used as justification for the support of same-sex marriage.

5 Conclusion

In this research, the intense conflict of the issue of same-sex marriage in Taiwan is examined by the method of corpus linguistics and quantitative evidence. In particular, the language use of the opposing stances is studied through the pair of near synonym *tóngzhì* and *tóngxìngliàn*. Two relevant research questions related to lexical semantics are addressed, i.e., what the semantic difference of these two words is and how these two words are characteristically used by each stance. For the first research question, it is found that the semantic difference of the near synonym pair, *tóngzhì* and *tóngxìngliàn*, can be demarcated along the axis of internal/external in a semantic field. While *tóngxìngliàn* as a sexual orientation denotes the concept of an individual's inner attribute, *tóngzhì* refers to the outer behavior of those individuals. In terms of the characteristic language use in each stance, it is discovered that the opponent of same-sex marriage tend to externalize the innateness of *tóngxìngliàn*. Homosexual is then attributed to all the visible aspects, of which the indulgence to sex is emphasized. As for the supporter of same-sex marriage, homosexual is often juxtaposed with other sexual orientations, which implies that it should not be considered to be a distinctive one compared to the others. In addition, by using the strategy of compounding, the statements they are arguing are nominalized.

In this study, we found that the boundary between semantic fields are dynamic and flexible. Especially when combined to another word to become a compound, the meaning of homosexual becomes negotiable. A compound as a nominal unit indirectly assumes the existence of the entity each stance aims to argue for, be it homosexual culture, behavior and life for the opponent of same-sex marriage or the human right of homosexual and homosexual marriage for the supporter of same-sex marriage.

It is hoped that this study has unfolded insightful aspects of the opposing stances from the perspective of lexical semantics.

References

Baker, Paul. (2004). 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of sociolinguistics, 8*(1), 88-106.

Baker, Paul. (2006). *Public discourses of gay men.* London: Routledge.

Chen, Yu-Qin. (2018). *Analyzing homosexuality: Lexical collocations and social attitudes in Taiwan's Internet news* (Unpublished master's thesis). National Taiwan Normal University.

Church, Kenneth, and Patrick Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16*(1), 22-29.

Cruse, Alan. (2010). *Meaning in language: An introduction to semantics and pragmatics* (3rd ed.). Oxford: Oxford University Press.

Engelhart, Jasmin. (2012). *Homosexual, gay and lesbian: A corpus study on words denoting homosexuality* (Unpublished bachelor's thesis). University of Vienna.

Evert, Stefan, and Andrew Hardie. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference.*

Glynn, Dylan. (2010). Synonymy, lexical fields, and grammatical constructions: A study in usage-based Cognitive Semantics. In Hans-Jörg Schmid and Susanne Handl (Eds.), *Cognitive Foundations of Linguistics Usage Patterns* (pp. 89-118). Berlin, Boston: De Gruyter Mouton.

Hardie, Andrew. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis

tool. *International Journal of Corpus Linguistics, 17*(3), 380-409.

Hunston, Susan. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Ji, Da-Wei. (2015). Fānyì de gōnggòng: àizī, tóng zhì, kùér [Publicity of translation: AIDS, homosexual, and queer]. *Bulletin of Taiwanese Literature, 20*, 75-112.

Li, Peng-Hsuan, Tsu-Jui Fu, and Wei-Yun Ma. (2019). Remedying BiLSTM-CNN Deficiency in Modeling Cross-Context for NER. *ArXiv, abs/1908.11046*.

Lin, Yi-Xuan. (2014). *A corpus-assisted investigation to the representation of gay marriages in Taiwan's newspapers (2005-2014)* (Unpublished master's thesis). National Chengchi University.

Romaine, Suzanne. (2001). A corpus-based view of gender in British and American English. *Gender across languages, 1*, 153-175.

Sang, Zi-Lan. (2014). *Fúxiànzhong de nǚtóngxìngliàn: Xiàndài zhōngguó de nǚtóngxìng àiyù* [The emerging lesbian: Female same-sex desire in modern China]. Taipei, Taiwan: National Taiwan University Press.

Tamagne, Florence. (2004). *A history of homosexuality in Europe: Berlin, London, Paris 1919-1939*. New York, NY: Algora.

Tracy, Karen. (2016). *Discourse, Identity, and Social Change in the Marriage Equality Debates*. New York, NY: Oxford University Press.

Wong, Andrew, and Qing Zhang. (2000). The linguistic construction of the tongzhi community. *Journal of Linguistic Anthropology, 10*(2), 248-278.

Wong, Andrew. (2005). The reappropriation of tongzhi. *Language in Society, 34*(5), 763-793.

Wong, Andrew. (2008). The trouble with tongzhi: The politics of labeling among gay and lesbian Hongkongers. *Pragmatics, 18*(2), 277-301.

A Simple and Efficient Ensemble Classifier Combining Multiple Neural Network Models on Social Media Datasets in Vietnamese

Huy Duc Huynh
University of Information Technology
VNU-HCM, Vietnam
16520508@gm.uit.edu.vn

Hang Thi-Thuy Do
University of Information Technology
VNU-HCM, Vietnam
16520339@gm.uit.edu.vn

Kiet Van Nguyen
University of Information Technology
VNU-HCM, Vietnam
kietnv@uit.edu.vn

Ngan Thuy-Luu Nguyen
University of Information Technology
VNU-HCM, Vietnam
ngannlt@uit.edu.vn

Abstract

Text classification is a popular topic of natural language processing, which has currently attracted numerous research efforts worldwide. The significant increase of data in social media requires the vast attention of researchers to analyze such data. There are various studies in this field in many languages but limited to the Vietnamese language. Therefore, this study aims to classify Vietnamese texts on social media from three different Vietnamese benchmark datasets. Advanced deep learning models are used and optimized in this study, including CNN, LSTM, and their variants. We also implement the BERT, which has never been applied to the datasets. Our experiments find a suitable model for classification tasks on each specific dataset. To take advantage of single models, we propose an ensemble model, combining the highest-performance models. Our single models reach positive results on each dataset. Moreover, our ensemble model achieves the best performance on all three datasets. We reach 86.96% of F1-score for the HSD-VLSP dataset, 65.79% of F1-score for the UIT-VSMEC dataset, 92.79% and 89.70% for sentiments and topics on the UIT-VSFC dataset, respectively. Therefore, our models achieve better performances as compared to previous studies on these datasets.

1 Introduction

The rapid development of science and technology in the world has created a vast amount of data. In particular, the growth of social networks continuously creates a huge amount of comments and posts which are valuable sources to exploit and analyze in the digital era. Text classification is a prerequisite for such works as analyzing user opinion in the network environment, filtering and removing malicious information, and detecting criminal risk. With great potential, text classification has attracted much attention from experts in the natural language processing community worldwide. In English, we easily search for a range of text classification publications in many fields. However, relatively few researches have been done on Vietnamese text. Most published articles focus on binary classification. However, a large amount of information today requires analysis in many more aspects (multi-label or multi-class). The lack of knowledge and techniques for the Vietnamese language makes us decide to conduct this research to classify multi-class text for Vietnamese social media datasets. These datasets are provided from the VLSP share-task and publications on text classification. In particular, there are various social media textual datasets such as UIT-VSMEC for emotion recognition (Ho et al., 2019) and UIT-VSFC for students' feedback classification (Nguyen et al., 2018b) and HSD-VLSP for hate speech detection (Vu et al., 2019). These are the datasets with multi-label and imbalance between the labels that have been published recently. They are suitable for the requirements that we would like to study.

The emergence of deep neural networks (Liu et al., 2017) and word embeddings have made text classification more efficient. Pre-trained word embeddings accurately capture semantics to assist deep

learning models improve the efficiency of classification. In this study, we implement deep learning models such as CNN (Kim, 2014), LSTM (Hochreiter and Schidhuber, 1997) and their variants to solve classification problems. Besides, we implement the BERT model (Devlin et al., 2018), which is a state-of-the-art model in many natural language processing tasks in recent years. BERT is trained through the transformer's two-dimensional context (a neural network architecture based on the self-attention mechanism to understand languages). BERT is in contrast to previous deep learning models that looked at a text sequence from left to right or combined left-to-right and right-to-left training. To improve the word representation, we create a normalized words dictionary, which helps recognize words included in pre-trained embedding but is not represented due to misspellings.

As a result, CNN model combined with fastText's pre-trained embedding (Grave et al., 2018), has been remarkably performance on Vietnamese social media datasets. Our study also proves the efficiency of BERT on Vietnamese students' feedback dataset. Besides, we combine single models to increase the efficiency of the classification. As a result, our ensemble model accomplishes higher results than the single model. Compared to previous studies done on the datasets, our models achieve better results.

2 Related Work

Nowadays, many organizations realize the importance of sentiment analysis for consumer's feedback. Through this feedback, they can evaluate the quality of their services or products and devise appropriate strategies. In order to predict the genre and rating of films through viewer ratings, Varshit battu and his collaborators (Varshit et al., 2018) conducted research on viewer comment data collected from many websites. They implemented various classification methods on their dataset to evaluate the effectiveness of methods. As a result, the CNN model achieved high results in many different languages.

The detection of emotions in texts has become an essential task in natural language processing. Su et al. (2018) studied the text emotional recognition problem based on semantic word vector and emotional word vector of the input text. Their proposed method used the LSTM model for emotion recognition by modeling the input text's contextual emotion evolution. They use five-fold cross-validation to evaluate the performance of their proposed method. As a result, their model achieved recognition accuracy of 70.66% better than the CNN-based method which was implemented in the same dataset.

In addition, hate speech detection is increasingly concerned because of the explosion of social networks. There has been an amount of successful research in this field. To complete the offensive task of categorizing tweets that were announced by SemEval competition in 2019. Nikolov and Radivchev. (2019) used different approaches and models towards offensive tweet classification. Their paper presented pre-processing data methods for tweets as well as techniques for tackling imbalanced class distribution in the provided test data. Their experiments show that the BERT model proved its outstanding advantages in text classification. Not only did it outperform conventional models on the validation set, but also based on the results from the test set, it did not cause the over-fitting issue.

In Vietnam, there have been some studies efforts for text classification tasks, as well as contributing Vietnamese data for the research community. Pham et al. (2017) announced a neural network-based toolkit namely NNVLP for essential Vietnamese language processing tasks, including part-of-speech tagging, chunking, named entity recognition. This toolkit achieved state-of-the-art results on these three tasks. With the two of UIT-VSMEC (Ho et al., 2019) and UIT-VSFC (Nguyen et al., 2018a) datasets we used in this study, their authors performed classification tasks using a variety of deep learning methods. On the UIT-VSMEC dataset, Ho et al. (2019) used Random Forest, SVM, LSTM, and CNN models to classify emotions of comments. They achieved 59.74% with seven labels and 66.48% with six labels by using the CNN model. With the UIT-VSFC dataset, Nguyen et al. (2018a) gained the highest result by the BiLSTM model with 92.03% on sentiment and 89.62% on the topic label.

The above studies have shown the superiority of the deep learning models in text classification, which is the premise for us to apply them to the Vietnamese datasets in this study. By modifying the models and implementing new models, we aim to

bring better results for these Vietnamese social media datasets.

3 Datasets

In this task, we conducted experiments on classification methods on three Vietnamese datasets, including UIT-VSMEC (Ho et al., 2019), UIT-VSFC (Nguyen et al., 2018b), and HSD-VLSP (Vu et al., 2019). UIT-VSMEC (Ho et al., 2019) is provided by Ho et al, the items in this dataset are the Vietnamese's comments from social networks. This dataset contains exactly 6,927 emotion annotated sentences with seven emotion labels: ENJOYMENT, SADNESS, ANGER, SURPRISE, FEAR, DISGUST, and OTHER. UIT-VSFC is provided by Nguyen et al. (2018b). This dataset was constructed from students' responses from a university for the sentiment classification task. This dataset includes 16,175 items with two classification parts that are important: sentiments and topics. Each sample of the training dataset is assigned one of three sentiment labels POSITIVE, NEGATIVE, or NEURAL and assigned one of four topics' labels LECTURERS, TRAINING PROGRAM, FACILITIES, OTHER. The HSD-VLSP dataset is a Vietnamese comments dataset about Hate Speech Detection on social networks provided by the VSLP 2019 shared-task (Vu et al., 2019). This dataset contains the comments and posts on Facebook social networks, including 25,431 items. Each data line of the training dataset is assigned one of three labels CLEAN, OFFENSIVE, or HATE. Table 1 shows examples for each dataset and its characteristics.

4 Methodology

In this task, we use several pre-processing techniques and deep learning models (CNN (Kim, 2014), LSTM (Hochreiter and Schidhuber, 1997) their variants, and BERT (Devlin et al., 2018)) to apply on the datasets. Each model has its own strength on each label. Therefore, after implementing single models, we propose a simple and efficient ensemble approach combining the best neural network models to improve the classifier performance. Figure 1 shows an overview of the ensemble model for Vietnamese social media text.

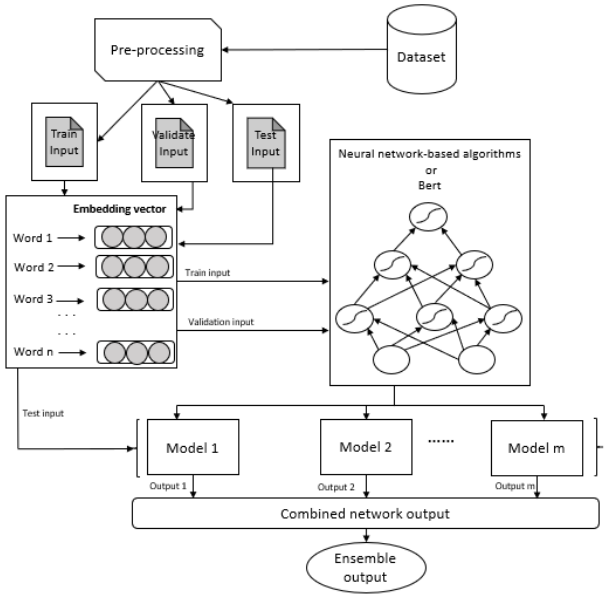


Figure 1: Overview of the ensemble model combining neural network-based model for Vietnamese social media datasets

4.1 Pre-processing

We implement several pre-processing techniques for Vietnamese comments/posts, as follows: All words are converted to lowercase. URLs and non-alphabetic characters are removed (includes number, excess whitespace). Also, we use a dictionary to convert abbreviations and slang words to normal to increase the performance of pre-trained embedding as well as use vnTokenizer to tokenize words. (Word segmentation is an important task that helps models achieve better results, especially for the Vietnamese language. More and more research works on this; these works use different methods and achieve high results, such as using CRFs and SVMs (Nguyen et al., 2006), Hybrid Approach (Huyen et al., 2008). In this study, we use vnTokenizer (Huyen et al., 2008) to tokenize words. Table 2 shows some examples in our dictionary.

4.2 Word Embeddings

In this task, we use three pre-trained embeddings for the Vietnamese language, which are currently assessed as the best performance embeddings. Those are fastText (Grave et al., 2018), Word2vec (Vu, 2016) with 300 dim and 400 dim. fastText (Grave et al., 2018) is used for many languages and brought

Datasets	Labels		Percentage (%)	Examples	
				Vietnamese sentences	English sentences
UIT-VSMEC	ENJOYMENT		28.36	Tốt quá!	Very good!
	SADNESS		19.31	Nay buồn!	Today, I feel sad!
	ANGER		16.59	Im đi!	Shut up!
	SURPRISE		6.92	Đẹp trai mà bị đá à!	You are handsome, but you broke up!
	FEAR		5.70	Tao sợ thật sự.	I'm really scared.
	DISGUST		4.46	Diễn viên già vãi -.-	The actor is so old -.-
	OTHER		18.66	Có thể bán lại cho tớ không?	Could you sell it for me?
UIT-VSFC	Sentiments	POSITIVE	49.80	Slide, giáo trình đầy đủ.	Slide and syllabus are full.
		NEGATIVE	45.80	Thầy chép bảng nhiều.	The lecturer writes on the board a lot.
		NEUTRAL	4.40	Tăng cường thiết bị..	Strengthen equipment.
	Topics	LECTURER	71.70	Thầy chép bảng nhiều.	The lecturer writes on the board a lot.
		PROGRAM	18.70	Slide, giáo trình đầy đủ.	Slide and syllabus are full.
		FACILITY	4.20	Tăng cường thiết bị.	Strengthen equipment.
		OTHER	4.70	Hài lòng về tất cả.	Satisfied about it all.
HSD-VLSP	CLEAN		91.49	Hôm nay trời nắng đẹp.	It is sunny today.
	OFFENSIVE		5.02	vkl.	Cuss.
	HATE		3.49	Nói đến vậy mà cũng không hiểu. Đồ ngu.	Saying that without understanding. Idiot.

Table 1: Overview statistics of the three Vietnamese social media textual datasets.

No.	Abbreviation	Vietnamese meaning
1	"Chờiiii"	"Trời"
2	"vkllllll"	"vkl"
3	"chetme"	"chết mẹ"
4	"kbh"	"không bao giờ"

Table 2: Examples of word normalization dictionaries

good results on social media data. Therefore, we use fastText (Grave et al., 2018) for the Vietnamese language in this task. Word2vec (Vu, 2016) is Word2vec model for the Vietnamese language and is trained on Vietnamese texts. Because they are Word2vec for the Vietnamese language, thus recognizing vast Vietnamese words and brings quite good results on datasets.

4.3 Convolutional Neural Network (CNN)

In this task, we use CNN (Kim, 2014) to classify emotions for Vietnamese text. CNN is an advanced deep learning model, which includes hidden layers such as pooling layers, convolutional layers, fully connected layers, and normalization layers. CNN achieved good results for this document classification task in general and the best on UIT-VSMEC (Ho et al., 2019) and HSD-VLSP (Luu et al., 2020).

4.4 Long Short-Term Memory (LSTM) and Variants

Like CNN (Kim, 2014), LSTM (Hochreiter and Schidhuber, 1997) is also a modern classification method. This method is strong in classification problems, and most of it has achieved high-performance classification results. Therefore, in this task, we decide to choose it to compare with other classification models. LSTM is a special kind of RNN (Medsker and Jain, 1999). LSTM's network architecture includes memory cells and ports that allow the storage or retrieval of information. We also use Bi-LSTM with Bidirectional (Schuster and Paliwa, 1997), BiLSTM can learn more contextual information extracted from two directions.

4.5 Gated Recurrent Units (GRU)

GRU is a variant of RNN and also achieves high results in classification problems. GRU has only two gates, that is the reset gate and the update gate. GRU does not have memory cells like LSTM (Hochreiter and Schidhuber, 1997). It has only the outputs to make decisions and to inform the next steps. These two gates filter the information from the inputs of the cell and provide an output that satisfies both the criteria of storage past information and the ability to make current decisions most accurately.

4.6 BERT (Bidirectional Encoder Representations for Transformers)

BERT (Devlin et al., 2018) is a transfer learning model that achieves state-of-the-art results on natural language processing tasks. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. The model can learn the context of a word based

on all of its surroundings. Fortunately, BERT is provided with a version for multilingual, including Vietnamese. We apply a BERT-based classifier with a pre-trained multilingual representation into the social media datasets in Vietnamese.

4.7 Our Proposed Ensemble Method

Ensemble approaches were very effective in previous studies (Huynh et al., 2020; Tran et al., 2020). We also find that each model has a high classification performance on certain labels so we create an ensemble model based on our implemented neural network models. For each sentence in the test set, the ensemble model selected the label according to the different models' votes. In the case of equal votes, the label was selected based on the model had the best classify performance for that label. Algorithm 1 shows the steps of the ensemble model.

Algorithm 1 - An ensemble method for Vietnamese social media textual data

Input: A social media text T and m top-performance models M_i $(0 < i < m + 1)$.
Output: Returning the label C such that our proposed ensemble method predicts.

procedure ENSEMBLE METHOD BASED ON VOTING
  Initialize dictionary D containing the votes of each label.
  for i = 1 to m **do**
    $L_i = M_i(T)$
    if $M_i(T)$ not in D **then**
      $D[L_i] = 1$
    else
      $D[L_i] = D[L_i] + 1$
    end if
    if len(D) != m **then**
      C = $L_i \in D.keys()$, where L_i is the label with the highest number of vote.
    else
      C is the label predicted by the model with the best performance.
    end if
  end for
  return C
end procedure

5 Experiments

5.1 Evaluation Metric

To make comparisons with previous studies, we used the corresponding measure based on previous studies' measure, including Weighted F1-score for UIT-VSMEC (Ho et al., 2019), Micro F1-score for UIT-VSFC (Nguyen et al., 2018b), with HSD-VLSP dataset, the organization has stopped providing a test set to ensure fairness in the next competition, so when the contest ended, we could not submit the predicted results to test the accuracy. When doing this study because of the lack of test set, we decide to use the k-fold cross-validation (k = 5) method and Macro F1-score (Luu et al., 2020) to evaluate the models we applied to HSD-VLSP.

5.2 Experimental Settings

In this study, we experiment with deep learning models that are evaluated as superior in the field of text classification, including LSTM, CNN, BERT, and their variants for the datasets. First, we clean the input data by removing expressions, numbers, special characters, and all words are converted to lowercase. Second, we use some pre-trained embedding that published for the Vietnamese language to represent words into vectors before putting them into deep learning models, including Word2vec and fastText's pre-trained embedding. In this step, with pre-trained embedding was trained on word-separated data, we separate the words for the datasets by the vnTokenizer (Huyen et al., 2008). fastText's pre-trained embedding demonstrates a more efficient representation of words with social datasets. However, we realize that there are some specific words to the classification classes, although represented in pre-trained embedding, they are abbreviated or written in slang, so they can not be recognized and ignored. Therefore, we try to optimize the performance of pre-trained embedding by creating word dictionaries to normalize the above words to the normal form so that they can be represented correctly. Whether it would improve our models' performance? The example of our dictionaries is presented in Table 2.

We also experiment with different values of parameters around the average value of sentences' length. To evaluate whether the pre-processing of

data has a significant effect on the classification results, we perform experiments on the processed data and the original data except for the UIT-VSFC dataset because it is cleaned before labeling. However, for the UIT-VSMEC, the pre-processing do not improve the efficiency compared to the original data. Therefore, we only show the results of the models on the original data for UIT-VSFC and UIT VSMEC dataset and results of the models on the preprocessing data for HSD-VLSP dataset in this study. After experimenting with single models, we combine our models using the max-voting method and model priority as presented in Algorithm 1.

5.3 Experimental Results

Our experiments achieve positive results, creating dictionaries that enhance our single model's performance 2%. With our single models, CNN model has shown its outstanding performance when combining with fastText's pre-trained embedding, which achieves the best results on social media commentary datasets. With the UIT-VSMEC dataset, the CNN model with three layers reach the best efficiency. Conversely, with the HSD-VLSP dataset, the CNN model has the best performance with five layers due to its larger size. Besides, BERT has higher performance than the other models when applying to the UIT-VSFC dataset. With a combination of the single models' strengths, our ensemble model has achieved high results. Table 3 shows our results through experiments performed.

5.4 Comparison With Previous Studies

The results of our single models as well as the ensemble model are better than those of previous studies that were conducted on the same data set. We follow the same metrics and test set that previous studies have used to make similar comparisons. Table 4, 5 and 6 show the best results we achieved compared to previous studies.

6 Result Analysis

Through the experiment on the UIT-VSMEC dataset, we find that the classification model got a high accuracy for SADNESS, and ENJOYMENT labels. The labels ENJOYMENT and SADNESS are the two labels that account for a high proportion in the dataset, so the correct classification rate for these labels is quite high. However, each model has its strengths in labels. CNN accomplish the highest accuracy classification rate on ENJOYMENT (73.70%) and is relatively accurate for the rest. BiLSTM has the best effect at DISGUST (65.90%) and SADNESS (63.79%) label. Although LSTM and GRU do not have superior results compared to other models on each label, they play an essential role in voting in the ensemble model. The ensemble model improves the predictions in the ENJOYMENT label from 73.70% to 75.64%, the SADNESS label from 63.79% to 67.20%, and from 63.04% to 71.73% for the FEAR label.

In the UIT-VSFC dataset, with the sentiments field, we find that the classifying model achieve a high right classifying rate on the NEGATIVE and POSITIVE label for the sentiment classification task. These two labels occupy approximately the same rate and are much higher than the NEUTRAL label. Therefore, when classifying, the NEUTRAL label is mistaken for the two, the classification results on the NEUTRAL label are low. With sentences that are too long, sentences that express many emotions in the sentence or sentences that are too short, lacking in subjects makes the model difficult to recognize the feelings of the sentence. The highest accuracy of the NEGATIVE and POSITIVE label is 95.95% and 94.71% when implementing BERT model. BiLSTM model achieves the highest performance for NEUTRAL label at 41.31% while CNN and LSTM have the classification results almost equal to the remaining models. Our ensemble model improves the accuracy of the NEGATIVE label from 95.95% to 97.01%.

For the topic classification task, correct classification was high on the LECTURER and FACILITY labels. The LECTURER label has the highest rate in the dataset with 71.70%, so the exact classification model is easy to explain. The LECTURER label and FACILITY label have the highest accuracy by using the BiLSTM model with 95.72% and 93.79%. The highest performance of PROGRAM and OTHER label is 81.29% and 48.42% when implementing BERT. The FACILITY label only accounts for the lowest rate in the dataset with 4.30%. This label contains words related to tools, facilities

Model	UIT-VSMEC		UIT-VSFC		HSD-VLSP
	Seven labels	Six labels	Sentiments	Topics	
BiLSTM + fastText (300 dim)	56.93	64.18	91.53	88.12	85.19
BiLSTM + Word2vec (300 dim)	56.16	62.24	90.84	88.53	78.99
BiLSTM + Word2vec (300 dim)	56.40	60.56	91.18	88.25	79.18
LSTM + fastText (300 dim)	52.12	59.02	91.06	87.42	85.56
LSTM + Word2vec (300 dim)	50.69	57.27	90.90	86.79	76.26
LSTM + Word2vec (400 dim)	49.32	56.58	90.58	87.36	80.84
GRU + fastText (300 dim)	52.88	59.45	90.77	88.02	84.95
GRU + Word2vec (300 dim)	54.02	59.70	90.74	87.71	77.84
GRU + Word2vec (300 dim)	54.30	59.75	90.87	87.68	78.38
CNN + fastText (300 dim)	**59.87**	**66.54**	90.42	87.68	**85.74**
CNN + Word2vec (300 dim)	32.65	33.76	89.16	85.94	79.01
CNN + Word2vec (400 dim)	23.76	47.33	89.79	86.16	81.63
BERT (Based-multilingual-cased)	49.22	55.97	**92.51**	**89.67**	65.11
Our proposed model	**65.79**	**70.99**	**92.79**	**89.70**	**86.96**

Table 3: F1-score performances of models on the test sets of various Vietnamese social media textual datasets.

Model	F1-score (%)	
	Seven labels	Six labels
CNN + Word2vec (Ho et al., 2019)	59.74	66.34
CNN + fastText (Our implementation)	**59.87**	**66.54**
Our proposed ensemble (GRU + CNN + BiLSTM + LSTM)	**65.79**	**70.99**

Table 4: The comparison with previous studies on UIT-VSMEC.

Model	F1-score (%)	
	Sentiments	Topics
BiLSTM + Word2vec (Nguyen et al., 2018a)	92.03	89.62
LD + SVM (Nguyen et al., 2018c)	92.20	-
BERT (Our implementation)	**92.51**	**89.67**
Our proposed ensemble (BERT + CNN + BiLSTM + LSTM)	**92.79**	89.38
Our proposed ensemble (BERT + CNN + BiLSTM)	92.13	**89.70**

Table 5: The comparison with previous studies on UIT-VSFC.

Model	F1-score (%)
Text-CNN (Luu et al., 2020)	83.04
Logistic regression* (Pham et al., 2019)	61.97
Logistic regression + Random Forest + Extra Tree* (Dang et al., 2019)	58.88
VDCNN, TextCNN, LSTM, LSTMCNN, SARNN* (Nguyen et al., 2019)	58.45
BiLSTM* (Do et al., 2019)	56.28
CNN + fastText (Our implementation)	**85.74**
Our proposed ensemble (CNN + BiLSTM + LSTM)	**86.96**

Table 6: The comparison with previous studies on HSD-VLSP. * indicates that the result is evaluated on a test set of the VLSP shared task 2019. Others use k-fold cross-validation to evaluate the model (k=5) following the study (Luu et al., 2020).

in the school. Hence, it was easy to identify and difficult to be confused with other labels. Although FACILITY label took up a small percentage, this label reached a high rate of correct categorization. The OTHER label is the one with the lowest accuracy rating because it is difficult to distinguish from the others. Our ensemble model does not increase the classification performance of each label superior to the single model, but it increases the model's overall prediction rate.

In the HSD-VLSP dataset, we find that the CLEAN label achieve the highest classification result, which is also the highest percentage in the dataset (accounting for 97.82%). The other two labels tend to be mistaken for the CLEAN label due to the effect of CLEAN was too high. The words that appear a lot in the CLEAN label, the classification model would default to their positive meaning, so when using these words in other meaningful sentences (hate, offensive), the model tend to categorize them into a positive sense. Moreover, HATE and OFFENSIVE labels show the relative same emotional state, which is confusing in the classification process, resulting in a low correct classification rate on these two labels. Through each fold, the classification results of the models change. In particular, the CNN model brings the most stable accuracy through each fold, so overall through five folds this model achieves the highest accuracy. CNN accomplish the highest accuracy classification rate on CLEAN at 99.42% and OFFENSIVE at 68.60%. The HATE label has the highest accuracy by using the LSTM model with 85.10%. Through each fold, LSTM shows superiority when classifying on HATE labels, so when ensemble, we give priority to the LSTM model to classify on label HATE. Therefore, the ensemble model helps to increase the accuracy on this label from 85.10 % to 85.39 %. Besides, the ensemble model makes the model's overall prediction rate increase.

7 Conclusion and Future Work

In this study, we propose a simple but effective ensemble method for social media text on the three different datasets: HSD-VLSP, UIT-VSFC, and UIT-VSMEC. These datasets are multi-label datasets with various labels depending on each dataset. Our ensemble model accomplish outstanding results higher than the deep neural network, as well as previous studies. The combination of many different models' strengths has achieved higher results than single models. In UIT-VSMEC, we achieve 65.79% of F1-score. With the UIT-VSFC dataset, our results are 92.79% in sentiment classification task and 89.70% in the topic classification task in terms of F1-score. Finally, in HSD-VLSP dataset, we reach 86.96% of the F1-score.

Besides, our single models also bring better results than previous studies. With single models, we find that, with small datasets, pre-processing does not bring expected results, the UIT-VSMEC dataset gain better performances when skipping the pre-processing step. On the contrary, with large datasets, pre-processing data has a good effect on our models, like data from the VLSP shared task. Also, with specific data domains, pre-trained word embeddings on the corresponding data domain will be more effective. Besides, improving word representation by dictionaries also helps to improve the efficiency of the models. Our dictionaries improve pre-trained embedding performance and increase 2% the effectiveness of our single models. This study indicate that the BERT model has been highly effective for Vietnamese language classification. However, with the use of pre-trained word embeddings training on Wikipedia data, it does not achieve the expected results when applying to the social media data like UIT-VSMEC and UIT-VSFC. With the datasets related to Vietnamese comments on social media, the CNN model has outperformed other single models. Choosing the number of layers is essential, with small datasets such as the UIT-VSMEC dataset, enhancing layers does not bring high efficiency but only makes the model more complicated. In contrast, the addition of layers reach excellent performance for large datasets such as the HSD-VLSP dataset.

For future works, it is necessary to improve the performance of the task, obtaining better results on the datasets, by studying other neural-based models. Besides, test each model with many different parameters to find more suitable parameters. It is implementing experiments of various data processing methods as well as improving coverage of word embeddings on the datasets to gain better results.

References

Allen, David M The Relationship between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, pages 125–127.

Battu Varshit, Batchu Vishal, Gangula Rama Rohit Reddy, Dakannagari Mohana Murali Krishna Reddy and Mamidi Radhika 2018. Predicting the Genre and Rating of a Movie Based on its Synopsis. *PACLIC 32. 56. 32nd Pacific Asia Conference on Language, Information and Computation.*

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*

Grave Edouard, Bojanowski Piotr, Gupta Prakhar, Joulin Armand, Mikolov Tomas 2018. Learning Word Vectors for 157 Languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Hang Thi-Thuy Do, Huy Duc Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen and Anh Gia-Tuan Nguyen 2019. Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model. *The Sixth International Workshop on Vietnamese Language and Speech Processing VLSP 2019.*

Hochreiter Sepp and Schmidhuber Jürgen 1997. Long short-term memory. *MIT Press, Neural computation*, 9, 8, pages 1735-1780.

Huu Quang Pham, Trung Son Nguyen and Pham Hoang Anh 2019. Automated Hate Speech Detection on Vietnamese Social Networks *The Sixth International Workshop on Vietnamese Language and Speech Processing VLSP 2019.*

Huyen Nguyen Thi Minh, Roussanaly Azim, Vinh Ho Tuong 2008. A hybrid approach to word segmentation of Vietnamese texts. *International Conference on Language and Automata Theory and Applications.*

K. V. Nguyen, Vu Duc Nguyen, P. X. V. Nguyen, T. T. T. Hong, N. L. Nguyen 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19-24.

Kim Yoon 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*, pages 1746-1751

Liu Jingzhou, Chang Wei-Cheng, Wu Yuexin and Yang Yiming 2017. Deep learning for extreme multi-label text classification. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115-124.

M. Schuster and K. K. Paliwa 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing.*

Medsker Larry, Jain Lakhmi C 1999. Recurrent neural networks: design and applications. *CRC press.*

Nikolov Alex and Radivchev Victor 2019. Nikolov-Radivchev at SemEval-2019 Task 6: Offensive Tweet Classification with BERT and Ensembles. *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691-695.

Nguyen Cam-Tu, Nguyen Trung-Kien, Phan Xuan-Hieu, Le Nguyen Minh, Ha Quang Thuy 2006. Vietnamese word segmentation with CRFs and SVMs: An investigation. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation.*

P. X. V. Nguyen, T. T. T. Hong, K. V. Nguyen, N. L. Nguyen 2018. Deep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus. *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 75-80.

Pham Thai-Hoang, Pham Xuan-Khoai, Nguyen Tuan-Anh, Le-Hong Phuong 2017. Nnvlp: A neural network-based vietnamese language processing toolkit. *Proceedings of The 8th International Joint Conference on Natural Language Processing.*

Son T Luu, Hung P Nguyen, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen 2020. Comparison Between Traditional Machine Learning Models And Neural Network Models For Vietnamese Hate Speech Detection. *IEEE RIVF 2020 Confererence*

Van Huynh, Tin, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen and Anh Gia-Tuan Nguyen 2020. Job prediction: From deep neural network models to applications. *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1-6.

Su Ming-Hsiang, Wu Chung-Hsien, Huang Kun-Yi and Hong and Qian-Bei 2018. LSTM-based text emotion recognition using semantic and emotional word vectors.. *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).*

Thai Binh Nguyen and Quang Minh Nguyen and Thu Hien Nguyen and Ngoc Phuong Pham and The Loc Nguyen and Quoc Truong Do 2019. VAIS Hate Speech Detection System: A Deep Learning based Approach for System Combination. *Association for Vietnamese Language and Speech Processing 2019*

V. D. Nguyen, K. V. Nguyen and N. L. Nguyen 2018. Variants of Long Short-Term Memory for Sentiment Analysis on Vietnamese Students' Feedback Corpus. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 306-311.

Van Thin Dang, Lac Si Le and Ngan Luu-Thuy Nguyen 2019. NLP@ UIT: Exploring Feature Engineer and Ensemble Model for Hate Speech Detection at VLSP 2019. *The Sixth International Workshop on Vietnamese Language and Speech Processing VLSP 2019.*

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen 2019. Emotion Recognition for Vietnamese Social Media Text. *PACLING 2019.*

Vu Xuan-Son, Vu Thanh, Tran Mai-Vu, Le-Cong Thanh, Nguyen Huyen T M 2019. HSD shared task in VLSP campaign 2019: Hate speech detection for social good. *Proceedings of VLSP 2019.*

Xuan-Son Vu 2016. Pre-trained Word2vec models for Vietnamese. *https://github.com/sonvx/word2vecVN.*

Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen 2020. UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network. *The 6th Workshop on Noisy User-generated Text (W-NUT) 2020.*

Text Mining of Evidence on Infants' Developmental Stages for Developmental Order Acquisition from Picture Book Reviews

Miho Kasamatsu[1] **Takehito Utsuro**[1] **Yu Saito**[2] **Yumiko Ishikawa**[3]
[1]Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, JAPAN
[2]Faculty of Child Studies, Seitoku University, Matsudo, 271-8555, JAPAN
[3]Cooperative Faculty of Education, Utsunomiya University, Utsunomiya, 321-8505, JAPAN

Abstract

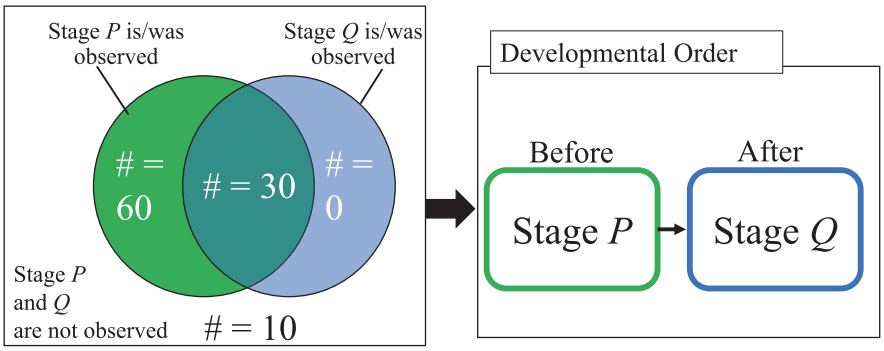
Herein, we study infants' reactions associated with the ordering of development mediated by picture books. We propose an approach that detects evidence on infants' reaction types from reviews of picture books. Existing developmental psychology research reports that infants' development occurs in a specific order; i.e., infants react according to their developmental stages. These existing findings must be extended because limited literature exists on studying picture books and infants' developmental order. However, the conventional approach in developmental psychology is time- and labor-consuming. Furthermore, existing findings show no information on the characteristics of picture books, making it harder to reproduce existing findings. To solve this problem, we collect infants' reactions by extracting the descriptions of infants' reactions from online reviews of picture books. Additionally, we investigate the characteristics of picture books that are related to these infants' developmental reactions. We collect over 150 reactions for each of the four types of reactions selected for the analysis and obtain typical characteristics of picture books for three reactions among the four types of selected reactions and find the proposed approach based on text mining of picture book reviews to be highly effective.

1 Introduction

Typically, educational books focus on specific subjects to be voluntarily learned such as science and sociology. However, picture books differ in the sense that because they are efficient in infants' cognitive development (Pardeck, 1986). They do not focus on specific educational subjects with their style of expressions, i.e., funny stories and pictures. Although the readers of picture books are parents or childcare personnel who read the books for infants who have yet to achieve literacy. The target audience comprises infants who perceive and interpret the incoming stimuli of the book being read to them and the pictures shown in the books. Thus, picture books differ from other educational books in the sense that those who read them are separated from those who perceive them.

Developmental psychology research states that infants exhibit various cognitive reactions to external stimuli based on their developmental stage (Sully, 2000; Piaget, 1962; Leslie, 1987; Walker-Andrews and Kahana-Kalman, 1999). Infants express such cognitive reactions when picture books act as the stimuli. Furthermore, considering that infants cannot understand the printed letters of picture books, this tendency of reacting to pictures might be amplified to a certain extent.

Herein, we investigate infants' reactions related to the ordering of development mediated by picture books. We propose an approach that detects evidence on infants' reaction types by applying the text mining technique. In this area of research, existing research on developmental psychology (Ishikawa and Maekawa, 1996) has reported that there is a specific order of infants' development, and infants react according to their developmental stages. Ishikawa and Maekawa (1996) focused on the ordering of infants' development mediated by picture books.



E.g., when the result of studying the reactions of 100 infants is as shown above, it can be interpreted as follows:
"When the stage Q is observed with an infant I, the stage P is always observed with the infant I "

The result in the left can be represented as above.

Figure 1: Overview of Ordering Analysis (Airasian and Bart, 1973)

Their findings include valuable insights regarding infants' developmental order, however, there are limitations to these findings: (i) no subsequent research exists based on infants' developmental order that is closely related to picture book reading, and (ii) existing findings include no information on the characteristics of picture books, making it harder to reproduce existing findings. Thus, extending these existing findings is necessary. However, the conventional approach in developmental psychology is quite time- and labor-consuming.

To investigate how the stimuli of picture books induce various reactions in infants, we apply the text mining technique to numerous reviews on picture books written by parents or childcare personnel. However, reviews on picture books have different characteristics compared with general book reviews. Such reviews include descriptions of an infant's reactions as well as descriptions of the reviewer's impressions of the book. Of these, descriptions of an infant's reactions are informative from the perspective of developmental psychology research. Thus, we analyze such descriptions extracted from picture book reviews. Furthermore, we focus on the infants' reactions related to the ordering of development mediated by picture books. We propose an approach that detects pieces of evidence of the types of infants' reactions by applying text mining. In addition, we investigate the characteristics of picture books that are closely related to those infants' developmental reactions. Specifically, we collect over 150 reactions for each of the four types of reactions selected for the analysis and obtain typical characteristics of picture books for three reactions out of the four types of selected reactions. We believe the proposed approach based on text mining of picture book reviews is highly effective.

2 Study of Infants' Developmental Order related to Picture Book Reading

In developmental psychology research, there is a specific order of infants' development, and infants show reactions in accordance with their developmental stages. Ishikawa and Maekawa (1996) focused on the relationship between the infants' developmental stages and the characteristics of picture books or the interaction induced by them. Ishikawa and Maekawa (1996) showed that there is a specific order of infants' development in the picture book readings and daily lives. For example, they indicated that before an infant learns how to turn pages and return to the proper page when he/she notices that he/she skipped a page, he/she who had not cared about skipping pages experiences preliminary stages such as "prefer picture books showing concrete daily life items and very few stories and sentences" and "prefer picture books having repetitive construction of pictures and sentences." To discover such types developmental orders, a questionnaire survey was conducted with 858 mothers having infants aged up to 74 months. In this survey, the participants were asked whether each of the 124 stages were observed

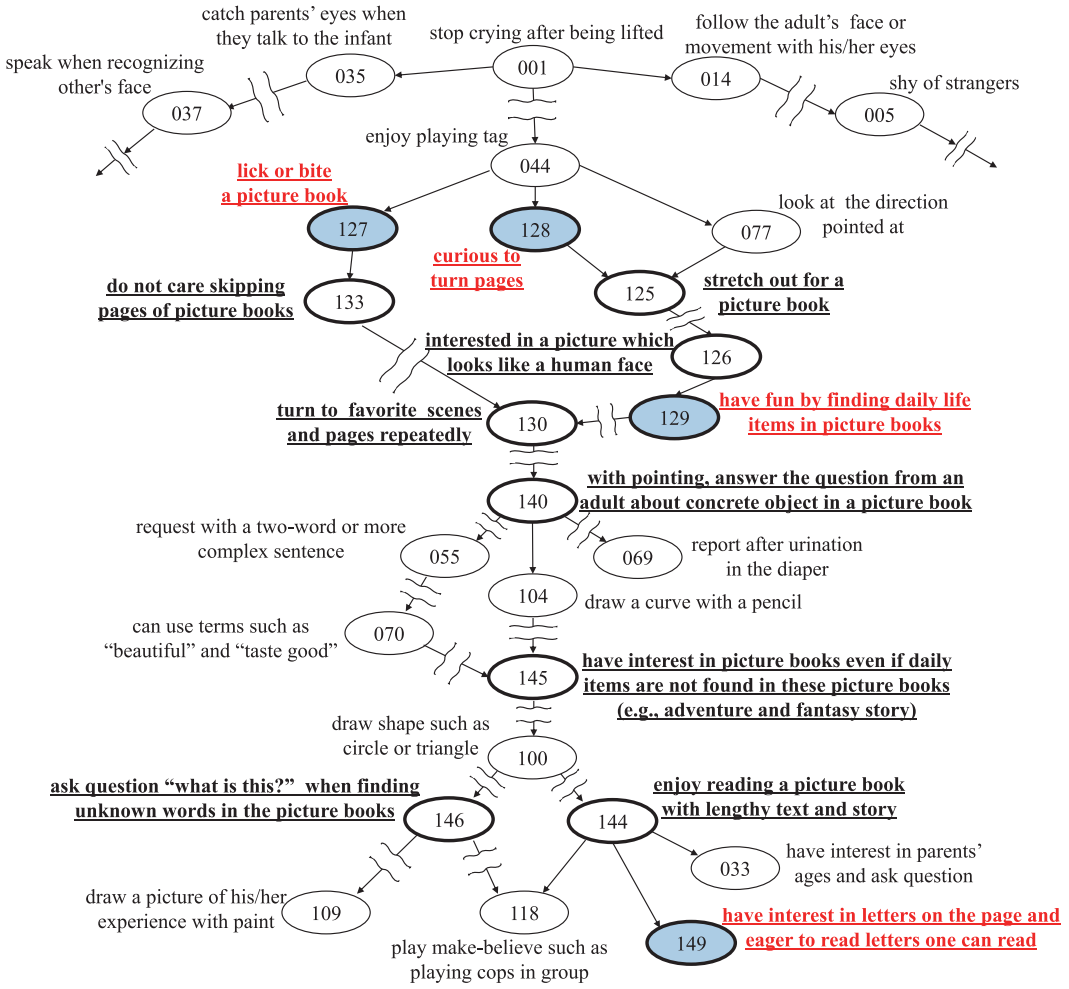


Figure 2: Excerpts of Developmental Stages and their Order (Ishikawa and Maekawa, 1996) (underlined stages are those related to picture book reading)

in their infants. The 124 stages include 25 stages regarding picture book reading and the remaining 99 stages concern the infants' daily lives such as "speak when you recognize someone's face." For each of the 124 stages, the participants select one of the following three answers: "is observed," "was observed," and "is not observed."

Ishikawa and Maekawa (1996) applied the ordering analysis technique (Airasian and Bart, 1973) to the result of this survey. They obtained the acquisition order of which stage comes first among those observed stages. Figure 1 shows the overall idea of the ordering analysis technique using the diagram notation and describes the situation wherein one stage P precedes another stage Q according to the numbers of infants observed throughout the survey. Figure 2 illustrates the excerpt of the result of the ordering analysis shown by Ishikawa and Maekawa (1996). The developmental stages are located as the earlier stages mentioned above and the later stages described below. The arrow in Figure 2 indicates that the stage at its endpoint follows the one at its starting point. Regarding picture book reading, 12 are underlined in Figure 2 among the 25 stages.

3 Purpose of this Paper

In general, the following limitations of developmental order exist in the work of Ishikawa and Maekawa (1996). (i) The scale of the developmental order proposed by Ishikawa and Maekawa (1996) is insufficient. (ii) Ishikawa and Maekawa (1996) included no information on which picture books were read

when infants showed specific reactions corresponding to a certain developmental stage.

To resolve (i), of the 124 developmental stages, clear developmental orders were detected in only 98 stages. For the remaining 26 stages, orders were not detected. It is unclear whether the orders detected in the 98 developmental stages discovered by Ishikawa and Maekawa (1996) are sufficient to exhaustively understand the orders among all developmental stages. Thus, the survey of Ishikawa and Maekawa (1996) needs to be reproduced and developmental stages that were not covered within their survey (Ishikawa and Maekawa, 1996) have to be included, and then an additional survey for discovering orders among newly detected developmental stages has to be further conducted. However, the conventional approach of the research of developmental psychology based on questionnaire surveys is time- and labor-consuming.

The issue of (ii) is another reason that renders the conventional approach of developmental psychology research time- and labor-consuming. To reproduce the survey of Ishikawa and Maekawa (1996), identifying which picture books were read when infants showed specific reactions corresponding to a certain developmental stage is necessary. For example, in the case of the developmental stages such as "prefer picture books that have concrete objects without stories and sentences" (developmental stage no. 135) and "enjoy reading a picture book with lengthy text and story" (developmental stage no. 144), preparing picture books that satisfy the requirements provided within the description of each developmental stage is necessary. However, Ishikawa and Maekawa (1996) included no information on which picture books were read in the results of the questionnaire survey.

Regarding the discussion above, this study proposes how to collect evidence on the infants' reactions related to the ordering of development mediated by picture books by applying the text mining technique to picture book reviews. To analyze the infants' reactions, text data of reviews on picture books were collected from EhonNavi,[1] a website specializing in picture books. For the typical developmental stages related to picture book reading

[1] `http://www.ehonnavi.net` (in Japanese)

(a) Principal Information

Start date of the service	Number of titles	Number of unique users per month	Number of members	Number of reviews
Apr. 2002	81,500	987,500	605,600	398,600

(b) Distribution of the Numbers of Reviews According to Infants' Age

Age of infants	0	1	2	3	4	5	6
Number of reviews	8,984	16,318	27,856	33,443	29,930	26,328	21,825

Table 1: Overview of EhonNavi

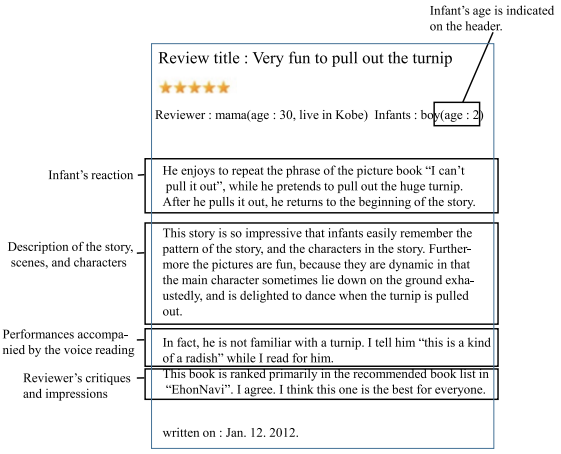


Figure 3: An Example of a Review

that were studied in Ishikawa and Maekawa (1996), we illustrate the procedure and the results of mining the evidence on the infants' reactions related to the ordering of development mediated by picture books from the collected reviews. Collecting the characteristics of picture books related to those infants' developmental reactions from online reviews on picture books is easy. This is an advantage of the proposed method based on text mining from online reviews on picture books over the conventional approach based on questionnaire surveys in developmental psychology research.

4 Website Specializing in Picture Books

To analyze the infants' reactions, text data of reviews (written in Japanese) on picture books were collected from EhonNavi, a website in Japan specializing in picture books. EhonNavi provides information concerning picture books such as their publishers, authors, and descriptions. It also provides numerous reviews written by the parents or childcare personnel. The number of picture books included in

EhonNavi is about 81,500. The number of reviews is approximately 398,600 as of June 2020 (shown in Table 1). There are other popular websites with numerous book reviews as well, such as Amazon[2] and Booklog.[3] Of these, EhonNavi has a unique characteristic in that its reviews are elaborate, thus reflecting the reactions of those who read the books as well as those who perceive them in detail. Additionally, another characteristic of EhonNavi showed that the age of the infant is attached to each review. All these characteristics are important for our work aimed at detecting the infants' reactions following their developmental stages. Therefore, we have used the reviews provided on EhonNavi for the analysis conducted herein. Figure 3 shows an example of a review on EhonNavi: the header of each review includes the age of the infant to whom the reviewer reads the picture book.

5 Mining Evidence on Infants' Developmental Stages related to Picture Book Reading from Reviews

5.1 Estimating Frequency Distribution of Infants' Reactions per Age

Of the 25 developmental stages regarding picture books studied in Ishikawa and Maekawa (1996), this study focuses on the four stages described in Table 2. They are frequently observed in picture book reviews collected from EhonNavi. These four stages are shown with blue circles and red letters in Figure 2. Of these four stages, stages 127, 128, and 129 are those observed for infants aged at early months, whereas stage 149 is the one observed for infants older than the former. For infants aged between 0 to 6, we estimated the frequency of the infants' reactions observed from the picture book reviews on EhonNavi. Figure 4 shows the corresponding distributions. Thus, we describe the procedure of collecting the reviews and semi-automatically estimating the frequency of infants' reactions in those collected reviews as follows:

Step 1. For each of the four stages, the row "search words" in Table 2 shows words used that we search for reviews which include descriptions by the reviewers satisfying the "manual criterion" of the corresponding stage. Here, we assume that a denotes the age group ranging from 0 to 6 years. In addition, we assume that q denotes a search word (e.g, "lick (舐め)") or more than one search word concatenated with "and" operator (e.g. "mouth and put (口 and 入れ)") listed at each subrow in the "search words" row of Table 2. Therefore, for each pair of the age a and the search word q, we collect reviews including those search words. We denote $h(a, q)$ as the number of collected reviews.

Step 2. For each of the four stages, we consider the row "manual criterion" of Table 2. We examine whether the reviews collected in step 1 satisfy the criterion. This means whether the collected reviews include descriptions of infants' reactions satisfying the condition of the corresponding developmental stage. Here, for each pair of age a and search word q, we denote $n(a, q)$[4] as the number of reviews randomly sampled for the procedure of manual examination, whereas $n_c(a, q)$ denotes the number of reviews that satisfy the criteria listed in Table 2.

Step 3. For each pair of age a and the search word q, we estimate the frequency $\hat{n}_c(a, q)$ of infants' reactions in those collected reviews as

$$\hat{n}_c(a, q) \quad = \quad h(a, q) \times (n_c(a, q) \,/\, n(a, q))$$

Finally, we illustrate the estimated frequency distribution in Figure 4. For each of the four stages, we sum up the estimated frequencies $\hat{n}_c(a, q)$ throughout all search words q as below and obtain the estimated frequency $\hat{n}_c(a)$ for each age a.

$$\hat{n}_c(a) \quad = \quad \sum_q \hat{n}_c(a, q)$$

For each of the four stages with the estimated frequency distribution per age, Figure 4 shows the periods in months wherein the pass rate[5] reported in Ishikawa and Maekawa (1996) is greater than or equals to 70% and is less than 80% (denoted as "70%–80%"), or greater than or equals to 80% (denoted as "≥80%"). From Figure 4, the periods in months where the pass rate of Ishikawa and Maekawa (1996) is greater than or equals to 80% mostly overlap with the peak in the estimated frequency distribution per age from the reviews. Regarding the peaks in the estimated frequency distribution per age, the estimated frequency is almost ≥90. The number of observations is sufficient when

[2] http://www.amazon.co.jp (in Japanese)

[3] http://booklog.jp (in Japanese)

[4] Herein, when $h(a, q) \geq 30$, $n(a, q)$ is fixed as 30 for all pairs of a and q. Otherwise, $n(a, q)$ equals to $h(a, q)$, i.e., all collected reviews are manually examined.

[5] In Ishikawa and Maekawa (1996), the pass rate is defined as the ratio of the answers "is observed" and "was observed" out of the overall answers in the questionnaire survey.

Stage No.	127	128	129	149
Stage description	lick or bite a picture book	curious to turn pages	have fun by finding daily life items in picture books	have interest in letters on the page and eager to read letters one can read
Search words (English translation of Japanese search words)	lick (舐め)	turn (めくる)	have fun (喜ぶ)	letter and read and not adult (文字 and 読む and not 大人)
	bite (噛)	turn and hand (めく and 手)	bus and have fun (バス and 喜)	letter and read (字 and 読)
	mouth and put (口 and 入れ)	turn and self (めくる and 自)	train and have fun (電車 and 喜)	letter and self (文字 and 自)
		turn and page (めくる and ページ)	dog and have fun (犬 and 喜)	letter and read (文字 and 読)
		flipping (ぺら)	cat and have fun (猫 and 喜)	letter and self and read (字 and 自 and 読)
		riffle (ぱら)		
Manual criterion for judging whether the case written in the retrieved review satisfies the condition of the corresponding developmental stage	An infant actually licks or bites a picture book.	An infant has an interest in contents other than the letters such as pictures, pop-up and lift-the-flap and tries to turn pages to find them. This criterion excludes cases where an infant turns pages in order to read letters in picture books by himself/herself.	An infant is pleased to discover correspondence between concrete objects in the real world and pictures in a picture book.	An infant reads and understands letters constituting sentences in a picture book.

Table 2: Words used in Searching for Picture Book Reviews including Cases Satisfying Each Developmental Stage

compared with the number of participants with the answers as "is observed" and "was observed" in the questionnaire survey of Ishikawa and Maekawa (1996).[6] Because the peaks of the frequency distribution per age estimated from the reviews and those of the distribution of the pass rates of Ishikawa and Maekawa (1996) are overlapping and the numbers of observations within the reviews are sufficient, then the picture book reviews are informative enough at least for the four stages studied herein.

5.2 Estimating Pass Rates per Age

Next, we estimate the pass rate for each age and four stages from the picture book reviews. Here, we collect reviews written after the picture book b is read for each pair of the picture book b and the age a. We denote $n_r(b, a)$ as the number of the collected reviews. Additionally, we denote $n(b, a)$[7] as the number of reviews randomly sampled for the procedure of manual examination. Here, $n_c(b, a)$ denotes the number of reviews that satisfy the criteria listed in Table 2. Then, for each pair of the picture book b and the age a, we estimate the frequency $\hat{n}_c(b, a)$ of infants' reactions in the collected reviews as

$$\hat{n}_c(b, a) \quad = \quad n_r(b, a) \times (n_c(b, a) \,/\, n(b, a))$$

Finally, we estimate the pass rate $pr(a)$ for age a for each of the four stages. We add the numbers $n_r(b, a)$ of the collected reviews and the estimated frequencies $\hat{n}_c(b, a)$ throughout all picture books b as shown below. Further, we obtain the estimated pass rate $pr(a)$ as their rate for each age a.

$$pr(a) \quad = \quad \sum_b \hat{n}_c(b, a) \,/\, \sum_b n_r(b, a)$$

Figure 4 shows the distribution of the estimated pass rates for each of the four stages. We observe that the pass rates estimated from the reviews are lower when compared to those reported in Ishikawa and Maekawa (1996). This is because not all parents described what they observed in the reviews. The fact is that only those parents who observed certain evidence of the developmental stages of their infants write the observed evidence in the reviews. The distributions of the pass rates estimated from the reviews coincide with the distributions of the pass rates reported in Ishikawa and Maekawa (1996).[8]

[6] The numbers of participants with the answers as "is observed" and "was observed" are 110 for stage 127, 88 for stage 128, 66 for stage 129, and 102 for stage 149.

[7] In this paper, when $n_r(b, a)$ is greater than or equals to 30, $n(b, a)$ is fixed as 30 for all the pairs of b and a. Otherwise, $n(b, a)$ equals to $n_r(b, a)$, i.e., all the collected reviews are manually examined.

[8] Of the four stages, the distribution of pass rates estimated from the reviews for stage 129 has higher pass rates in the higher ages from 4 to 6 years when compared with the distribution of the estimated frequency. This is because the estimated frequencies for those higher ages are small, which causes much higher pass rates.

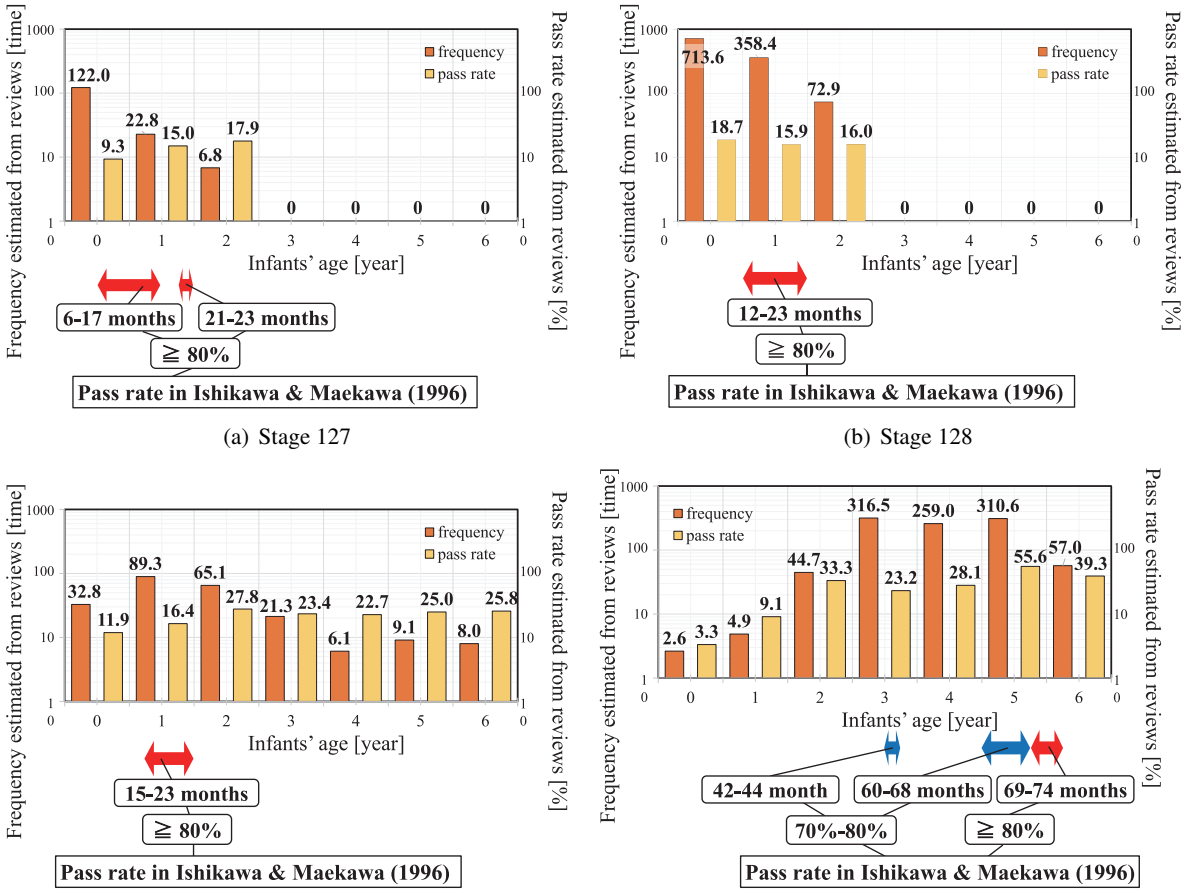


Figure 4: Frequencies and Pass Rates of Infants' Reactions Estimated from Reviews and Pass Rates in Ishikawa and Maekawa (1996)

5.3 Analyzing the Characteristics of Picture Books

The major advantage of the proposed approach based on text mining of picture book reviews is that it is easy to collect the characteristics of picture books that are read when infants show specific reactions corresponding to a developmental stage. Thus, this section studies the characteristics of picture books that are read when the reactions corresponding to stages 128, 129, and 149[9] are observed from the four stages. For each of the three stages, let B_0 denote the set of picture books b that satisfy the following condition:

> at least one review (written after the picture book b is read) is manually examined in step 2 of the previous section, and the review includes descriptions of infants' reactions satisfying the condition of the corresponding developmental stage.

Furthermore, we examine the characteristics of each picture book $b\,(\in B_0)$ such as "vehicles in the book," "animals in the book," "lift-the-flap book," "secure biding book," "colorful," and "featuring words" as shown in Figure 5. For each of the examined characteristics, we count the number of picture books satisfying the characteristics. Finally, we consider the characteristics that have the number greater than or equal to 5 picture books and illustrate them in Figure 5 with infants' age as the vertical axis. In addition, the three developmental stages 128, 129, and

[9]Infants' reactions that are specific to stage 127, i.e., "lick or bite a picture book," usually have no strong relation to any specific characteristics of picture books, but are observed when picture books of any type are read, when suitable for infants around the ages of 3 years or younger. Thus, we discard stage 127 for the analysis on the characteristics of picture books.

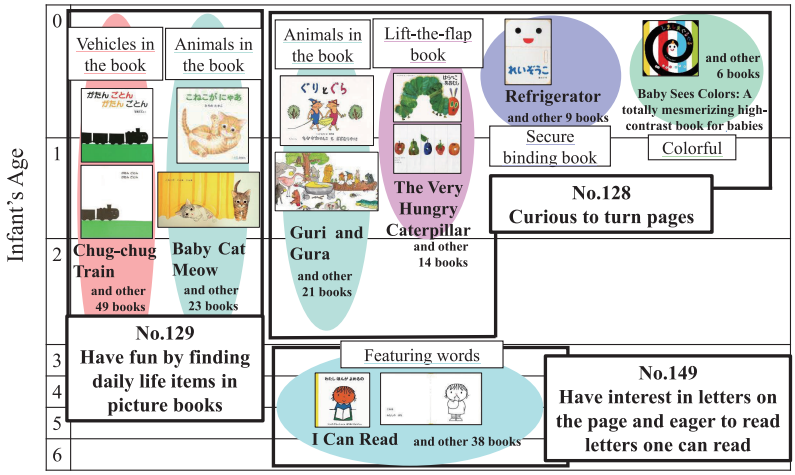


Figure 5: Typical Characteristics of Picture Books for Each Developmental Stage

149 illustrate the areas covering the corresponding characteristics. It is clear to understand the layout of characteristics shown in Figure 5, as the stage 128, "curious to turn pages," is observed in the picture books with characteristics such as "colorful," "secure binding book," and "lift-the-flap book." The stage 129, "have fun by finding daily life items in picture books," is observed in picture books with the characteristics such as "vehicles in the book" and "animals in the book." Further, the stage 149, "have interest in letters on the page and eager to read letters one can read," is observed in picture books with characteristics such as "featuring words."

6 Related Work

Previous work on analyzing infants' developmental reactions detected from reviews on picture book reading includes an earlier work on how to detect infants' major nine developmental reactions from reviews based on a simple keyword matching approach (Uehara et al., 2015). Further, later works considered on pointing behavior (Uehara et al., 2016) or hand/finger gestures (Uehara et al., 2017). Some focused on the relationship between characteristics of picture books and infants' developmental reactions (Baba et al., 2017a) and clustering of picture books (Baba et al., 2017b). The proposed approach is novel when compared to previous works on analyzing infants' developmental reactions detected from reviews on picture book reading. This is because we focused on the study of Ishikawa and Maekawa (1996) that concentrated on the ordering of infants' development mediated by picture books. The proposed approach shows how to detect evidence on infants' developmental stages related to the ordering of development mediated by picture books from reviews. Related work includes studies on picture book recommendations based on the similarity of storylines of picture books, infants' interest, and language developmental stages (Hattori et al., 2016; Yasuo et al., 2017). Their approach is based on the texts and pictures of picture books themselves but not on reviews of picture books.

7 Conclusion

In this work, we studied the infants' reactions related to the ordering of development mediated by picture books. We proposed an approach that detects evidence on the types of infants' reactions from reviews on picture books. We further examined the characteristics of picture books that are related to those infants' developmental reactions. Future work includes scaling up this study into the remaining 21 developmental stages studied in the work of Ishikawa and Maekawa (1996). For the remaining 21 stages, the proposed approach based on search words applies to a few of them. Furthermore, another approach based on characteristics of picture books (e.g., "picture book with a relatively long story") also applies to other types of stages.

References

P. W. Airasian and W. M. Bart. 1973. Ordering theory: A new and useful measurement model. *Educational Technology*, 13(5):56–60.

M. Baba, H. Uehara, M. Kasamatsu, T. Utsuro, and C. Zhao. 2017a. Analyzing characteristics of picture books based on an infant's developmental reactions in reviews on picture books. In *Proc. 3rd ALLDATA*, pages 57–62.

M. Baba, H. Uehara, and T. Utsuro. 2017b. Clustering picture books based on an infant's developmental reactions in reviews on picture books. In *Proc. 11th IMCOM*, pages 1–8.

T. Hattori, T. Kobayashi, S. Fujita, Y. Okumura, and K. Aoyama. 2016. Pitarie: Picture book search with interdisciplinary approach. *NTT Technical Review*, 14(7):1–8.

Y. Ishikawa and H. Maekawa. 1996. Comprehension of picture books and developmental order: A basic study on the use of picture books as an aid to child development. *Bulletin of Disability Sciences*, 20:83–91. (in Japanese).

A. M. Leslie. 1987. Pretense and representation: The origins of theory of mind. *Psychological Review*, 94(4):412–426.

J. Pardeck. 1986. *Books for Early Childhood: A Developmental Perspective*. Greenwood Pub. Group.

J. Piaget. 1962,. *Play, Dreams, and Imitation in Childhood*. W W Norton & Co Inc.

J. Sully. 2000. *Studies of Childhood*. Free Association Books.

H. Uehara, M. Baba, and T. Utsuro. 2015. Detecting an infant's developmental reactions in reviews on picture books. In *Proc. 29th PACLIC*, pages 64–71.

H. Uehara, M. Baba, and T. Utsuro. 2016. Extracting children's behavioral characteristics for acquiring language from texts of picture book reviews. *The International Journal of Networked and Distributed Computing*, 4(4):212–220.

H. Uehara, M. Baba, and T. Utsuro. 2017. Analyzing developmental characteristics of infants' finger/hand gestures — text analysis of picture book reviews —. In *Proc. 23rd Annual Meeting of the Association for Natural Language Processing*, pages 430–433.

A. S. Walker-Andrews and R. Kahana-Kalman. 1999. The understanding of pretence across the second year of life. *British Journal of Developmental Psychology*, 17(4):523–546.

M. Yasuo, M. Matsushita, T. Hattori, and S. Fujita. 2017. Measuring similarity of story lines for picture book search. In *Proc. 22nd TAAI*, pages 25–28.

Expressing the Opposite: Acoustic Cues of Thai Verbal Irony

Nimit Kumwapee
Department of Linguistics,
Faculty of Arts,
Chulalongkorn University
nimit.kumwapee@gmail.com

Sujinat Jitwiriyanont
Department of Linguistics and
Southeast Asian Linguistics Research Unit,
Faculty of Arts, Chulalongkorn University
sujinat.j@chula.ac.th

Abstract

The present study examined the acoustic cues of sarcasm in Thai. Four Thai speakers participated in two reading tasks: neutrality and sarcasm reading. Speech rate, F0 mean, F0 range, amplitude mean, amplitude range, F0 slope, and F0 intercept were measured and analyzed. The results indicated that sarcasm was produced at a faster speech rate and with a higher F0 mean, a wider F0 range, and a higher amplitude mean. Amplitude range reported no statistical significance. F0 slope alone was an insignificant cue, but F0 slope together with F0 intercept could distinguish between sarcasm and neutrality. Regarding gender differences, male speakers decreased their speech rate and increased their F0 mean while female speakers increased their F0 range when expressing sarcasm. Also, both male and female speakers increased their amplitude mean when producing sarcasm.

1 Introduction

Verbal irony is a linguistic device utilized to convey an opposite meaning from the literal meaning embedded in its linguistic form (Searle, 1991). To illustrate, the utterance "Your cooking is the best." means that the person to whom the utterance is addressed is literally the best cook while can also suggest that the person is the worst cook if the speaker intends to say otherwise. Additionally, irony is regarded as flouting the Maxim of Quality (Grice, 1989). In other words, the speaker of the utterance above was lying to the addressee that they are good at cooking. Sperber and Wilson (1981) propose another framework that irony is an echoic expression of thoughts that conveys a dissenting attitude such as skepticism, mockery, or contempt. Thus, the speaker of the utterance above could be mocking the person they address. Despite having more definitions than presented here, irony can be viewed as a form of communication expressing a different/opposite facet of meaning with different attitudes.

Different meanings within a statement can be expressed via different methods employed whether it be different word choice, syntactic structure or means of communication. That is, some words convey a more formal meaning than another and some structures emphasize on different pieces of information. Different tones of voice are used to signify different meanings apart from the original statement. During communication, meaning is not just encoded in linguistic forms. It is both linguistic and non-linguistic cues that play a role in an effective communication among interlocutors. Linguistic cues are encoded within sounds, words, and structures as previously explained. Non-linguistic cues, on the other hand, involve gestures, facial expressions, and the situations/contexts in which the communication takes place. Hellbernd and Sammler (2016) found that extralinguistic cues such as speech prosody also function in conveying intentions. They found that prosodic cues such as F0 rise, mean F0, mean intensity, and duration provided a foundation for listeners to recognize the intention of speakers. This shows that acoustic cues also help convey information that is not encoded within the linguistic form—intentions—which in turn helps decode the

intended meaning of an utterance. Bryant and Fox Tree (2002) similarly examined the role of contextual and prosodic information in the recognition of verbal irony and found that participants rated a sentence as more sarcastic[1] when provided with acoustic contents or irony-biasing contexts. This shows that both acoustic cues and contextual information work hand in hand during the process of inferring or decoding ironic intents. Hence, prosodic cues do not only function as local linguistic cues, but also function as pragma-linguistic cues that help in the signaling and interpreting processes of intents in communication. Speakers undoubtedly employ such cues when expressing verbal irony.

1.1 Acoustic Cues and Sarcasm

There were a wide range of studies regarding the acoustic characteristics of sarcasm. Each study examined different acoustic parameters of a different language, but speech rate, F0 mean, and F0 range appeared as variables in all studies. Majority of studies found that sarcasm was produced with a slower speech rate when compared to neutral speech regardless of language. However, F0 mean and F0 range appeared to be language-dependent and varied across studies. For example, sarcasm was expressed with a lower F0 mean in English (Cheang and Pell, 2008; Bryant, 2010; Chen and Boves, 2018), Spanish (Rao,2013) and Cantonese (Lan et al., 2019), but was produced with a higher F0 mean in French (Lœvenbruck et al., 2013) and Italian (Anolli et al., 2002). Although there were different findings concerning F0 mean in English and Cantonese (Rockwell, 2007; Cheang and Pell, 2009), such difference could be a result of different methodologies used in each study. Apart from speech rate, F0 mean, and F0 range, amplitude mean and range –previously found to be insignificant cues in English –were also found to be another significant cue for sarcasm in various works (Anolli et al., 2002; Cheang and Pell, 2009; Lan et al., 2019). To recap, these studies support the previous point that speakers employ different acoustic cues when expressing verbal irony and these cues are found to contrast with neutral speech and vary across language.

Apart from the differences in acoustic cues between speech types found, the studies by Rao (2013), Chen and Boves (2018), and Lan et al. (2019) also found the differences between sarcasm produced by male and female speakers. However, there were differences across language. To elaborate, Rao (2013) found that Mexican Spanish male speakers would significantly decrease their speech rate and suppressed their F0 range more than female speakers when producing sarcasm. Additionally, Chen and Boves (2018) discovered that British English male speakers showed larger durational difference than female speakers whereas female speakers would lower their mean pitch when expressing sarcasm. Lan et al. (2019) found a similar pattern to Chen and Boves (2018) in term of the durational difference; however, they also found that female speakers showed a larger F0 mean difference than male speakers and F0 range was found significantly smaller only in female speakers. Hence, speakers of different gender could employ acoustic cues differently when producing sarcasm and such differences are language dependent like the acoustic characteristics of sarcasm themselves.

1.2 Acoustic Study of Sarcasm in Thai

Majority of works that studied sarcasm/verbal irony in Thai concerned pragmatics (Panpothong, 1996; Kongchang, 2017; Bunnag, 2017) and stylistics (Anansapsuk, 2016). Nevertheless, no studies examined sarcasm in Thai from an acoustic perspective. An acoustic study of Thai most relevant to the study of verbal irony was carried out by Sonboonta (2010). Sonboonta (2010) studied the acoustic characteristics of the What-word /ʔarai/ in direct and indirect speech acts and found that despite the fact that no difference between speech acts could be established, F0 of a positive indirect speech act was found to be higher than that of a negative indirect speech, suggesting that there was difference in acoustic values of different speech acts to some extent. However, the study only examined the acoustic characteristics of a single word produced by female speakers. Consequently, an acoustic study of Thai sarcasm in the sentence level produced by both male and female speakers is needed so as to generalize the characteristics of Thai sarcasm and to provide a broader

[1] Majority of studies seem to use the term sarcasm in Searle's sense, and both are used interchangeably despite the distinctions. When sarcasm is used in this study, it refers to the general sense of verbal irony.

picture about acoustic cues of sarcasm from the perspective of the Thai language. It is also important to note that most of the languages studied are non-tonal with the only exception of Cantonese. An acoustic study of Thai sarcasm would also add to the existing literature the acoustic characteristics of sarcasm from the point of view of a tonal language such as Thai. The aims of this study are (1) to study the acoustic cues of Thai sarcasm in the sentence level and (2) to investigate whether there is a difference in acoustic cues of sarcasm between male and female speakers.

2 Methodology

2.1 Participants

Two male and two female Thai speakers participated in this study (M = 24 years old, SD = 0.71). All participants spoke Bangkok Thai fluently. They were picked from a self-selection process and voluntarily participated without being paid. The participants reported no speech hearing and production problems.

2.2 Materials

Thirty sentences and a biasing situation for each sentence were prepared for this study. Each sentence did not end with any final particles and was affirmative sentence with the number of words ranging from five to thirteen (M = 7.27 words/sentence, SD = 2.1). Two sets of materials were prepared: (1) Baseline Reading Set and (2) Biased Reading Set. The former was just the thirty baseline sentences without any context provided whereas the latter was a set of each baseline sentence preceded by its biasing context as presented in Table 1.

2.3 Data Collection

Two production tasks were conducted: baseline sentence reading and biased sentence reading.[2] The reading was self-recorded by each participant using a mobile application with a sampling rate between 41-48 kHz. The participants were instructed to record a sample of their voice and send back to check whether the voice level was not too loud or too low and whether they placed their cellphone too close or too far before the recording began. Afterwards, they received the Baseline Reading Set and they were instructed to read each sentence once with 2-3s pause between each. After they sent back their baseline recording, they received the Biased Reading Set and were instructed to read the situation context silently before producing the bolded and underlined target sentence. Total tokens (4 participants x 2 readings x 30 sentences) acquired were 240 tokens.

[2]While some of the previous studies used a perception test to select only the productions that sound sarcastic and some did not, the current study did not use a perception test because it focused on production, the signalling of sarcasm through acoustic cues.

2.4 Acoustic Value Extraction and Calculation

Acoustic measures such as utterance duration, maximum pitch, minimum pitch, mean pitch, pitch listing, maximum intensity, minimum intensity/amplitude, and mean intensity/amplitude were extracted manually for each token using Praat. At this stage, mean F0 (Hz) and mean amplitude (dB) were automatically measured without any further calculation needed. Afterwards, speech rate (word/s) was calculated by dividing the syllable number of a sentence by the utterance duration. F0/amplitude range was calculated by subtracting the maximum value by the minimum value. F0 values gained from the pitch listing function in Praat were converted into a data point set. The data point set was then used to calculate a slope (Hz/ms) and its intercept (Hz) using the Least Square Regression method. The extraction and calculation proceeded as above for all 240 tokens.

2.5 Acoustic Analyses

An average of each acoustic cue was calculated for each participant, for each gender, and for all participants. All data were converted into z-scores. The acoustic characteristics of sarcasm and neutrality between participants, between gender, and for overall participants were then observed from the z-scores.

2.6 Statistical Analyses

Raw acoustic values stored as a data table with information about gender and speech type were imported into R to perform a statistical analysis. ANOVA test for each acoustic variable considering speech type as a factor was then carried out to test the difference in acoustic values between types of speech whereas Paired t-test for each acoustic cue of each speech

Biasing context	Baseline sentence
You went out with a group of friends and then they invited a person you did not like to join as well. After finishing, your friend asked you how it was. You ironically replied:	/c^{h}ăn miː k^{h}waːm sùk mâːk/ “I am very happy.”
You assigned an employee to arrange a set of documents. Although there were a few documents, he took half a day to finish. You then said:	/k^{h}ăw t^{h}am ŋaːn rew ciŋ ciŋ/ “He works really fast.”

Table 1: Examples of Baseline Sentences together with Their Biasing Contexts.

type was conducted to test whether there was a difference between gender. The alpha level used in this study was 0.05.

3 Results

3.1 General Acoustic Cues between Neutrality and Sarcasm

Figure 1 showed that the values of speech rate, F0 mean, F0 range, amplitude mean, and amplitude range for sarcasm was higher than that of neutrality. However, the statistical significance was found only in the case of speech rate, F0 mean, F0 range, and amplitude mean. To elaborate, sarcasm is significantly produced at a faster speech rate than neutrality, $F(1,238) = 13.45$, $p = .0003$. When expressing sarcasm, speakers increased their F0 mean significantly, $F(1,238) = 5.308$, $p = .0221$, and exhibited a significant wider F0 range than when they produced neutral speech, $F(1, 238) = 6.176$, $p = .0136$. As for amplitude, sarcasm was significantly produced with a higher amplitude mean, $F(1,238) = 6.985$, $p = .0088$, but there is no significant difference found in amplitude range, $F(1,238) = 0.016$, $p = .901$. This suggested that when Thai speakers produced sarcasm, they did so by increasing the speech rate, F0 mean, F0 range, and amplitude mean. Although the values of amplitude range appeared to be higher in sarcastic speech, amplitude range might not be a reliable cue when it came to distinguish sarcasm and neutrality.

Figure 2 illustrated the relationship between F0 slope (x-axis) and F0 intercept (y-axis) obtained from the Least Square Regression method. Sarcasm seemed to stay in a higher region for both male and female speakers (See Figure 2). 65.83 % of F0 slope

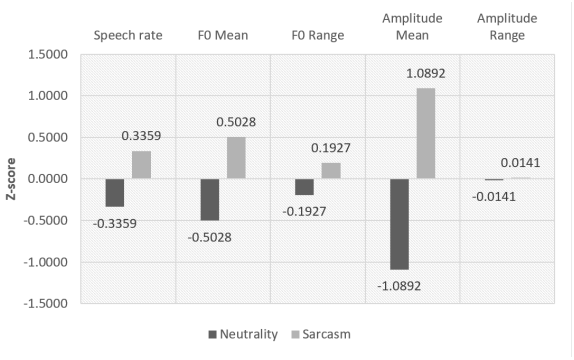


Figure 1: Five Acoustic Values (Z-Scores) of Two Speech Types.

in sarcastic speech were less than their original neutral slopes, suggesting that speakers employed different F0 contours for sarcasm. 83.33 % of F0 intercept in sarcasm were higher than their neutral counterparts, supporting that sarcasm was produced with a higher F0 mean than neutral speech. However, there was no significant difference in F0 slope between speech types. Still, there was a significant difference in F0 intercept between neutrality and sarcasm, $F(1,238) = 10.58$, $p = .0013$. Additionally, there was a significant difference in F0 slope together with F0 intercept between neutral and sarcastic speech, $F(1,238) = 10.55$, $p < .01$. The interaction between F0 slope and F0 intercept was also found to be statistically significant, $F(1,238) = 6.85$, $p < .01$. This signified that the two values taken together as a parameter could distinguish between sarcasm and neutrality. Moreover, sarcasm appeared to exhibit more variability than neutrality. To illustrate, F0 slope of sarcasm showed more variability (M= -0.3122, SD= 0.4963) than that of neutrality (M= -

0.2134, *SD*= 0.377). Likewise, F0 intercept of sarcasm also showed more variability (*M*= 207.6355, *SD*= 66.9692) than that of neutrality (*M*= 180.5076, *SD*= 61.568). It could be that speakers manipulated their F0 when they produced a sarcastic sentence more than when they made a neutral sentence.

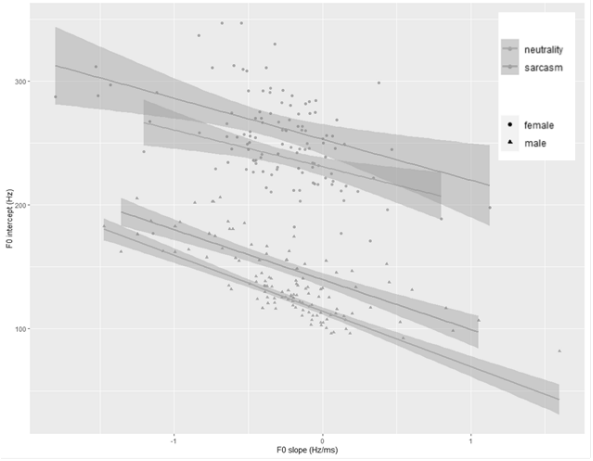


Figure 2: Relationship between F0 Slope and F0 Intercept by Speech Type and Gender with Linear Fitted Smooths (Upper lines of each pair represent sarcasm) and Confidence Bands.

3.2 Acoustic Cues between Sarcasm and Neutrality by Participants and Speech Types

As for speech rate (See Figure 3 Panel A), although M1 speaker seemed to significantly emphasize his sarcastic reading, resulting in a slower speech rate for sarcasm than neutrality, the overall speech rate of sarcasm was still faster than neutral speech. With regard to gender, male speakers produced a sarcastic speech with a significantly slower rate than neutral speech, $t(1,118) = 2.4092$, $p = .0175$. On the other hand, female speakers produced a sarcastic speech with a significantly faster rate than neutral speech, $t(1,118) = -8.9696$, $p < .0001$. This illustrated that there was gender difference despite the general characteristics of speech rate for overall participants.

Regarding F0 Mean, male speakers seemed to significantly increase their F0 mean in sarcastic speech, $t(1,118) = -9.8553$, $p < .0001$. However, there was no significant difference in F0 mean between sarcasm and neutrality within female speakers, $t(1,118) = -1.7472$, $p = .0832$. This was due to F2 speaker whose F0 mean did not differ much between sarcasm and neutrality (See Figure 3 Panel B).

For F0 range, there was no significant difference in F0 range between sarcasm and neutrality in male speakers, $t(1,118) = -0.6338$, $p = .5274$, possibly because M2 speaker produced sarcasm with a narrower F0 range (See Figure 3 Panel C). However, female speakers significantly exhibited a wider F0 range ($t(1,118) = -3.2322$, $p = .0016$) when producing sarcasm as can be seen in both F1 and F2 speakers.

For amplitude mean (See Figure 3 Panel D), both male and female speakers increased their amplitude mean when expressing sarcastic speech, $t(1,118) = -2.0723$, $p = .0404$, and $t(1,118) = -2.6534$, $p = .0091$, respectively. However, M2 speaker showed a different pattern from the rest of the participants. Despite this difference, the overall characteristic of amplitude means for overall speakers, male speakers, and female speakers was still statistically significant.

For amplitude range, the difference between male and female speakers could not be generalized because there were both cases that amplitude range was wider (M1, F1) and narrower (M2, F2) for sarcasm within both group (See Figure 3 Panel E).

4 Conclusion and Discussion

This study examined speech rate, F0 mean, F0 range, amplitude mean and amplitude range along with F0 slope and F0 intercept of sarcasm and neutrality. Speech rate, F0 mean, F0 range, and amplitude mean were found to be a significant cue that differentiated between speech types. Speech rate for sarcasm in Thai is different from other languages because it is expressed at a faster speech rate (See Table 1). However, one speaker appeared to emphasize and stress each word significantly, resulting in a slower speech rate. The difference in speech rate might be a result from different emotions expressed in different contexts such as in the study by Tumtavitikul and Thitikannara (2006) that showed that there was a difference in duration for speech produced with different emotions. The current work was in line with Yimngam et al. (2011) that neutral speech was produced at the slowest speech rate from other types of emotions. This study did not control situational or emotional contexts because it

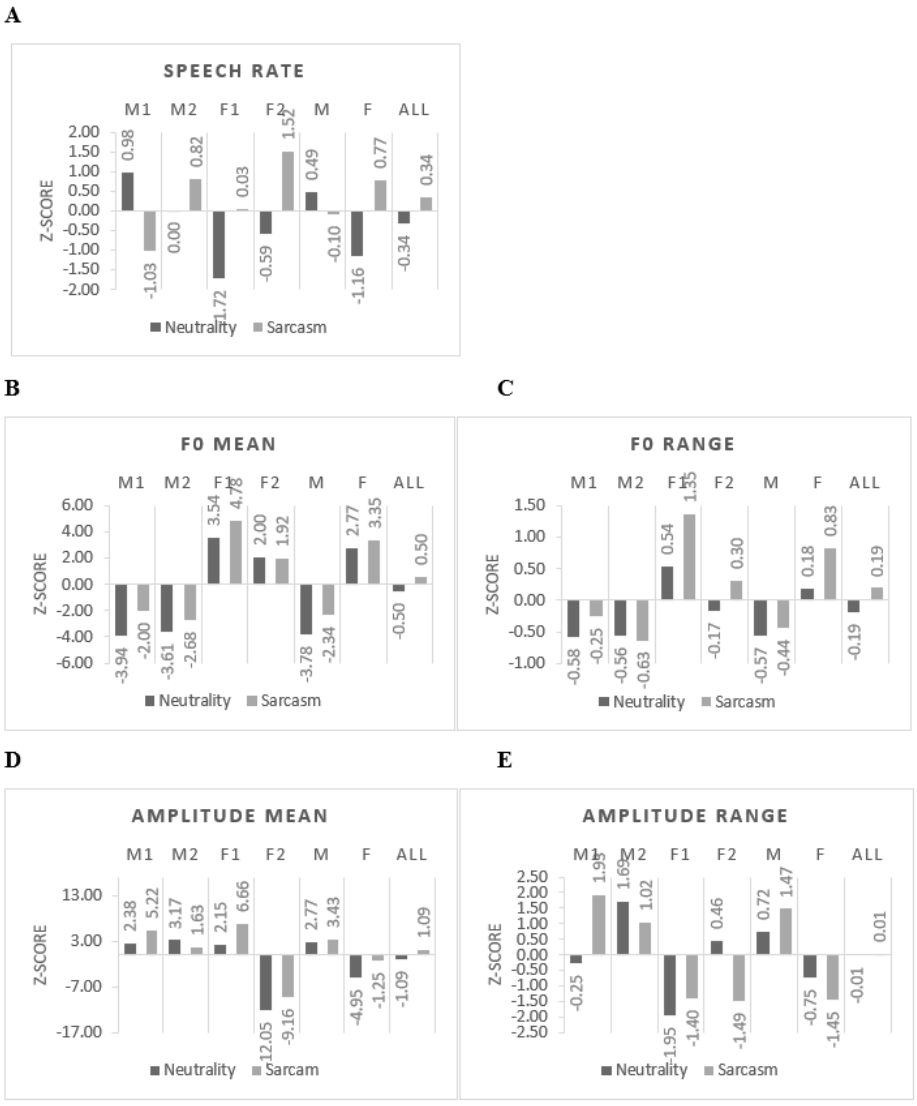


Figure 3: Acoustic values of Sarcasm and Neutrality across Participants and Genders and for Overall.

aimed to explore the general characteristics of sarcastic speech. The variation in sarcasm in term of types and emotions awaits further study.

For F0, the result in F0 mean that sarcasm is produced with a higher pitch corresponds with studies in French (Lœvenbruck et al., 2013) and Italian (Anolli et al., 2002). Surprisingly, the result of F0 mean in this study is different from Cantonese (Lan et al., 2019), which is also a tonal language. Regarding F0 range, that sarcasm is expressed with a wider F0 range in this study is accordance with studies of French (Lœvenbruck et al., 2013), Italian (Anolli et al., 2002), and Cantonese (Cheang and pell, 2009; Lan et al., 2019). Regarding amplitude, that sarcasm is expressed with a higher amplitude is in line with the studies in Italian (Anolli et al., 2002) and Cantonese (Cheang and pell, 2009; Lan et al., 2019). Although sarcasm seems to be expressed with a wider amplitude range as in Italian (Anolli, et al., 2002) and Cantonese (Lan et al., 2019), there is no significant difference in amplitude range found in the current study.

As for F0 slope, there is no significant difference between speech types. However, F0 slope together with F0 intercept distinguishes between neutrality and sarcasm. Sarcasm seems to exhibit a larger vari-

ability in F0 slope than neutral speech. The variability might be due to different emotional contexts in which sarcasm is expressed. Study of F0 contour by Gu and Lee (2007) showed that different emotional speech displayed different sentential F0 declination. Additionally, different emotions affect the pattern of F0 contour or tone as in Li (2015). Hence, the variability could arise from the emotional differences which in turn result in different F0 contour patterns. Even though there was no significant difference found for F0 slope, F0 movement and contour are found to be specific to different emotions as in the works by Paeschke and Sendlmeier (2000) and by Paeschke (2004). Moreover, that there was no significance found is possibly due to the nature of a tonal language that F0 movement of a sentence could not be so different from its original neutral sentence as Wu (2019) found that unnaturalness in F0 affected perceivability. As some manipulations of F0 seem to exist, future works could explore whether there is a difference in intonation patterns of different speech types or whether sarcasm affects the intonation patterns in Thai.

This study also examined gender as a factor and found that within the general acoustic characteristics found for sarcasm, there are gender differences across different acoustic cues. That male speakers produce sarcasm with a slower speech rate is in line with the studies by Rao (2013), Chen and Boves (2018), and Lan et al. (2019), but female speakers shows more durational difference between sarcasm and neutrality than male speakers in this study. Additionally, this study found that male speakers increase their F0 mean significantly when expressing sarcasm whereas female speakers exhibit a wider F0 range. Likewise, both genders increase their amplitude mean significantly when making a sarcastic sentence. Nevertheless, this study found that male speakers rely not only on decreasing their speech rate as found in Rao (2013), Chen and Boves (2018) and Lan et al. (2019), but also on increasing their F0 mean. Whereas Lan et al. (2019) found that F0 range was significantly smaller only in female speakers, this study found a different pattern that F0 range was significantly wider only in female speakers. These similarities and differences support that male and female speakers use different acoustic cues for sarcasm in different languages and gender should also be included as one of the variables so as to find out if there is to be difference within the overall acoustic characteristics.

In this study, sarcasm was found to be expressed with a faster speech rate, higher F0 mean, wider F0 range, and higher amplitude mean. The gender difference was also observed across acoustic cues of sarcasm. Also, this study found that there are variations within sarcasm in term of emotional and situation contexts that awaits future works. Future works may explore various sarcastic sentence patterns apart from an affirmative type and consider different types of sarcasm expressed with different emotions such as anger or joy or in different settings such as friendly or unfriendly so as to provide a more comprehensive picture about the variations and characteristics of sarcastic speech in Thai.

References

Anansapsuk, Ornkanya. (2016). Verbal irony in Udom Taepanich's 7-10 episode stand-up comedy. Master thesis, M.A (Thai). Bangkok: Graduate School, Chulalongkorn University.

Anolli, L., Ciceri, R., and Infantino, M. G. (2002). From "blame by praise" to "praise by blame" : Analysis of vocal patterns in ironic communication. *International Journal of Psychology*, 37(5): 266-276.

Boersma, P., and Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved from https://www.praat.org/

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, 47(7): 545–566.

Bryant, G. A., and Fox Tree, J. E. (2002). Recognizing Verbal Irony in Spontaneous Speech. *Metaphor and Symbol*, 99-117.

Bunnag, Orawee. (2017). The functions of verbal irony for commenting via Facebook of mobile operator in Thailand. *Vacana Journal*, 5(2): 24-40.

Cheang, H. S., and Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, 50(5): 366-381.

Cheang, H. S., and Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, 126(3): 1394-1405.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, Mass.: Harvard University Press.

Gu, W., and Lee, T. (2007). Quantitative analysis of F0 contours of emotional speech of Mandarin. In *SSW*, 228-233.

Hellbernd, N., and Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act

perception. *Journal of Memory and Language*, 88: 70-86.

Kongchang, Rumrada. (2017). Responding strategies to verbal irony in Thai: a case study of speakers of equal status. Master thesis, M.A (Thai). Bangkok: Graduate School, Chulalongkorn University.

Lan, Chen, Hui, Pak, Xu, Wenwei and Mok, Peggy. (2019). Revisiting Acoustic Markers of Sarcasm in Cantonese. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)*, 77-81.

Li, A. (2015). Emotional Intonation and Its Boundary Tones in Chinese. In *Encoding and Decoding of Emotional Speech*, 133-164. Springer, Berlin, Heidelberg.

Lœvenbruck, H., Jannet, M. A. B., d'Imperio, M., Spini, M., and Champagne-Lavau, M. (2013). Prosodic cues of sarcastic speech in French: slower, higher, wider. Institut Universitaire de France, (August), 3537-3541.

Paeschke, A. (2004). Global trend of fundamental frequency in emotional speech. In *Speech Prosody 2004*, International Conference.

Paeschke, A., and Sendlmeier, W. F. (2000). Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.*

Panpothong, Natthaporn. (1996). A pragmatic study of verbal irony in Thai. Doctoral dissertation, University of Hawai'i.

R Development Core Team (2008). R: A language and environment for statistical computing. [Computer program]. Version 4.0.0, retrieved from http://www.R-project.org.

Rockwell, P. (2007). Vocal features of conversational sarcasm. *Journal of Psycholinguistic Research*, 29(5): 361-369.

Searle, John R. (1991). Metaphor. In Steven Davis (ed.), *Pragmatics: A reader*, 519–39. Oxford and New York: Oxford University Press.

Sonboonta, Wanlapapaporn. (2010). An Acoustic Analysis of Thai Wh-word [?arai] in direct speech acts and indirect speech acts. Master thesis, M.Ed (Educational Linguistics). Bangkok: Graduate School, Srinakharinwirot University

Sperber, Dan, and Wilson, Dierdre. (1981). Irony and the Use-Mention Distinction. In Peter Cole (ed.), *Radical Pragmatics*, 295-318.

Tumtavitikul, A., and Thitikannara, K. (2006). The intonation of Thai emotional speech. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, New Zealand, (December), 6-8.

Wu, M. (2019). Effect of F0 contour on perception of Mandarin Chinese speech against masking. PloS one, 14(1).

Yimngam, S., Premchaisawadi, W., and Kreesuradej, W. (2011). Prosody analysis of Thai emotion utterances. In *International Conference on Application of Natural Language to Information Systems*, (June), 177-184. Springer, Berlin, Heidelberg.

Identifying Authors Based on Stylometric measures of Vietnamese texts

Ho Ngoc Lam
Ho Chi Minh City University of Education
ngoclam0706@gmail.com

Vo Diep Nhu
VNUHCM-University of Science
vodiepnhu@gmail.com

Dinh Dien
VNUHCM-University of Science
ddien@fit.hcmus.edu.vn

Nguyen Tuyet Nhung
VNUHCM-University of Social Sciences and Humanities
velvetsnow.nguyen@gmail.com

Abstract

Author identification has many applications in investigating or resolving authorship disputes. Research on author identification has been conducted in many high resource languages, such as English, Chinese, Spanish, etc. However, for Vietnamese, studies are limited because of the lack of relevant language resources. This paper represents the topic of author identification with the application of stylometric methods: Mendenhall's characteristic curve, Kilgariff's squared method (Kilgariff's Chi-Squared), the Delta method of John Burrows. The study applied three different methods based on a corpus extracted from Vietnamese online newspapers, categorized by each author and achieved results from 50% to 100% depending on the method and number of linguistic features.

1 Introduction

International integration, along with the exponential growth of the Internet, has led to an increase in plagiarism, imitation of celebrities' writing style, and copyright disputes.

Due to the enormous amount of information, looking for the style and characteristics of written works in order to identify the author's style is a huge challenge. Globally, there have been numerous studies which find out models to identify the author's style in many languages. However, there are very few studies in natural language processing applying writing style in Vietnamese to attribute authorship.

Stylometry, beginning with attempts to settle authorship disputes, was first developed by Augustus De Morgan in 1851 based on word length. By the late 1880s, Thomas C. Mendenhall had analyzed the word length distribution for works written by Bacon, Marlowe, and Shakespeare to determine the true author of plays supposedly written by Shakespeare. In 1932, George Kingsley Zipf discovered the connection between ranking and the frequency of words, later stated in Zipf's law. In 1944, George Yule created a way to measure frequency of words, used to analyze vocabulary richness, namely Yule's characteristic. In the early 1960s, most research papers refer to Mosteller and Wallace's works on the Federalist Papers, which was considered as a basis of using computation in stylometry. In the next several decades, with the increasing number of digital texts, as well as the growth of the Internet, machine learning techniques, and neural networks, accessing information led to the development of natural language processing tools. Semantics continued to grow in the 21st century, and due to the overwhelming amount of information, copying texts also became more popular, leading to the growth of stylometry which is used in plagiarism detection, author identitication, author profiling, etc.

In this paper, we use a corpus of Vietnamese online texts to attribute authorship using the following measures: Mendenhall's characteristic curves, Kilgariff's Chi-Squared, John Burrows's Delta measure.

2 Related work

Broadly, there are two categories of stylometry:

Adversarial Stylometry: When translated, a piece of writing has its style imitated, and going through many translators makes its characteristics less distinct. These changes make detecting the original style more difficult.

Detecting stylistic similarities includes the following tasks:

Stylochronometry: In time, an author may change his/her writing style due to changes in vocabulary, lifestyle, environment, age, etc. Studies have sharp distinction because they depend on a language in a specific time period and on a particular author. Author Profiling: extracting the characteristics of a text to gain information about an author such as gender, age, region, time of writing.

Authorship Verification: Based on characteristics readily available in the training data, determining whether two texts were written by the same author. Authorship Attribution: an individual or group of authors has characteristic styles that are developed subconsciously. Based on these distinctions, we will identify the true author(s) of texts in a corpus.

3 Experimentation

In authorship identification using corpus-based approach, we use the NLTK Python package to process the corpus in order to execute the methods of author attribution. Due to limitations in the number of the texts per author, we will choose only 10 authors whose texts contain appropriate number of sentences and words and closely similar in size. Depending on methods and characteristic numbers, our results vary between 50% and 100%.

3.1 Corpus

We use a corpus of Vietnamese online texts, including 1304 texts extracted from several Vietnamese online newspapers (largely from VnExpress), Facebook, and blogs. These texts are written by 10 authors, who give their own opinion or share their own experiences on social issues. The corpus was pre-processed to eliminate links, images, captions, and tokenized semiautomatically. The process of tokenization was carried out with CLC toolkit, an automatic tool developed by Computational Linguistics Center (VNUHCM-University of Science). Then we manually checked the whole corpus and correct the mistakes. The number of texts and tokens of each author are displayed in Table 1 below.

No.	Authors	Corpus			Test set
		Texts	Tokens (Include punctuations)	Tokens (Exclude punctuations)	Tokens (Exclude punctuations)
1	Author59	162	118,148	101,375	43,287
2	Author83	45	84,010	72,613	14,525
3	Author88	151	110,930	96,404	38,316
4	Author97	91	99,910	85,781	27,693
5	Author203	108	94,452	79,747	21,659
6	Author1028	152	87,341	75,029	16,941
7	Author1035	184	122,028	101,938	43,850
8	Author1050	133	120,337	102,653	44,565
9	Author1262	121	173,865	148,085	89,997
10	Author1289	157	118,267	102,940	44,852
Total		**1304**	**1,129,288**	**966,565**	**385,685**

Table 1. Information of the corpus and test set

3.2 Stylometric measures

Measure 1: Mendenhall's characteristic curves

Mendenhall once wrote that an author's "stylistic signature" could be found by measuring the frequency with which he or she used words of different lengths. These characteristic curves give results quickly and visually, allowing the researcher to draw a conclusion on the author's style. Applying this method, our group worked on our dataset of works by ten chosen authors. To standardize the size of the text while applying this method, we made the token number in works from each author's bibliography 58,088 token (punctuations removed). On each author's bibliography, we sequentially did the following: calculating the length of each token, calculating the frequency of calculated length in the bibliography, and visualize the data. Besides the visualized data, we use Caroll's index R to measure each author's lexical diversity to have an overview of style:

$$\boldsymbol{R = \frac{V}{N}}$$

V: vocabulary size (number of word types)

N: text size (number of word tokens)

Figure 1. Equation for lexical diversity

Measure 2: Kilgariff's Chi-squared

In the dataset whose authors are known, namely

Known: let denote the file of i^{th} candidate author K_i (i = 1, 2, … 10)

Let denote the unknown author's file U.

Calculate Chi-squared for each of the ten candidate authors.

1. First, build a joint corpus J, including K_i and U, and identify the 500 most frequent words in it.

2. Calculate the proportion of the joint corpus made up of the candidate author's tokens (AuShare).

AuShare = len(token K_i)/len(token Jcorpus)

3. Look at the 500 most common words in the candidate author's corpus and compare the number of times they can be observed to what would be expected if the author's file and the disputed file were both random samples from the same distribution.

4. Calculate how often we really see each of the 500 most common words, cw[x] (x = 1, 2, … 500), in K_i and U respectively with:

- Kcw_ob: observed number of cw in K_i
- Kcw_ob: observed number of cw in U

5. Calculate how should we see each cw in K_i and U respectively with:

- Kcw_ex: expected number of cw in K_i
- Ucw_ex: expected number of cw in U

6. Calculate a chi-squared distance of K_i and U:

$$\chi^2 = \chi^2_{K_i} + \chi^2_U$$

Respectively calculate chi-squared of K_i and U:

$$\chi^2_{K_i} = \sum_x \frac{(Kcw\_ob - Kcw\_ex)^2}{Kcw\_ex}$$

$$\chi^2_U = \sum_x \frac{(Ucw\_ob - Ucw\_ex)^2}{Ucw\_ex}$$

Figure 2. Equations for the chi-squared statistic of K_i and U.

The smaller the chi-squared value, the more similar the two corpora. Therefore, we will calculate a chi-squared for the difference between each file of the candidate author dataset Known and disputed file U; the smaller value will indicate which of Known is the most similar to U.

Measure 3: John Burrows' Delta measure

The Delta measure, proposed by John F. Burrows as a tool to solve the problem of copyright, measures the difference between two sets of text.

1. Combine all files in Known into a single corpus and get n frequency distribution words (test in n=20, n=30 respectively)

2. Calculating n[y] (y =1, 2, …, n) presence for each subcorpus K_i.

3. Calculating n[y] means (μy) and standard deviations (σy).

4. Calculating z-scores:

$$\boldsymbol{Z_i = \frac{C_i - \mu_i}{\sigma_i}}$$

C_i: the observed frequency

μ_i: means

σ_i: standard deviation

Figure 3. z-scores calculate the z-score in the test set.

5. Calculating features and z-scores for our test file

6. Calculating Delta

Find Delta point to compare the test set with each author.

$$\Delta_c = \sum_i \frac{|\boldsymbol{Z_{c(i)} - Z_{t(i)}}|}{\boldsymbol{n}}$$

$Z_{c(i)}$: z-score for feature i in subcorpus 'c'

$Z_{t(i)}$: z-score for feature i in the test set

Figure 4. Delta measure

3.3 Results

Measure 1: Mendenhall's characteristic curves

The results are shown in Figure 5. We observe that each author has the following features: Author59's longest word contains 17 characters, while that of Author203 and Author1050 only has 14.

Every author uses words having between 2 and 4 characters the most. The most prevalent word has 3 characters.

Author1035 and Author1262 yield different results from the other authors. Each of them uses 3letter words the most, followed by 4-letter words instead of 2-letter words like the other eight authors. The authors' lexical diversity: When examined with the same 58,088 tokens, the author having the highest lexical diversity (Caroll index R) is Author1035 with 0.146 whereas the one with the lowest diversity is Author83 with 0.092. The results are shown in Table 2.

No.	Authors	Lexemes	Vocabulary richness (R)
1	Author59	5961	0.103
2	Author83	**5327**	**0.092**
3	Author88	6781	0.117
4	Author97	6922	0.119
5	Author203	8206	0.141
6	Author1028	6467	0.111
7	Author1035	**8456**	**0.146**
8	Author1050	6136	0.106
9	Author1262	7498	0.129
10	Author1289	6259	0.108

Table 2. Vocabulary Richness

STT	Authors / Authors	Author59	Author83	Author88	Author97	Author203	Author1028	Author1035	Author1050	Author1262	Author1289
1	Author59	**2,549.649**	6,775.924	7,995.894	7,241.717	4,740.681	5,447.395	8,319.614	9,016.777	23,040.670	14,850.658
2	Author83	9,678.276	**2,359.664**	8,415.828	6,905.708	4,433.277	4,812.310	9,282.713	9,058.835	23,650.604	13,255.750
3	Author88	8,473.703	6,471.436	**2,154.291**	7,169.004	3,825.872	5,402.131	7,798.770	7,443.396	14,637.402	8,640.629
4	Author97	9,013.078	7,526.163	10,346.465	**1,874.816**	4,372.070	4,435.556	7,266.010	10,843.907	25,638.256	17,720.348
5	Author203	6,527.751	5,967.048	4,146.237	5,678.605	**2,267.813**	3,991.429	5,421.921	5,989.636	16,631.327	9,929.698
6	Author1028	10,234.820	8,515.822	11,195.689	7,291.693	6,462.341	**3,348.786**	8,587.273	12,500.521	23,682.286	17,239.991
7	Author1035	7,622.525	8,220.965	9,796.556	6,694.397	4,728.346	4,963.033	**3,366.667**	9,957.941	21,031.091	16,496.578
8	Author1050	9,416.742	6,611.444	6,650.636	8,935.540	4,336.624	6,179.540	9,129.678	**1,787.953**	17,136.267	10,778.297
9	Author1262	16,449.314	10,470.761	9,555.598	13,669.770	7,137.432	9,292.752	12,779.942	13,189.267	**2,850.582**	14,699.759
10	Author1289	10,584.509	6,665.983	5,901.024	8,819.146	5,094.171	6,564.798	10,761.940	8,324.811	19,483.718	**3,339.960**

Table 3. Chi-squared results

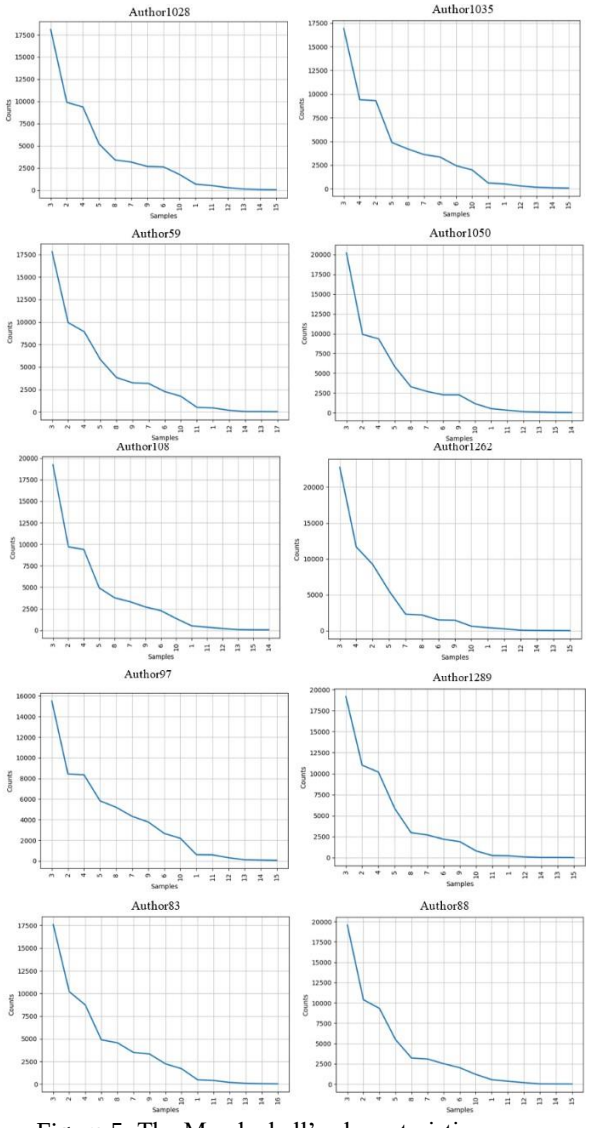


Figure 5. The Mendenhall's characteristic curves

Measure 2: Kilgariff's Chi-squared

Grieve (2007) assumed that the lower the chisquared measure between two texts, the more likely that they were written by the same author. Therefore, the known text giving the smallest Chi-squared value would be written by the author most likely to have written the unknown text. Table 3 below shows the Chi-squared results when we tested the text sample for each of the authors.

Measure 3: John Burrows' Delta

Table 4 and Table 5 display the results of Delta measure when we tested on each of the authors' text (according to the headings). Examined by rows, the smaller the Delta value is, the closer to the author's style the test work is.

After we tested on 30 signatures, the result yields 40%, matching the prediction on 4 out of 10 authors: Author59, Author97, Author1262, Author1289. The results are shown in Table 4.

The 30 signatures include: 'là', 'không', 'và', 'của', 'có', 'một', 'người', 'tôi', 'những', 'cho', 'được', 'các', 'thì', 'trong', 'với', 'đó', 'đã', 'cũng', 'để', 'phải', 'mà', 'ở', 'như', 'khi', 'này', 'mình', 'đến', 'về', 'sẽ', 'đi'.

Authors / Authors	Author 59	Author 83	Author 88	Author 97	Author 203	Author 1028	Author 1035	Author 1050	Author 1262	Author 1289
Author59	***0.003***	***0.005***	0.059	0.017	0.013	***0.005***	***0.005***	0.048	0.095	0.039
Author83	0.014	0.014	0.050	0.027	0.004	0.007	0.018	0.036	0.086	0.027
Author88	0.055	0.057	0.007	0.070	0.040	0.050	0.059	***0.007***	0.043	0.016
Author97	0.019	0.018	0.081	***0.004***	0.035	0.024	0.017	0.066	0.116	0.058
Author203	0.019	0.021	0.043	0.034	0.004	0.014	0.023	0.031	0.078	0.022
Author1028	0.018	0.017	0.049	0.031	***0.003***	0.011	0.022	0.035	0.085	0.026
Author1035	0.042	0.042	0.103	0.029	0.057	0.049	0.039	0.091	0.139	0.082
Author1050	0.061	0.061	0.003	0.074	0.046	0.054	0.065	0.011	0.039	0.020
Author1262	0.062	0.064	***0.002***	0.076	0.046	0.057	0.063	0.014	***0.037***	0.023
Author1289	0.027	0.027	0.037	0.040	0.012	0.020	0.030	0.023	0.072	***0.014***

Table 4. Experimental results of 30 most frequent lexemes

After we tested on 20 signatures, the result yields 50%, matching the prediction on 5 out of 10 authors: Author83, Author203, Author1035, Author1262, Author1289. The results are shown in Table 5. The 20 signatures include: 'là', 'không', 'và', 'của', 'có', 'một', 'người', 'tôi', 'những', 'cho', 'được', 'các', 'thì', 'trong', 'với', 'đó', 'đã', 'cũng', 'để', 'phải'.

Authors / Authors	Author 59	Author 83	Author 88	Author 97	Author 203	Author 1028	Author 1035	Author 1050	Author 1262	Author 1289
Author59	0.039	0.053	0.042	0.033	0.064	***0.004***	0.081	0.042	0.096	0.068
Author83	***0.012***	***0.002***	0.087	0.081	0.115	0.047	0.132	0.091	0.147	0.020
Author88	0.061	0.067	0.022	0.016	0.050	0.029	0.067	0.026	0.082	0.088
Author97	0.076	0.091	0.004	0.007	0.026	0.042	0.044	0.004	0.059	0.106
Author203	0.098	0.112	0.023	0.029	***0.007***	0.063	0.022	0.019	0.038	0.127
Author1028	0.078	0.093	***0.004***	0.009	0.024	0.044	0.041	***0.003***	0.057	0.108
Author1035	0.134	0.149	0.059	0.065	0.032	0.100	***0.015***	0.056	0.002	0.164
Author1050	0.076	0.090	0.005	***0.007***	0.027	0.041	0.044	0.004	0.060	0.105
Author1262	0.137	0.151	0.061	0.068	0.034	0.102	0.017	0.058	***0.002***	0.166
Author1289	0.013	0.007	0.090	0.081	0.115	0.047	0.132	0.091	0.147	***0.017***

Table 5. Experimental results of 20 most frequent lexemes

4 Discussion

Among the three measures mentioned above, Delta measure does not yield good results as we expected.

In Chi-square statistic, we convert everything to lowercase so that we won't count word tokens that begin with a capital letter because they appear at the beginning of a sentence and lowercased tokens of the same word as two different words. Sometimes this may cause a few errors, for example when a proper noun and a common noun are written the same way except for capitalization, but usually it increases accuracy. In addition, Chi-squared is a coarse method. For one thing, words that appear very frequently tend to carry a disproportionate amount of weight in the final calculation. Sometimes this is fine; other times, subtle differences in style represented by the ways in which authors use more unusual words will go unnoticed. (Laramée, 2018).

The algorithm based on taking the number of the most common words (words with highest frequency) in the corpus as a feature. In the VnExpress corpus, we get texts from the "Perspective" section, which offers a wide variety of topics, such as finance, society, lifestyle, health, etc. Not all authors write about the same topics, and relativity among topics leads to an inconsistency of topics in the corpus.

Even though we have processed on the sets with the same token number of each author, the disparity in topics may be the reason why the chosen features are biased towards certain authors, rather than representing the whole corpus.

4.1 Conclusion

Research in authorship identification in Vietnamese text is uncommon despite its high applicability in many fields. In fact, researchers face difficulties in finding a corpus with sufficient size and information about authors.

In this paper, we have presented three different measures of authorship identification; these are the basic methods of determining an author's style such as lexical diversity, number of characters in a word, and word frequency (to find the most frequent words). The Chi-squared measure yields 100% accuracy; whereas Burrows' Delta measure yields 40% accuracy with 30 features, and 50% accuracy with 20 features.

In future research, we will the examining on a corpus with a wide variety of topics to increase lexical variety. At the same time, we will prepare a richer annotated corpus so as to work on authorship identification using machine learning.

References

Alex I. Valencia Valencia, Helena Gomez Adorno, Christopher Stephens Rhodes & Gibran Fuentes Pineda. 2019. *Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-grams Model*. Notebook for PAN at CLEF.

Andrea Bacciu, Massimo La Morgia, Eugenio Nerio Nemmi & Valerio Neri. 2019. *Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features*. Notebook for PAN at CLEF.

Antonio Pascucci, Vincenzo Masucci & Johanna Monti. 2019. *Computational Stylometry and Machine Learning for Gender and Age Detection in Cyberbullying Texts*. IEEE.

Carmen Klaussner & Carl Vogel. 2015. *Stylochronometry: Timeline Prediction in Stylometric Analysis*. Springer International Publishing, Switzerland.

Divjak, D. 2019. *Frequency in Language Memory, Attention and Learning*. Cambridge University Press.

Eder, M. 2015. *Rolling stylometry*. Oxford University Press on behalf of EADH.

Hoover, D.L. 2004. *Testing Burrows's Delta. Literary and Linguistic Computing*. 19(4):453-475.

Imene Bensalem, Paolo Rosso & Salim Chikhi. 2014. *Intrinsic Plagiarism Detection using N-gram Classes.* Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Cornell University, arXiv: 1810.04805v2.

Kyung-Ah Sohn, Alemu Molla Kebede & Kaleab Getaneh Tefrie. 2014. *Anonymous Author Similarity Identification*. IEEE Symposium on Security and Privacy.

K. Surendran, O. P. Harilal, P. Hrudya, Prabaharan Poornachandran & N. K. Suchetha. 2017. *Stylometry Detection Using Deep Learning*. Springer Nature Singapore Pte Ltd.

Love, H. 2002. *Attributing authorship: An introduction.* Cambridge University Press, pp. 133.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming & F. J. Smith. 2003. *Extension of Zipf's Law to Word and Character N-grams for English and Chinese*. The Association for Computational Linguistics and Chinese Language Processing, 77-102.

Le Thanh Nguyen & Dinh Dien. 2019. *English-Vietnamese Cross-Language Paraphrase Identification Method.* Springer.

Le Thanh Nguyen, Nguyen Xuan Toan & Dinh Dien. 2016. *Vietnamese plagiarism detection method*. University of Florida, ACM, 44–51. https://doi.org/10.1145/3011077.3011109.

Mahmoud Khonji & Youssef Iraqi. 2017. *De-anonymizing Authors of Electronic Texts: A Survey on Electronic Text Stylometry*. Preprints.

Mendenhall, T. C. 1887. *The Characteristic Curves of Composition. Science*, 9(214): 237-249.

Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt & Damon McCoy. 2014. *Doppelgänger Finder: Taking Stylometry To The Underground.* IEEE Symposium on Security and Privacy.

Shaina Ashraf, Hafiz Rizwan Iqbal & Rao Muhammad Adeel Nawab. 2016. *Cross-Genre Author Profile Prediction Using Stylometry-Based Approach.* Notebook for PAN at CLEF.

Shuiyuan Yu, Chunshan Xu & Haitao Liu. 2018. *Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation*. Cornell University, arXiv:1807.01855.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang & Damon Woodard. 2017. *Surveying stylometry techniques and applications.* University of Florida, ACM Comput, Surv. 50, 6, Article 86. https://doi.org/10.1145/3132039.

Zipf, G.K. 1968. *The psycho-biology of language: an introduction to dynamic philology*. M.I.T. Press.

Marking Trustworthiness with Near Synonyms: A Corpus-based Study of "Renwei" and "Yiwei" in Chinese

Bei Li **Chu-Ren Huang** **Si Chen**

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Hong Kong SAR

benita-bei.li@connect.polyu.hk churen.huang@polyu.edu.hk sarah.chen@polyu.edu.hk

Abstract

We conducted a corpus-based study on near synonymous cognitive verbs "renwei 認為" and "yiwei 以為" with a similar meaning of "to think" in Chinese. The motivation of this study is that information trustworthiness in propositions varies with the use of "renwei" or "yiwei", which concerns the issue of speakers' stances or attitudes. We examined the epistemic modality attributed to "renwei" and "yiwei" in written discourse, by testing their different functions of evidential marking, negative forms and hedging device. Results showed that "renwei" was associated with specialized and professional subjects and consequently demonstrated its relatively high degree of evidentiality. Then, specific negative polarity shifters exclusively collocating with "yiwei" denoted larger scope of negation of entire propositions. Lastly, only "renwei" functioned as a hedge to mitigate claims lacking full commitment in the political discourse. Therefore, the evidence of data proved that the epistemic marker "renwei" was associated with a higher degree of information trustworthiness, contributing to the annotation system in NLP for the detection of information credibility.

Keywords: information trustworthiness; near-synonyms; epistemic modality

1 Introduction

Epistemic modality is an important topic on semantics, functioning in judging and evaluating the propositional content with regard to the degree of certainty, possibility or necessity in our communication. Speakers or writers use epistemic markers, like cognitive verbs (e.g. "think"/"believe") or modal auxiliaries (e.g. "may"/"could"), to express their attitudes or stances-taking toward the propositions. Recently the linguistically decoded information involving cognition has been an important topic to be faced in natural language processing (NLP). For example, in the statement "I thought he always supported me", the speaker's intention is to emphasize that "he" does not support the speaker now through the use of the past tense of the epistemic verb "think". Consequently, the trustworthiness of information seems complicated to decode in NLP with regard to the speaker's attitudes and stances. The present study investigates one type of epistemic markers, the cognitive verbs in Chinese, to analyze the credibility of information epistemic verbs attributed using a corpus-based approach.

Moreover, the use of near synonymous epistemic verbs makes it more complicated for text detection or information retrieval in Chinese. In the present study, the cognitive verbs "renwei" (認為) and "yiwei" (以為) will be comparatively analyzed to detect the information trustworthiness contained in the statements they constituted respectively. The pair of near synonyms "renwei" and "yiwei" has a similar meaning of "to think" with the function of marking speakers' epistemic modality. Specifically, "ren4" (認) is a verb referring to the act of identification or recognition, and "wei2" (為) is a verb denoting the action or the state of being (apply to both "renwei" 認為 and "yiwei" 以為). The preposition "yi3" (以) of "yiwei" means "with" or "by". The grammatical constructions consisting of "renwei" or "yiwei" are similar to the epistemic parenthetical "(I) think" in English. The subject and a complement-taking

psych verb (e.g. “renwei” or “yiwei”) form the matrix clause which induces the coming complement clause.

What motivates the investigation on this pair of near synonyms is the variations in conversational implicatures raised by these two epistemic markers in a Chinese context. In sentence (1), the mood and attitude toward the complemented content are neutral to a certain extent, whereas sentence (2) illustrates the negation of complementizer regarding the stance-taking of the addresser. As mentioned above, sentence (2) highlights the fact that “he” does not support “me” now.

(1) 他認為社會福利問題關係重大。
“He **thinks** that social welfare issues are of great importance.”
(2) 我以為他始終支持我的立場。
“I **thought** he always supported me.”

Therefore the present study aims at examining the information trustworthiness in propositions attributed to cognitive verbs. The functions of evidential marking, negative forms and hedging device in epistemic modality are three strategies adopted to test the differences between the two epistemic markers “renwei” and “yiwei”.

Evidentiality: As a sub-category of epistemic modality, the essence of evidentiality is the way that speakers inherently encode the information sources of a clause (Chafe, 1986). There was no agreement upon the category of evidentiality, as several scholars regarded evidentiality as a grammatical structure (e.g. Willett, 1988; Palmer, 2001), while some others expanded into any evidential expressions referring to speakers’ judgments by linguistic coding of the validity of information (e.g. Chafe, 1986; Su et al., 2010). Following the latter broad sense, studies on evidentiality dealt with the epistemological status of information or evidence at addressers’ disposal (De Haan, 2005; Nuyts, 2006).

In speech communication, speakers often use evidential markers to emphasize the explicit source of information, such as “It’s said…”, “People say…”, and “according to…”, in order to improve the credibility of information to a certain degree. According to Rubin et al. (2006), the lexical verb “think” together with “seem" and “sound” were grouped into the moderate degree of certainty as evidential markers. However, we need to explore whether Chinese lexical verbs “renwei” and “yiwei” have the function of evidential markers, and if any, what factors may affect the trustworthiness they expressed.

Negative forms: Negation is a complicated issue in natural language with the involvement of logic, semantics and pragmatics. (Horn, 1989, Speranza and Horn, 2010). A typical form of negation is realized by negative function words like “not”. For example, propositions like “Sam didn’t smile” are contradictory with “Sam smiled” (Israel, 2004). Besides, negative statements could also be triggered by negative polarity shifters (NPIs), in which the scale-reversing context licenses without explicit negation (Penka and Zeijlstra, 2010). For instance, in the statement “The car failed to climb the hill”, the NPI “fail” reverses the direction of the proposition.

A series of studies on negation and polarity by Horn and his colleagues (e.g. Horn, 1972; 1989; 2001; 2009; Horn and Kato, 2000) investigated the variety of negation scopes and the sophisticated entailment inferences of propositions designated by NPIs and their licensed context. Horn’s polarity theory about negative strengthening illustrated that both litotes and affixal negation were stronger than the non-conventionalized strengthening. That is, the proposition (***p’***) “She is not happy today” has a stronger negative attitude than the proposition (¬ ***p***) “It is false that she is happy today”. The proposition (***p’’***) “She is unhappy today” also entailed a stronger negation than ¬ ***p***.

However, whether the negative forms of Chinese epistemic near synonymous verbs have diversities in their implicatures and the truth of propositions is still remained to be explored. What’s more, the topic of polarity shifters is still challenging in Chinese NLP tasks due to the sheer number as well as their invalidation of automatic approaches (Xu and Huang, 2015).

Hedges: Hedging devices as a rhetorical strategy can be realized by particular lexical items, specific grammatical structures or prosodic variations of utterances in conversation. Speakers use hedges to refer to the fuzziness, vagueness, indirectness or

approximation of information they are conveying. In a broad sense, the term hedge is related to all kinds of linguistic means expressing the lack of full commitment (Fraser, 2010).

In previous literature on hedging devices, phrases like "I think" "I guess" and "I wonder" in English were grouped into the type of plausibility shields (Markannen and Schröder, 2010), and the current study followed this categorization. Pragmatic purposes were one important function of hedges which mainly involved the mitigation of claims, showing politeness to listeners, avoiding the criticism of prediction and etc. (Lakoff, 1975; Hyland, 1996; Brown and Levinson, 1987; Taweel et al., 2011).

Additionally, studies on epistemic parentheses as hedges involved a variety of discourses such as juristic judgments (KOŹBIAŁ, 2020) and political speech (Taweel et al., 2011). However, the majority of literature concerned with the conversations in spoken language, little study has paid attention to the genre of newspapers especially on specific discourse by a corpus-based approach.

Therefore, the purpose of the current study is to test the information trustworthiness between the pair of epistemic markers "renwei" and "yiwei" in written discourse. It was first conducted on the perspectives of evidentiality and negative forms to differentiate the credibility of epistemic modality for the near synonyms. Further, the study analyzed the near synonyms in political discourse by corpus to examine the rhetorical strategy of hedging devices for "renwei" and "yiwei". Consequently, the research questions of the current study are:

(1) Whether the information trustworthiness is differently contained in statements attributed to the markers "renwei" and "yiwei" in Chinese newspapers?
(2) If so, which one is more trustworthy, and which one is more opinionated?
(3) What are their differences in various strategies of epistemic modality?

2 Methods

2.1 The Chinese Word Sketch Approach

The methodology adopted was a corpus-based approach to the comparative analysis of epistemic markers "renwei" and "yiwei" using two datasets. The first one was Chinese Gigaword2 Corpus including gigaword_xin (xin), gigaword_cna(cna) and gigaword_zbn(zbn), collected from the newspapers of Mainland China, Taiwan, and Singapore respectively. Specifically, gigaword_xin (xin) included the texts of journalism formed by more than 200 million words from the Xinhua News Agency of Beijing from the period 1991 to 2002. For gigaword_cna (cna), more than 380 million words constructed the source data of journalism from the Central News Agency of Taiwan from 1990 to 2002. The newspapers of Lianhe Zaobao of Singapore contributed to the data of gigaword_zbn (zbn).

The tool used to process this dataset was the online system Chinese Word Sketch (available at http://wordsketch.ling.sinica.edu.tw/cws/). Three functions of the online system were utilized for the current study. First, "Concordance" provided a new query of the word or phrase and displayed the overview of entries containing keywords. This function also played a role in setting specific "left context" and "right context" to the keywords for the search of surroundings.

Then, the function of "Word Sketch Difference" demonstrated both the similarities and differences of two lexical items in terms of their grammatical structures and collocation patterns. By this approach, two keywords could be input simultaneously to obtain the "Common Patterns" and "Only Patterns" collocations for the pair of near synonyms.

Lastly, "Word Sketch" was the third function involved for detailed descriptions of grammatical relations and collocation patterns for a specific lexical item. The values of frequency and salience with respect to specific collocations contributed most to the following findings.

2.2 The Sinica Corpus Approach

Sinica Corpus (Academia Sinica Balanced Corpus of Modern Chinese) provided the dataset of Chinese newspapers on politics for the study. Data were processed by the online tool Sinica Corpus (version 4.0 available http://asbc.iis.sinica.edu.tw/) developed by Chen and Huang (2016) on the basis of the dataset in Chen et al. (1996). This dataset collected a total of 1,396,133 sentences, 11,245,330 word tokens, extracted from Chinese text of various articles in Taiwan from 1981 to 2007.

The topics of Sinica Corpus involved philosophy (8%), science(8%), society(38%), art(5%), life(28%) and literature(13%). Moreover, it functioned in narrowing the scope of the query in terms of the genres of a data source. Currently, Sinica Corpus was adopted to contrast the functions of epistemic markers in hedging devices (Section 3.4) for the genre of newspapers on political topics.

To sum up, this study utilized two approaches, the Chinese Word Sketch and Sinica Corpus, to compare the grammatical features as well as the collocation patterns of the pair of near synonyms "renwei" and "yiwei". The next section of results and analysis concerns the information trustworthiness tested by the evidentiality, negative forms and hedging device in epistemic modality.

3 Results and Analysis

3.1 Overall distributions of "renwei" and "yiwei"

For the dataset we adopted here, there were a total of 745051 entries found for the keyword "renwei" and 22286 entries for "yiwei". However, the relatively less frequent distribution for the lexical item "yiwei" still involved a number of invalid data.

Because there are two types of combinations consisting of the two Chinese characters 以 "yi3" and 為 "wei2"/"wei4" but have different meanings from "to think". One combination is the polyphony 為 ("wei4" here) followed by 以 "yi3" constructing the phrase "以為 yiwei" with the meaning of "to do something for…". The other kind of combination is the ellipsis of the phrase "以之為 yi3 zhi1 wei2" which means "to regard something as…". While the current study only considered the lexical item 以為 "yiwei" which possesses a similar syntactic meaning of "to think" with 認為 "renwei", and excluded the other two formations.

Obviously, "renwei" and "yiwei" showed dramatic differences in the use of frequency in the written discourse. Despite that they shared similar semantic meanings, a tendency to use in written discourse was demonstrated for "renwei" instead of "yiwei". The following sub-sections will analyze the factors leading to this tendency and effects on their information trustworthiness.

3.2 Differences in Evidentiality

	Collocations Frequency		Collocations Salience	
	renwei	yiwei	renwei	yiwei
Subject	228268	4433	21.9	14
他	37895	337	62.2	27.8
專家	10238	12	53.5	4.5
一般人	212	98	35.1	51.1
我	5676	295	51.0	41.6
人士	9004	28	44.6	7.6
他們	6488	127	43.0	23.8
她	3808	111	40.4	25.9
我們	2848	99	36.6	25.1
人們	647	122	27.9	36.5
大家	944	151	28.8	36.4
學者	2132	6	36.4	3.8
筆者	75	20	31.7	33.8

Table 1. "Common Patterns" collocations of subjects for "renwei" and "yiwei" by "Word Sketch Difference" function

By the function of "Word Sketch Difference", the overall and comparative descriptions on grammatical relations and collocations demonstrated the subject, modifier and sent subject of keywords with the

frequency and salience. Table 1 shows the "Common Patterns" of "Subject" collocated with the keywords "renwei" and "yiwei". The first column is the list of subjects collocated with the two keywords, followed by the frequency and salience of each collocation pattern in the second and third columns, in which there are two sub-columns sequentially displaying the details for "renwei" and "yiwei".

Based on "Common Patterns", although "renwei" and "yiwei" shared the subjects of pronouns like "他/她" (he/she) , "他們" (they) and "我" (I) with relatively approximate salience, "renwei" showed an overwhelmingly frequent distribution of subjects from professional fields, such as experts (專家) and personages（人士) in Table 1. For instance, the salience for the collocations of experts (專家) is 53.5 for "renwei" whereas it is only 4.5 for "yiwei".

Moreover, different distributions of subjects were apparently approved in the "Only Patterns" for the keyword "renwei" by the function of Word Sketch. Tables 2 lists the "Subject" of "Only Patterns" which exclusively collocated with "renwei" rather than "yiwei".

	Collocations Frequency	Collocations Salience
Subject	228268	21.9
觀察家	1704	60.9
分析家	2180	59.7
輿論	2677	58.3
與會者	820	49.4
分析師	803	44.8
經濟學家	762	44.0
行家	218	38.9
科學家	1215	35.2

Table 2. "Only Patterns" collocations of subjects for "renwei" by "Word Sketch" function

In Table 2, people of professional fields with specialized skills, as subjects, collocated with the epistemic verb "renwei" with a quite high salience value, such as the observer (觀察家) (salience of 60.9), the analyst (分析家) (salience of 59.7) and the economist (經濟學家) (salience of 44.0). Two exampled sentences (3) (4) and their translations were listed below.

(3) 但西方觀察家認為，實際數字遠超過此數。
"But the Western observer **thought** that the actual number far exceeded this number."
(4) 經濟學家一致認為，新協定成立之初，必然會有人失業。
"Economists **agreed** that there would inevitably be people who were unemployed when the new pact is created."

Results indicated the differences in the credibility of evidential markers expressed by the cognitive verbs "renwei" and "yiwei". That is, the information source from professionals was associated with a high certainty of evidentiality. In journalism written discourse, the proposition predicated by "renwei" could express the writers' or speakers' attitudes of trust toward the complemented information. While the findings of "yiwei" demonstrated the absence of functions of evidential marking and expressing information trustworthiness from the addressers.

3.3 Differences in Negative Forms

Modifiers	No. of Entries		MI		Only Patterns	
	ren wei	yiwei	ren wei	yi wei	ren wei	yiwei
不 "bu"	456	9	61.9	20.1	Yes	NA
原 "yuan"	NA	6	NA	25.8	NA	Yes

Table 3. Distributions of NPIs "bu" and "yuan" collocating with "renwei" and "yiwei" by "Word Sketch Difference" function

The function of "Word Sketch Difference" also demonstrated the distinct tendency in collocating with negative polarity items (NPIs) among "renwei" and "yiwei". Table 3 involved two NPIs, 不 bu4, "no", and 原 yuan2 "barely", modified "renwei" and "yiwei" with diverse distributions of frequency. Specifically, the construction like "bu" followed by

"renwei" is the "Only Pattern" while the exclusivity is not available (NA) for the condition of "yiwei". Similarly, the form consisting of "yuan" and "yiwei" is the "Only Pattern" and this modifier is not the "Only Pattern" for "renwei".

Results indicated that differences in information trustworthiness may be triggered by different NPIs in terms of the scope of negation, epistemic modality and speakers' attitudes. On the one hand, the NPI like "bu", a function word of negation, triggered a negative context where the epistemic word "renwei" was negated by the speaker as the sentence (5a) exampled. The speaker denied the immediate scope of "bu", the verb "renwei", instead of the subject or object of a sentence, and therefore the constituent negation canceled the epistemic value of the proposition. Whereas the NPI "bu" rarely collocated with "yiwei" as the sentence (5b) was peculiar in communication.

(5a) 目前我們不認為畫像中有任何人是嫌犯。

(5b) * 目前我們不以為畫像中有任何人是嫌犯。

"At the moment we don't **think** anyone in the portrait is a suspect."

On the other hand, results of the polarity shifter "yuan" denoted a distinct degree of information trustworthiness from that of the NPI "bu" expressed. As Table 3 showed, "yuan" exclusively collocated with "yiwei" like sentence (6a) exampled, while this polarity shifter failed to collocate with "renwei" as the sentence (6b) illustrated. The phrase "yuan" followed by "yiwei" did not create an explicit negative context but still denoted implicatures on negative semantic meanings of the entire proposition. That is, the adverbial NPI "yuan" expressed the negative meaning through the comparison of tenses, in which the speaker's present attitude changed and opposed to his or her own thought in the past. The polarity shifter "yuan" consequently reversed the direction of commitments instead of negating the constituents of clauses like the negative form "bu" followed by "renwei" in example (5a). Hence, the trustworthiness of information expressed by the combination "yuan" and "yiwei" was discriminated from the negative form "bu" and "renwei" due to their varieties in the scope of negation.

(6a) 人們原以為，只要大家遵守某些規定，這些發展就可在同一制度下彼此相輔相成。

(6b) *人們原認為，只要大家遵守某些規定，這些發展就可在同一制度下彼此相輔相成。

"People originally **thought** that these developments can complement each other under the same system as long as everyone complies with certain regulations."

Apart from NPIs "bu" and "yuan", "renwei" and "yiwei" also demonstrated differences in negative imperative sentences in terms of the NPIs "buyao" and "bie". Table 4 presented the frequency distributions of these two NPIs collocating with "renwei" and "yiwei" respectively.

Modifiers	No. of Entries		MI		Only Patterns	
	ren wei	yiwei	ren wei	yiwei	ren wei	yiwei
不要 "buyao"	16	48	24.6	56.5	NA	NA
別 "bie"	NA	15	NA	43.7	NA	Yes

Table 4. Distributions of NPIs "buyao" and "bie" collocating with "renwei" and "yiwei" by "Word Sketch Difference" function

For the NPI "buyao", though it could modify both "renwei" and "yiwei", the latter one showed overwhelmingly frequent use of collocating with this modifier in a negative imperative sentence. As for the NPI "bie", however, collocating with the verb "yiwei" was its "Only Pattern" against "renwei". We further conducted new individual queries for the collocations of "renwei" and "yiwei" in which "bie" was the left context of the keywords "renwei" or "yiwei" respectively. There were 211 entries found for "yiwei" whereas only 9 entries for "renwei", even though the overall distribution (section 3.1) of the latter keyword was considerably much more than that of the former one. As sentence (7) illustrated, the speaker expressed strong negative attitudes towards the propositions by using the NPI "bie" in imperative sentences.

(7) 如果您現在仍吃減肥菜，但沒有不適症狀，別以為自己沒事。

"If you are still on a diet but don't have symptoms of discomfort, don't **think** you are fine."

3.4 Differences in Hedging Device

By Sinica Corpus, the present research firstly set newspapers as the searching range which was consistent with the type of Chinese Gigaword2 Corpus used in the Chinese Word Sketch approach (adopted from Section3.1. to Section3.3). Then the political discourse was analyzed to contrast the performances of hedging strategy realized by the markers "renwei" and "yiwei". Table 5 lists the topic distributions of the keywords, and "renwei" was much more frequently used than "yiwei" in nearly all topics.

Topics	"renwei"	"yiwei"
philosophy	61	14
science	64	1
society	2669	274
art	43	7
life	533	109
literature	96	66
Total	3466	471

Table 5. Topic distributions of "renwei" and "yiwei" in Sinica Corpus

In the topic of society, six sub-topics of 政黨 "political party", 內政 "domestic affairs", 軍事 "military affairs", 政治現象"political phenomena", 國際關係 "international relations" and 國家政策 "national policy" were set to locate the genre into political discourse in Sinica Corpus. Table 6 shows the frequency of keywords used in sub-topics of political discourse. The small number of entries (33) indicated that the marker "yiwei" was rarely used by addressers in newspapers especially on the political topic. Therefore the following part will analyze the role of hedging device for "renwei" by cases.

Sub-topic	"renwei"	"yiwei"
Political Discourse	689	33

Table 6. Politic sub-topics distributions of "renwei" and "yiwei" in Sinica Corpus

The epistemic marker "renwei" demonstrated the pragmatic function of hedging devices to express fuzzy information under the consideration of speakers' stances or attitudes. For example, speakers may mitigate their claims by showing the uncertainty of the statement like the sentence (8) indicated. Together with the possibility, the epistemic marker "renwei" emphasized that the speaker was unconfident in the realization of peace.

(8) 特使普利馬可夫透過一名通譯說：我依然感到樂觀，我認為和平有希望達成。
"I remain optimistic, and I **think** there is hope for peace," Special Envoy Primakov said through a translator."

Also, the marker "renwei" used as hedging reflected the power relation in the political domain. The addressers of governmental agencies were entailed high power relation involving governing authority. In newspapers on politics, therefore, the use of "renwei" was a way to mitigate their strong commands, like sentence (9) exampled.

(9)即使政府認為電子化政府與電子商務勢在必行，也大可透過市場競爭的模式，讓業者各自推動其智慧卡。
"Even if the Government **considers** that electronic government and e-commerce are imperative, it can also enable operators to promote their smart cards through the market competition model."

Moreover, the epistemic marker "renwei" tended to co-occur with other grammatical structures to hedge the attitudes toward either the facts or opinions of statements together. As for findings of the current study, the modal verbs, conditional clauses or possibilities were all used together with the marker of epistemic modality.

4 Discussion and Conclusions

This study has adopted a corpus-based method to compare the information trustworthiness between two epistemic modalities attributed to "renwei" and "yiwei". The results obtained provide the literature with multiple perspectives to retrieve the information credibility from cognitive verbs.

First, on the aspect of evidentiality, statements attributed to the marker "renwei" are relatively trustworthy because the source of information is more authoritative or professional. It seems partially consistent with previous findings (González et al. 2017) that uncertainty epistemic markers and direct evidential were preferred to be used in oral conversation, while written discourse was associated with a relatively high degree of certainty. However, further comparative studies on spoken Chinese are needed before we conclude that the information conveyed by "yiwei" is more opinionated. Because the datasets adopted here were collected from newspapers, the written discourse, and the distribution of "yiwei" was dramatically less frequent than that of "renwei".

Moreover, the relatively low degree of information trustworthiness attributed to "yiwei" could also be accounted for the polarity theory that negative attitudes towards entire propositions were associated with specific NPIs (e.g. "yuan") exclusively collocating with "yiwei". As Riemer and Dittmer (2016) discussed on computational treatments, the interaction between modality and negation was also proved in text understanding in the current study.

Then, this research found only "renwei" used as hedging in the political discourse. Its functions of hedging the certainty of information and mitigating strong statements were proved in Chinese political discourse. The effects of power relation on differentiating near synonyms were proved among epistemic verbs, which was also applied to the previous study on other Chinese near synonyms verbs (Wang and Huang, 2018).

Furthermore, the methodology may also contribute to the comparative analysis of near synonyms with a new perspective. That is, the variations in epistemic modality were thoroughly investigated in terms of multiple strategies, which took the factor of speakers' cognitive status into consideration. This contribution perfectly meets the requirement of more precise processing of natural language rather than neglect the attitudes or implicatures of speakers.

In conclusion, the present study provided a new approach to detect the information trustworthiness of text understanding attributed to Chinese epistemic verbs. The near synonymous cognitive verbs each has its own inherent epistemic commitment; yet a speaker can further enhance the (un) trustworthiness with both linguistic (adjuncts, (im) personal subjects, etc.) and extra-linguistic (power-relations, meta-information of speaker identities) cues. The empirical evidence may contribute to future annotation projects, especially for studies on Chinese near synonyms in NLP, as the ability to distinguish and annotate diverse information credibility has been necessary due to the overwhelming information nowadays.

Acknowledgments

We thank Dr. Xuefeng Gao for his invaluable contributions, and three anonymous reviewers for their constructive comments.

References

Brown, Penelope, and Stephen C. Levinson. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press, 1987.

Chafe, Wallace. "Evidentiality in English conversation and academic writing." *Evidentiality: The linguistic coding of epistemology* 20 (1986): 261-272.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul: Kyung Hee University. pp.167-176. https://www.aclweb.org/anthology/Y96-1018.pdf

Chen, Keh-Jiann, and Chu-Ren Huang. Academia Sinica Balanced Corpus. In Rint Sybesma (Ed.), Encyclopedia of Chinese Language and Linguistics. Brill. 2016.

De Haan, Ferdinand. "Encoding speaker perspective: Evidentials." *Linguistic diversity and language theories* 379 (2005): 397.

Fraser, Bruce. "Hedging in political discourse." *OKULSKA, U., CAP, P., Perspectives in Politics and Discourse, Capitolo* 8 (2010).

González, Montserrat, Paolo Roseano, Joan Borràs-Comes, and Pilar Prieto. "Epistemic and evidential marking in discourse: Effects of register and debatability." *Lingua* 186 (2017): 68-87.

Horn, Laurence. "On the semantic property of logical operators in English." *Published by Indiana University Linguistics Club, University of California Los Angeles* (1972).

Horn, Laurence. "A natural history of negation." (1989).

Horn, Laurence R. A natural history of negation. CSLI. 2001.

Horn, Laurence R. "WJ-40: Implicature, truth, and meaning." *International review of pragmatics* 1.1 (2009): 3-34.

Horn, Laurence R., and Yasuhiko Kato, eds. *Negation and polarity: Syntactic and semantic perspectives*. OUP Oxford, 2000.

Hyland, Ken. "Writing without conviction? Hedging in science research articles." *Applied linguistics* 17.4 (1996): 433-454.

Israel, Michael. "The pragmatics of polarity." *The handbook of pragmatics* (2004): 701-723.

Koźbiał, Dariusz. "Epistemic Modality: A Corpus-Based Analysis of Epistemic Markers in EU and Polish Judgments." *Comparative Legilinguistics* 41.1 (2020): 39-70.

Lakoff, George. "Hedges: A study in meaning criteria and the logic of fuzzy concepts." *Contemporary research in philosophical logic and linguistic semantics*. Springer, Dordrecht, 1975. 221-271.

Markkanen, Raija, and Hartmut Schröder, eds. *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*. Vol. 24. Walter de Gruyter, 2010.

Nuyts, Jan. "Modality: Overview and linguistic issues in Frawley." *The expression of Modality–Berlin–New York: Mouton de Gryuter* (2006).

Palmer, Frank Robert. *Mood and modality*. Cambridge University Press, 2001.

Penka, Doris, and Hedde Zeijlstra. "Negation and polarity: an introduction." *Natural Language & Linguistic Theory* 28.4 (2010): 771-786.

Riemer, Manuel, and Livia Dittmer. "An introduction to the special issue." *Ecopsychology* 8.3 (2016): 163-166.

Rubin, Victoria L., Elizabeth D. Liddy, and Noriko Kando. "Certainty identification in texts: Categorization model and manual tagging results." *Computing attitude and affect in text: Theory and applications*. Springer, Dordrecht, 2006. 61-76.

Taweel, Abeer Q., Emad M. Saidat, Hussein A. Rafayah, and Ahmad M. Saidat. "Hedging in Political Discourse." *Linguistics Journal* 5.1 (2011).

Speranza, John L., and Laurence R. Horn. "A brief history of negation." *Journal of Applied Logic* 8.3 (2010): 277-301.

Su, Qi, Chu-Ren Huang, and Helen Kaiyun Chen. "Evidentiality for text trustworthiness detection." *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*. 2010.

Xu, Ge, and Churen Huang. "Mining Chinese Polarity Shifters." *Workshop on Chinese Lexical Semantics*. Springer, Cham, 2015.

Wang, Xiaowen, and Chu-Ren Huang. "From Near Synonyms to Power Relation Variations in Communication: A Cross-Strait Comparison of "Guli" and "Mianli"." *Workshop on Chinese Lexical Semantics*. Springer, Cham, 2018.

Willett, Thomas. "A cross-linguistic survey of the grammaticization of evidentiality." *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 12.1 (1988): 51-97.

Empirical Study of Text Augmentation on Social Media Text in Vietnamese

Son T. Luu
University of Information Technology
VNU-HCM, Vietnam
sonlt@uit.edu.vn

Kiet Van Nguyen
University of Information Technology
VNU-HCM, Vietnam
kietnv@uit.edu.vn

Ngan Luu-Thuy Nguyen
University of Information Technology
VNU-HCM, Vietnam
ngannlt@uit.edu.vn

Abstract

In the text classification problem, the imbalance of labels in datasets affect the performance of the text-classification models. Practically, the data about user comments on social networking sites not altogether appeared - the administrators often only allow positive comments and hide negative comments. Thus, when collecting the data about user comments on the social network, the data is usually skewed about one label, which leads the dataset to become imbalanced and deteriorate the model's ability. The data augmentation techniques are applied to solve the imbalance problem between classes of the dataset, increasing the prediction model's accuracy. In this paper, we performed augmentation techniques on the VLSP2019 Hate Speech Detection on Vietnamese social texts and the UIT - VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. The result of augmentation increases by about 1.5% in the F1-macro score on both corpora.

1 Introduction

In recent years, the growth of hate speech has become a crime, not only face-to-face action but also online communication (Fortuna and Nunes, 2018). The development of social network nowadays had made this situation worse. The threading of harassment comments and harassment speech makes the user stop expressing their opinions and looking up for other ideas (Vu et al., 2019). Fortuna and Nunes (2018) defined hate speech as the language attacking, diminishing, and inciting violence or hate against individuals or groups based on their characteristics, religion, nations, and genders. To solve this problem, many datasets are constructed to detect and classify user comments on social network sites such as Facebook, Twitter, and Facebook in many languages[1].

The HSD-VLSP dataset (Vu et al., 2019) provided by the VLSP 2019 Shared task about Hate speech detection on Social network[2] contained nearly 25,000 comments and posts of Vietnamese Facebook users and has three labels. However, the distribution of three classes in the dataset is imbalanced. Besides, the UIT-VSFC dataset (Nguyen et al., 2018) that was used for predicting the feedback from students contained about 16,000 sentences and was annotated for two different tasks: sentiment analysis and topic analysis. Same as the HSD-VLSP dataset, the distribution of labels on the UIT-VSFC dataset is also imbalanced. We use the data augmentation techniques to generate new comments that belong to minority classes from the original dataset to tackle those restrictions. We conduct experiments on the augmented dataset and compare it with the original dataset to indicate data augmentation effectiveness. Those augmentation techniques include synonym replacement, random insertion, random swapping, and random deletion (Wei and Zou, 2019).

[1] http://hatespeechdata.com/
[2] https://www.aivivn.com/contests/8

The rests of the paper are structures as below. Section 2 introduces recent works in hate speech detection. Section 3 gives an overview of two datasets include the HSD-VLSP dataset and the UIT-VSFC dataset. Section 4 presents the methods and models used in our paper. Section 5 shows our experiment results when applied to the text augmentation techniques. Section 6 concludes the paper.

2 Related works

Duyen et al. (2014) conducted an empirical study about the sentiment analysis for Vietnamese texts based on machine learning to study the influences on the models' accuracy. However, besides the impact of the model's ability, and the feature selection such as word-based, syllable-based, and extracting essential words, the imbalance in the dataset also affects the result. The imbalance in label distribution happens regularly (Ali et al., 2015) when one class seems to be more interested than the other. For example, in social media networks, the abusive and hateful comments are often hidden by the users or administrators, since the clean comments take the majority part. The VLSP2019 hate speech dataset (Vu et al., 2019) and the UIT-VSFC dataset (Nguyen et al., 2018) also suffer the imbalance in class distribution. The detail of those datasets is showed in Section 4.

Wang and Yang (2015) provided a novel method for enhancing the data used for behavior analysis using social media texts on Twitter. Their approaches include using the lexical embedding and frame-semantic embedding. The obtained results showed that using the data augmentation brings significantly better results than no data augmentation (using Google New Lexical embedding brings 6.1% improvement in F1-score and using additional frame-semantic embedding from Twitter brings 3.8% improvement in F1-score.

Ibrahim et al. (2018) presented different data augmentation techniques for solving the imbalance problem in the Wikipedia dataset and an ensemble method used for the training model. The result achieved a 0.828 F1-score for toxic and nontoxic classification, and 0.872 for toxicity types prediction.

Rizos et al. (2019) introduced data augmentation techniques for hate speech classification. The authors 's proposed methods increased the result of hate speech classification to 5.7% in F1-macro score.

Finally, Wei and Zou (2019) provided EDA (Easy Data Augmentation) techniques used to enhance data and boost performance on the text classification task. It contains four operations: synonym replacement, random insertion, random swap, and random deletion. In this paper, these operations are applied to the HSD-VLSP 2019 dataset and the UIT-VSFC dataset to increase the classification models' ability.

3 Datasets

3.1 The HSD-VLSP dataset

The hate speech dataset was provided by the VLSP 2019 shared task about hate speech detection for social good (Vu et al., 2019). The dataset contains a total of 20,345 comments and posts crawled from Facebook. Each comment is labeled by one of three labels: CLEAN, OFFENSIVE, and HATE. Table 1 showed the overview information about the dataset.

	Num. comments	**Avg. word length**	**Vocab. size**
CLEAN	18,614	18.69	347,949
OFFENSIVE	1,022	9.35	9,556
HATE	709	20.46	14,513
Total	20,345	18.28	372,018

Table 1: Overview of the HSD-VLSP dataset

According to Table 1, the number of CLEAN comments take a majority part in the dataset, the number of OFFENSIVE comments and HATE comments are much fewer. Thus, the distribution of labels in the dataset is imbalanced.

3.2 The UIT-VSFC dataset

The Vietnamese Students' Feedback Corpus for Sentiment Analysis (UIT-VSFC) by Nguyen et al. (2018) are used to improve the quality of

education. The dataset contains nearly 11,000 sentences and consists of two tasks: sentiment-based classification and topic-based classification. The sentiment-based task comprises three labels: positive, negative, and neutral. The topic-based task comprises four labels corresponding to lecturer, training program, facility, and others. Table 2 describes the overview about the UIT-VSFC training set.

	Num. comments	**Avg. word length**	**Vocab. size**
Total	11,426	10.2	117,295
Sentiment based task			
Positive	5,643	8.2	46,807
Negative	5,325	12.6	67,193
Neutral	458	7.1	3,295
Topic based task			
Lecturer	8,166	9.7	79,854
Training program	2,201	12.2	27,039
Facility	497	12.3	6,130
Others	562	10.9	4,272

Table 2: Overview of the UIT-VSFC training set

According to Table 2, the number of data in the neutral label is lower than positive and negative on the sentiment-based task. So is the topic-based task when the *facility* and *others* labels are much lower than the two remain labels. In brief, the imbalance data happened on the neutral label for the sentiment-based task, and the *facility* and the *other* labels for the topic-based task.

4 Our proposed method

4.1 The augmentation techniques

In this paper, we implement the EDA techniques introduced by Wei and Zou (2019). Those techniques will get a sentence as input and perform one of these following operations to generate new comments:

- **Synonym replacement (SR)**: This operation creates a new sentence by randomly choosing n words from the input sentence and replaces them by their synonyms, excluding the stop words. In our experiments, we use the Vietnamese wordnet[3] from Nguyen et al. (2016) for synonym replacement and the Vietnamese stopword dictionary[4] for removing stop words in the sentence.

- **Random Insertion (RI)**: This operation generates new data by first finding a random word in the input sentence, which is not a stop word, then taking its synonym and putting it into the sentence's random position. The synonyms are taken from the Vietnamese wordnet[3].

- **Random Swap (RS)**: This operation makes a new sentence by choosing two random words in the input sentence and swap their position.

- **Random Deletion (RD)**: This operation creates a new sentence by accidentally deleting p words in the sentence (p is the probability defined before by the user).

According to Wei and Zou (2019), n indicates the number of changed words for SR, RI, and RS methods, which calculated as $n = \alpha * l$, where α is the percentage of replacement word in the sentence and l is the length of the sentence. For the RD method, the probability of deletion words p equal to α. The α is defined by the user.

Table 3 shows examples of data between original and after augmented by EDA techniques in the HSD-VLSP dataset.

[3] https://github.com/zeloru/vietnamese-wordnet

[4] https://github.com/stopwords/vietnamese-stopwords

Comments	Type
Original: con này xấu trai vl *(this guy is f*cking ugly)* **Augmented**: con xấu trai vl	RD
Original: Đcm nản vl *(This is f*cking bored)* **Augmented**: Đcm nhụt chí vl	SR
Original: Đume đau răng vl *(Toothache got damn hurt!)* **Augmented**: Đume răng đau vl	RS
Original: Đm Lắm chuyện vl *(F*uck those curious guys)* **Augmented**: Đm thứ Lắm chuyện vl	RI

Table 3: Several example of the augmented data on the HSD-VLSP dataset

4.2 The classification model

Aggarwal and Zhai (2012) defined the text classification problem as a set of training data $D = \{X_1, ..., X_N\}$, in which each record is labeled with a class value drawn from a set of discrete classes indexed by $\{1..k\}$. The training data used to construct a classification model. With a given test dataset, the classification model is used to predict a class for each instance in the test dataset. Our paper used the Text-CNN model (Kim, 2014) for the HSD-VLSP dataset and the Maximum Entropy model (Nigam et al., 1999) for the UIT-VSFC dataset to study the effectiveness of data augmentation on those two datasets. In practice, the idea of Logistic Regression is maximizing the cross-entropy loss of the actual label in the training dataset (Jurasky and Martin, 2000), which is the same as the Maximum Entropy model (Nigam et al., 1999). Thus, we use the term Maximum Entropy instead of Logistic Regression in our results.

5 Empirical results

5.1 Experiment configuration

For the HSD-VLSP corpus, we use cross-validation with five folds for the Text-CNN model and the Maximum Entropy model. Following the same manner in the previous study (Luu et al., 2020), for each fold, we keep the test set and enhance the training set with EDA techniques.

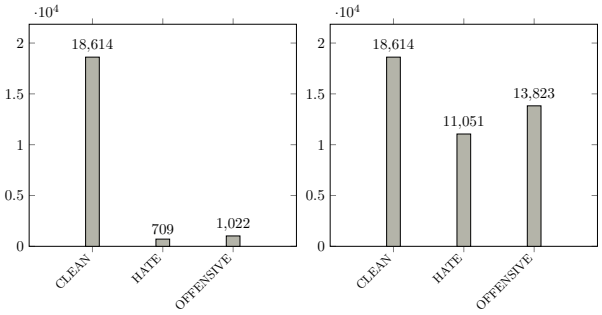
For the UIT-VSFC dataset, we used the data divided into the training, development, and test sets by Nguyen et al. (2018). Then we run the EDA techniques on the training set and use the test set to evaluate the result.

5.2 Data augmentation result

We first applied the EDA techniques on the entire original HSD-VLSP dataset to enhance the data on HATE and OFFENSIVE labels. Table 4 describes the information about the HSD-VLSP dataset after making data augmentation.

	Num. comments	Avg. word length	Vocab. size
CLEAN	18,614	19.3	360,958
OFFENSIVE	13,823	11.3	157,517
HATE	11,051	23.6	260,841
Total	43,488	17.9	779,316

Table 4: The augmented HSD-VLSP courpus



Original dataset Augmented dataset

Figure 1: Number of comments on before and after augmentation in the HSD-VLSP dataset

It can be inferred from Table 4 that after applying the EDA techniques, the number of data and the vocabulary size on the HATE and OFFENSIVE labels and increased significantly (Words calculate the vocabulary size, and we use the pyvi[5] for tokenizing). Figure 1 illustrates the distribution of three classes before and after us-

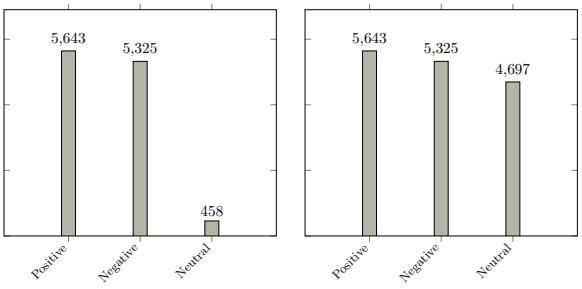
[5] https://pypi.org/project/pyvi/

ing data augmentation techniques on the HSD-VLSP dataset. After using EDA techniques, the data on three labels are well-distributed.

Besides, we apply the EDA on the UIT-VSFC training set to enhance the data on the neutral label for the sentiment-based task, and on *facility* and *other* labels for the topic-based task. Table 5 describes the UIT-VSFC training set after enhanced. Comparing with the original dataset as described in Table 2, the number of comments and the vocabulary size of the neutral label on the sentiment-based task increased significantly. Same as the sentiment-based task, the number of comments and vocabulary size of the *facility* and *other* labels are also dramatically increased.

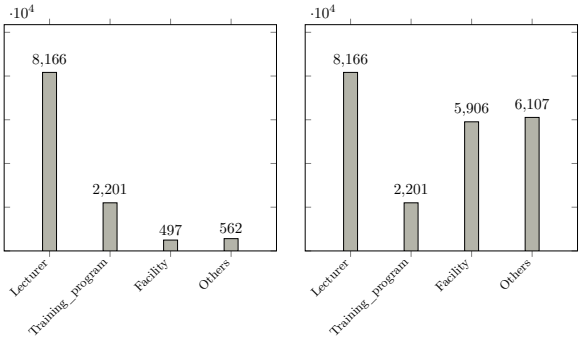
	Num. comments	**Avg. word length**	**Vocab. size**
Sentiment-based task			
Positive	5,643	8.2	46,807
Negative	5,325	12.6	67,193
Neutral	4,697	8.1	38,349
Total	15,665	9.7	152,349
Topic-based task			
Lecturer	8,166	9.7	79,854
Training program	2,201	12.2	27,039
Facility	5,906	13.7	81,299
Others	6,107	13.3	54,722
Total	22,380	10.8	242,914

Table 5: The augmented UIT-VSFC training set



Original training set Augmented training set

Figure 2: The distribution of the sentiment-based task's labels of the UIT-VSFC dataset before and after enhanced



Original training set Augmented training set

Figure 3: The distribution of the topic-based task's labels of the UIT-VSFC dataset before and after enhanced

Figure 2 and Figure 3 illustrate the UIT-VSFC training dataset before and after enhanced data on sentiment based task and topic based task respectively. For the two tasks, after augmentation the distribution of data between labels are balanced.

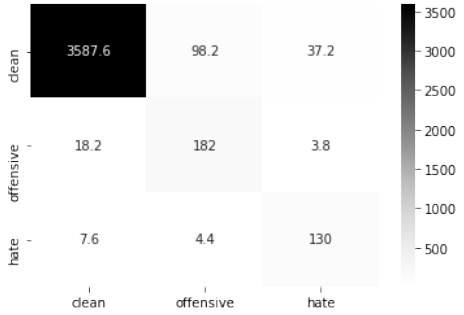
5.3 Model performance results

We implement the Text-CNN model on the entire original HSD-VLSP dataset and the augmented HSD-VLSP dataset. Table 6 shows the result by F1-macro score. Comparing with the original results (Luu et al., 2020), the accuracy of the HSD-VSLP dataset after using augmented techniques are higher than the original dataset. According to Figure 4, the number of right prediction on the *offensive* and the *hate* labels are increased.

Methodology	**F1-macro (%)**
Text-CNN (original) (Luu et al., 2020)	83.04
Text-CNN (augmented)	**84.80**
Maximum Entropy (original) (Luu et al., 2020)	64.58
Maximum Entropy (augmented)	**75.27**

Table 6: Empirical result by the Text-CNN model on the HSD-VLSP dataset

The original dataset



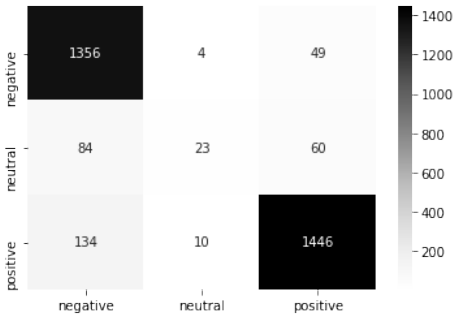
The augmented dataset

Figure 4: Confusion matrix of Text-CNN model before and after enhanced data on the HSD-VLSP

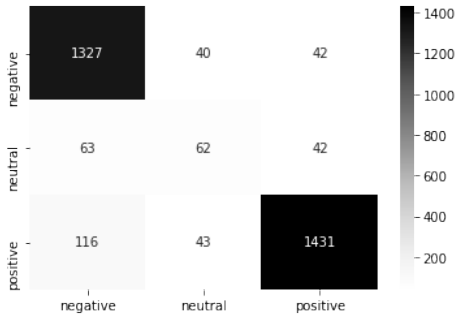
Methodology	**F1-micro (%)**	**F1-macro (%)**
Sentiment-based task		
Maximum Entropy (original)	87.94	68.47
Maximum Entropy (augmented)	**89.07**	**74.32**
Text-CNN (original)	**89.82**	75.57
Text-CNN (augmented)	89.38	**77.16**
Topic-based task		
Maximum Entropy (original)	84.03	71.23
Maximum Entropy (augmented)	**86.03**	**74.87**
Text-CNN (original)	86.63	75.23
Text-CNN (augmented)	86.32	74.86

Table 7: Empirical result of the UIT-VSFC dataset

Besides, Table 7 shows the result of the UIT-VSFC dataset on the sentiment-based and the topic based tasks, respectively, before and after enhanced data. The original F1-micro score of the UIT-VSFC on both sentiment-based and topic-based tasks are referenced from (Nguyen et al., 2018).



The original training set



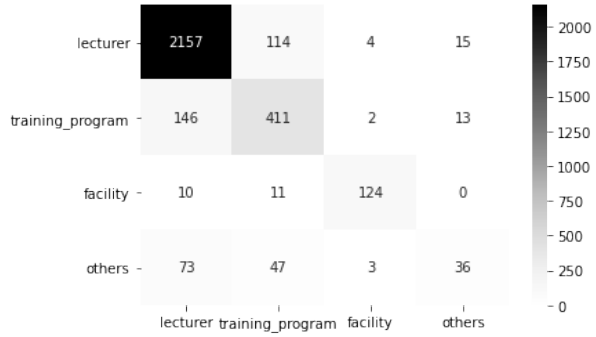
The augmented training set

Figure 5: Confusion matrix of the Maximum Entropy model on the UIT-VSFC dataset before and after enhanced data for the sentiment-based task

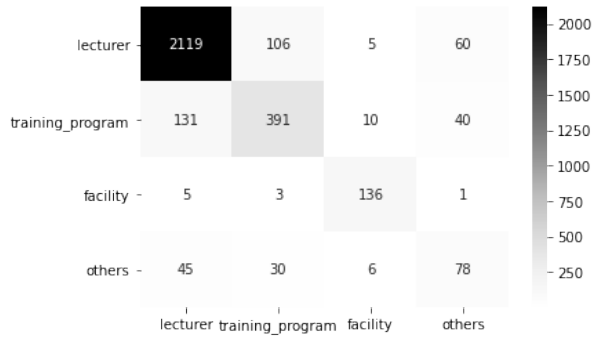
According to Table 7, for the sentiment-based task, the UIT-VSFC dataset, after enhanced on the training set, gave better results than the original training set by the Maximum Entropy model on both F1-micro and F1-macro scores. The Text-CNN model gave better results by the F1-macro score when the training data are enhanced. For the topic based task, the result of Maximum Entropy are better after enhanced data. The Text-CNN results after augmented data, in contrast, are not as better as the original data.

In addition, Figure 5 illustrates the confusion matrix of the UIT-VSFC dataset trained by the Maximum Entropy model before and after enhanced data for the sentiment-based task, and Figure 6 indicates the confusion matrix of the UIT-VSFC dataset for the topic based task

trained by the Maximum Entropy model. According to Figure 5, the ability of true prediction on the neutral label is increased after enhanced data. Nevertheless, according to Figure 6, the results before and after augmented data are just slightly different. Indeed, the enhanced data does not affect much on the performance result of the topic based task.



The original training set



The augmented training set

Figure 6: Confusion matrix of the Maximum Entropy model before and after enhanced data on the UIT-VSFC dataset for the topic-based task

Overall, for the HSD-VLSP hate speech dataset, the data augmentation techniques increase the models' performance. For the UIT-VSFC corpus, the data augmentation increased models' performance on the sentiment-based task by both Maximum Entropy and Text-CNN, while it does not impact the topic-based task.

5.4 Error analysis

According to Figure 6, on the UIT-VSFC dataset on the topic based task, the prediction of the *training_program* label seems to be inclined to the *lecture* label, and *others* label seem to be inclined to the *training_program* and the *lecturer* labels. Table 8 listed examples of those cases. It can be inferred from Table 8 that, most of cases the model predicted wrong to the *lecture* label because the texts have words related to lecture such as: teacher, teaching, lesson, and knowledge. So does the *training_program* label with the appearance of words related to training program topic such as: subjects, requirements, and outcomes.

No.	Texts	True	Predict
1	cô nhiệt tình, giảng bài hiệu quả (English: The teacher is so enthusiastic and teaches very well)	1	0
2	tiến độ dạy hơi nhanh (English: The teaching process is fast)	1	0
3	sinh viên khó tiếp thu kiến thức (English: Student feel difficult to understand the knowledge)	3	0
4	các yêu cầu của môn cần ghi rõ (English: The subject's requirements should be well described)	3	1

Table 8: Error analysis in the test set of the UIT-VSFC dataset on topic-based task. Label description: 0 - lecturer, 1 - training program, 2 - facility, 3 - others

6 Conclusion

The imbalance in the datasets impact the performance of the machine learning models. Therefore, this paper focuses on the techniques that decreased the skewed distribution in the dataset by enhancing minority classes' data. We implemented the EDA techniques on the VLSP hate speech and the UIT-VSFC datasets and studied

data augmentation's effectiveness on the imbalanced dataset. The results show that, when the data on the minority labels are increased, the model's ability to predict those labels is higher. However, the data augmentation techniques pull down the accuracy of other labels. Therefore, it is necessary to consider whether it is appropriate to apply the data augmentation techniques in a specific problem.

In the future, we will construct the lexicon-based dictionary for sentiment analysis in the Vietnamese language, especially the abusive lexicon-based words like Hurtlex (Bassignana et al., 2018) for hate speech detection to improve the ability of the machine learning model. We will also implement modern techniques in text classification such as the BERT model (Devlin et al., 2019) and the attention model (Yang et al., 2016) to increase the performance. Furthermore, in the hate speech detection problem, we will construct a new dataset which is more diverse in data sources and more balanced among classes.

Acknowledgments

We would like to give our great thanks to the 2019 VLSP Shared Task organizers for providing a very valuable corpus for our experiments.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer.

Aida Ali, Siti Mariyam Hj. Shamsuddin, and Anca L. Ralescu. 2015. Classification with class imbalance problem: a review. In *SOCO 2015*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

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.

N. T. Duyen, N. X. Bach, and T. M. Phuong. 2014. An empirical study on sentiment analysis for vietnamese. In *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, pages 309–314.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July.

M. Ibrahim, M. Torki, and N. El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Daniel Jurasky and James H Martin. 2000. Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey*.

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, October. Association for Computational Linguistics.

S. T. Luu, H. P. Nguyen, K. Van Nguyen, and N. Luu-Thuy Nguyen. 2020. Comparison between traditional machine learning models and neural network models for vietnamese hate speech detection. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6.

Phuong-Thai Nguyen, Van-Lam Pham, Hoang-Anh Nguyen, Huy-Hien Vu, Ngoc-Anh Tran, and Thi-Thu Ha Truong. 2016. A two-phase approach for building vietnamese wordnet. In *Proceedings of the 8th Global WordNet Conference. Bucharest, Romania*, pages 259–264.

K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L. Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholom, Sweden.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th*

ACM International Conference on Information and Knowledge Management, CIKM 19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen T M. Nguyen. 2019. HSD shared task in VLSP campaign 2019: Hate speech detection for social good. In *Proceedings of VLSP 2019*.

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, Lisbon, Portugal, September. 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 6383–6389, Hong Kong, China, November. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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, San Diego, California, June. Association for Computational Linguistics.

Attention-based Domain Adaption Using Transfer Learning for Part-of-Speech Tagging: An Experiment on the Hindi Language

Rajesh Kumar Mundotiya, Vikrant Kumar, Arpit Mehta and **Anil Kumar Singh**
Department of Computer Science and Engineering, IIT(BHU), Varanasi, India
{rajeshkm.rs.cse16, vikrantkumar.cse18}@iitbhu.ac.in,
{arpitmehta.cse18, aksingh.cse}@iitbhu.ac.in

Abstract

Part-of-Speech (POS) tagging is considered a preliminary task for parsing any language, which in turn is required for many Natural Language Processing (NLP) applications. Existing work on the Hindi language for this task reported results on either the General or the News domain from the Hindi-Urdu Treebank that relied on a reasonably large annotated corpus. Since the Hindi datasets of the Disease and the Tourism domain have less annotated corpus, using domain adaptation seems to be a promising approach. In this paper, we describe an attention-based model with self-attention as well as monotonic chunk-wise attention, which successfully leverage syntactic relations through training on a small dataset. The accuracy of the Hindi Disease dataset performed by the attention-based model using transfer learning is 93.86%, an improvement on the baseline model (93.64%). In terms of F_1-score, however, the baseline model (93.65%) seems to do better than the monotonic-chunk-wise attention model (94.05%).

1 Introduction

Deep learning has been consistently providing promising results on a large variety of language processing problems. Textual processing includes diverse applications of NLP such as text classification, dialect identification and classification, sequence labelling problems (such as Named Entity Recognition and Extraction, Chunking and POS tagging) and machine translation.

However, for performance improvement obtained on the preliminary NLP tasks – POS tagging and Chunking – especially under a low resource scenario, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) have been used more. An efficient way of information modeling by Gated Recurrent (GRU) and Long Short Term Memory (LSTM), a variant of RNN, has also been tried.

Earlier work on POS tagger for canonical Hindi text achieved considerable results of about 97.10% on Universal Dependency dataset (Plank et al., 2016), which belongs to a single domain. The performance reduces radically after deploying this existing trained model to a different domain-specific data or out-of-domain data. Domain-specific data such as Tourism and Disease has its own distributions and having a minimal amount of annotated dataset, considered as low resources, which also causes an Out-of-vocabulary (OOV) words issue.

OOV is a major problem in low resources text processing, faced while training a model on one domain of a language and trying it to another domain of the same language. This problem is partly countered by incorporating character level information into the model.

Lately, Transfer Learning has been shown to enhance the performance of the model by transferring learned features (general features as well as domain-specific features) which were obtained during training the model. The general features are transferred to the target domain through an initializer or feature extractor. These methods are beneficial as they benefit from the pre-trained model via neurons (Zennaki et al., 2019). Yang et al. (2017), Meftah et

al. (2018) have followed the Transfer Learning approach on English (following Subject-Verb-Object sentence structure), while there is not much work for Hindi (following Subject-Object-Verb sentence structure) using such models.

The proposed architecture of (Ma and Hovy, 2016) is employed as a baseline model for the purposes of our work. It encodes character level information by CNN. Authors have strengthened the baseline model through attention mechanism: self-attention and monotonic chunk-wise attention as the contribution. The motivation behind using these attention mechanisms is that it exhibits adequate improvement on neural machine translation, especially for low resource regime (Chiu and Raffel, 2017; Bahdanau et al., 2014; Goyal et al., 2020). Also, the experimental datasets required can be smaller in size. The improvement in capturing syntactic information is due to the attention mechanisms. The results obtained by the attention mechanism provide an improvement over the original baseline results.

2 Baseline Model

We use as our baseline the above mentioned model using a discriminative tagging model proposed by Ma et al. (2016), together with character-level information encoded by CNN, illustrated in Figure 1.

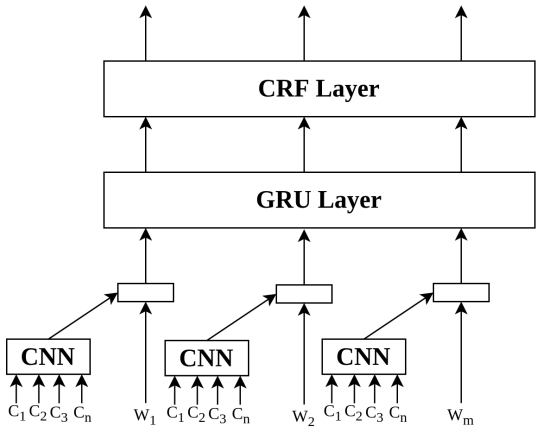


Figure 1: Baseline model for POS Tagging

In this model, the preservation of both syntactic and semantic information of words is achieved by a combination of two vectors obtained at word-level and character-level (Murthy et al., 2018).

The character-level information captures orthographic and morphological features by applying CNN (Murthy et al., 2018), where characters are initially represented by a one-hot encoder and passed to convolution layer. The convolution layer holds n-gram information followed by max-pooling layer, where n is given by filter size. Maximum relevant information over the different features perceived through this layer, which are the distinct features of the word, represented at the character level, are passed to a fully connected layer. This layer used a Rectifier Linear Unit (ReLU) as a non-linear activation function to produce character-level word vector. The word vector is assigned by random initialization which is learnt during model training. The concatenated character and word-level vector is fed to the Bidirectional GRU. The obtained output from forward and backward GRUs at each time-step are combined before being fed to a Conditional Random Fields (CRF) layer. The CRF layer generate a probability score over the labels at each time-step.

3 Attention Based Model

Since the last few years, attention mechanisms have been providing promising results in NLP applications as well, e.g. Machine Translation gets a better alignment between the source and the targets words after applying the attention mechanism (Bahdanau et al., 2014; Chiu and Raffel, 2017). Here, we use two attention mechanisms into the baseline model: self-attention (Cheng et al., 2016) and Monotonic Chunkwise Attention (MOCHA) (Chiu and Raffel, 2017) to enhance the capabilities of capturing syntactic relations from input words.

3.1 Attention Mechanism

Self-attention or intra-attention (Cheng et al., 2016) became popular after a Transformer model came into existence for Neural Machine Translation (Vaswani et al., 2017). The Transformer proposed by Vaswani et al. (2017) completely relied on self-attention, which uses different positions of the input to obtain the attention score. The primary reason for calling self-attention as intra-attention is a dependency on itself for score calculation, which is calculated by applying softmax over the additive or dot product of the current vector with previous at-

tention score. These intra-word dependencies are helpful for capturing the syntactic relations among words during labelling.

Monotonic chunk-wise attention (Chiu and Raffel, 2017) is also an extension of Hard monotonic attention. It provides flexibility to the attention score calculation. In this method, the calculation of energy score is based on the chunk (a particular static word window size) rather than entire word input (usually following soft attention) or a particular time-step of input (generally following Hard monotonic attention). The energy score uses chunk energy (soft attention over a limited window) and monotonic energy (Bahdanau attention (Bahdanau et al., 2014) with a sigmoid function instead of softmax) to calculate the attention score. This attention score is calculated for each time-step input.

3.2 Attention-based Model

The previous extensions to the attention mechanisms are based on the encoder-decoder architecture, prevalent in end-to-end neural machine translation systems. In our work, two Bidirectional GRU layers are exploited for incorporating the attentions in the baseline model for POS tagging. The first GRU layer is treated as an encoder for attention and the remaining layer as a decoder for the attention-based extended baseline model. The dropout layer is also used between attention input and output to prevent overfitting. The rest of the model architecture from the input data by CNN and word vector to predictions by CRF are the same as the baseline model, as shown in Figure 2.

3.3 Domain Adaption model

Domain adaption has been performed with supervised, unsupervised and semi-supervised settings until now for many tasks including POS tagging. We have used relatively little annotated data to build a robust POS tagger for the target domain by using Transfer Learning. Transfer Learning procedure closely follows the Meftah et al. (2018) settings. The attention-based model has been trained on the first domain for POS tagging while performing transfer learning. The optimal learned parameters during this training are passed for the training of another domain. That is a standard procedure of transfer learning where all labels are considered as equal.

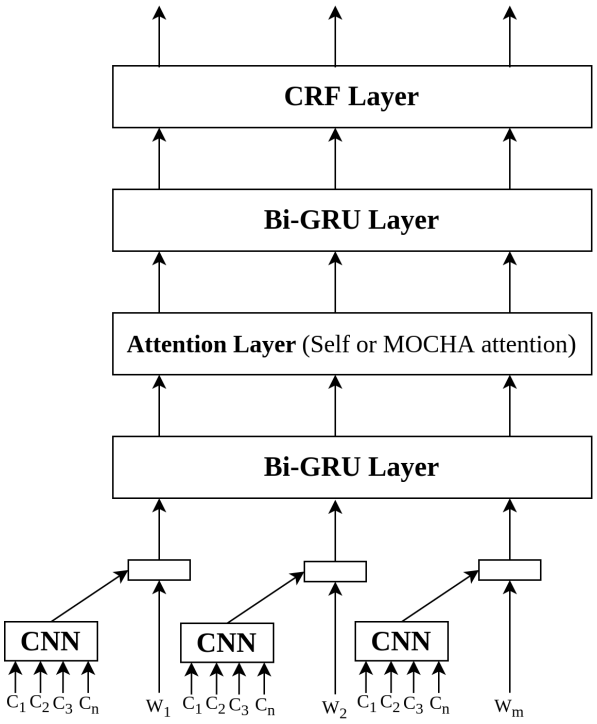


Figure 2: Attention-based extended baseline model for POS Tagging

Here, the optimal parameters θ_s from the training of source domain are used for initialization of the target domain's parameters θ_t. After this initialization ($\theta_s \rightarrow \theta_t$), the model is fine-tuned for the target domain, as shown in Figure 3.

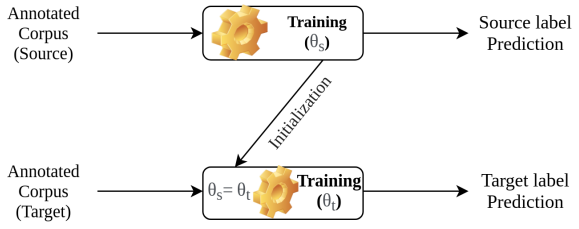


Figure 3: Domain adaption via transfer learning approach

4 Experimental Setup

4.1 Dataset

For performing the experiments of domain adaption, we have used Disease and Tourism domains of the Hindi Treebank dataset[1]. The dataset follows the

[1] http://tdil-dc.in/index.php?lang=en

Bureau of Indian Standards (BIS) tagset. The statistics of the dataset are mentioned in Table 1. As the size of the dataset is small, and out of which overlapped types are 1579, we have extracted Treebank for our experiments.

Domain	Sentences	Types
Tourism	3022	7100
Disease	1494	4987
Overlapping	-	1579

Table 1: Hindi Treebank data statistics according to domain

Since the size of the Disease domain dataset is smaller compared to Tourism, it is considered a source domain, while the other is considered the target domain for domain adaption.

4.2 Settings

The source and target domain datasets are divided in a 70%–30% ratio for performing validation of the trained model. The maximum length of sentences and words has been fixed for training the model, which is 52 and 22, respectively. However, gradient calculation avoided the padded sentences and words, which in turn prevents overfitting. The character vector size 32 are obtained after applying two filters 64 and 124, each with the size of 3, with a dropout of 30%. The model trained with the word vector and GRU unit of 100 and 128, respectively. As annotation corpus is tiny, the model tends to overfit quickly. Hence, dropout and early stoppage have applied with the value of 50% and 30 as patience, respectively. The parameters and hyper-parameters used in training are briefly mentioned in Table 2.

(Hyper-)parameter	Value
Char. vector	30
Word vector	100
Batch Size	32
Filters	[64, 124]
Filter size	3
CNN Dropout	0.3
GRU unit	128
Dropout	0.5
Early stoppage	20
Optimizer	Adam

Table 2: The value of parameters and hyper-parameters used in model training

5 Result and Analysis

The baseline model is also robust towards the POS tagging as the obtained results on the Disease dataset for isolated training has improved by domain adaptions even tough overlapping vocabularies are relatively small (1579 types). The baseline model gets up from 93.64% to 94.29% as in isolation and domain adaption training, respectively, which is the highest accuracy among reported results in Table 3.

The self-attention-based model has degraded the performance due to their nature of attention score calculation and limitation of sentence length. On the other hand, MOCHA-based model has improved the POS tagging system's performance due to the nature of chunk consideration during attention score calculations. We have used a chunk size of 8 in model setup. The MOCHA-based model obtained an accuracy of 93.86%, which has a slight improvement over the baseline model depicted in the Table 3.

Model (%)	Accuracy	F_1-score
Baseline Model	93.64	94.05
Baseline Model + DA	**94.29**	94.20
Self-Attention + DA	91.11	90.46
MOCHA + DA	**93.86**	93.65

Table 3: Obtained results from baseline model and attention based models, where DA indicates Domain adaption settings

The baseline model and monotonic chunk-wise attention model achieved 94.05% and 93.65%, respectively as best F_1-score for domain adaption. However, after tuning the hyper-parameter for DA, learning rate (0.01 for Baseline, 0.02 for MOCHA and 0.004 for Self-attention) of these model have improved the performance. We have used Variable length of training size (200, 400, 600 and 900) for DA training by using these models that show Self-attention model performs better (94.63% F_1-score on training size of 900) than other models (94.20% and 93.65% F_1-score on training size of 900 for Baseline and MOCHA model, respectively), as illustrated in Figure 4.

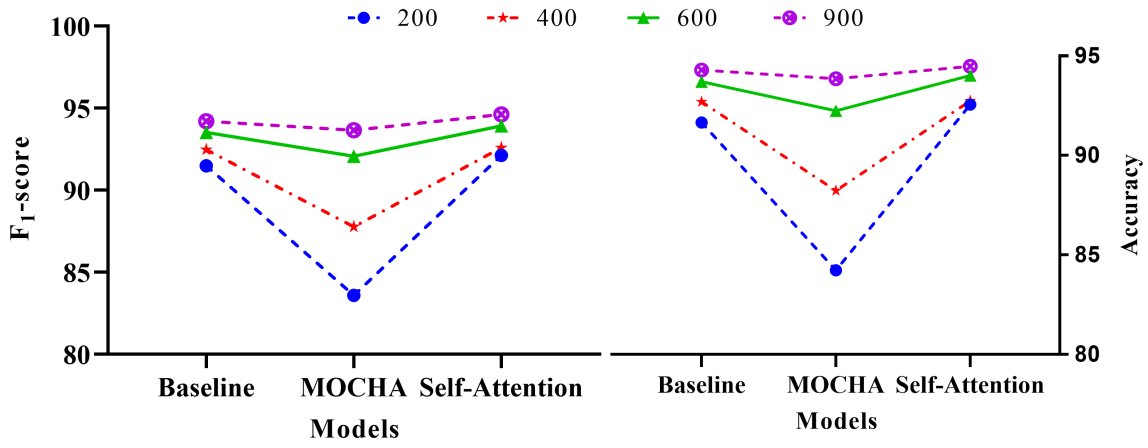


Figure 4: Accuracy and F_1-score comparison on Variable length of training data size for DA on the Baseline, Self-attention and MOCHA-based model

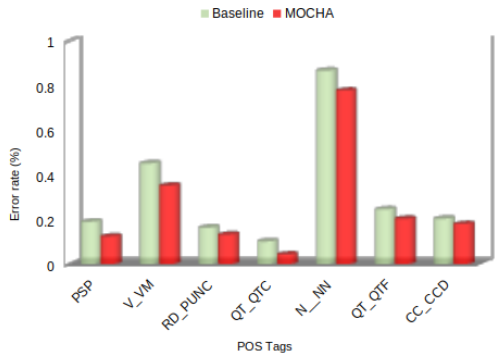


Figure 5: Error-rate comparison between selective most and less frequent POS tags obtained from predictions of baseline and MOCHA-based model

As evident from Table 3, the MOCHA-based model is precise over the baseline model. From the analysis of predictions file of the baseline model and MOCHA-based model, we found that error-rate reduced on the selective tags. Postposition (PSP), Main Verb (V_VM), Punctuation (RD_PUNC), Cardinal Quantifier (QT_QTC) and Co-ordinator Conjunction (CC_CCD), General Quantifier (QT_QTF), Common Noun (N__NN) are selective most and less frequent POS tags, respectively. These tags have reduced error rate, the difference among these are shown in Figure 5. Hence, It shows that MOCHA-based model more accurate to predictions of right POS tags on scarce words as well. On other POS tags, the error rate of the MOCHA-based model found to be comparable to the baseline model.

6 Related Work

A short chronological overview of the related work presented here to provide the context of our work. Blitzer et al. (2006) used Structural Correspondence Learning (SCL) to automatically induce correspondence to the features of a different domain in order to transfer POS tagger from Wall Street Journal (financial news) to MEDLINE (biomedical abstracts).

Collobert et al. (2011) presented a task-independent, a learning algorithm and unified convolutional neural network architecture, pertaining to various NLP tasks as POS tagging, Chunking, Named Entity Recognition and Semantic Role Labelling. They jointly trained models of POS tagging, Chunk and NER tasks with the additional linkage in trainable parameters for transferring knowledge learned in one task to another.

Zhang et al. (2014) showed type-supervised domain adaptation for the Chinese word segmentation and POS tagging, using domain-specific tag dictionaries. Unlabeled target domain dataset has improved target domain accuracy by providing annotated source domain dataset. They have obtained a 33% error reduction on target domain tagging by unlabeled sentences and a lexicon of 3000 words.

Yu et al. (2015) used an effective confidence-

based self-training approach to select additional training samples for domain adaptation of a dependency parser and were able to improve parsing accuracy for out-of-domain texts by 1.6% on texts from a chemical domain.

Mishra et al (2017) used unlabeled data for POS tagging applying for feature transfer via transfer learning from resource-rich to resource-poor language across eight Indian languages, each having 25K sentences and gained an average accuracy of 81%.

Yang et al. (2017) explored transfer learning for neural sequence tagging, where source task with large annotated dataset was exploited to enhance the performance of the target task with smaller dataset. They examined the effect of Transfer Learning on recurrent neural networks across domains, applications and languages, and obtained significant improvement.

Meftah et al. (2018) used GRU, CRF and CNN for character level feature representation as model components for POS tagging as a sequence labelling problem. To address the data scarcity, they examined the effectiveness of Cross-Domain and Cross Task Transfer Learning.

Li et al. (2019) proposed a domain embedding approach to merge the source and the target domain training data. The results demonstrated that it is more effective than multi-task learning approaches and both direct corpus concatenation (as traditional approach). Contextualized word representation with fine-tuning is used to utilize unlabeled target-domain data, which further increased its cross-domain parsing accuracy.

We have used a similar CNN architecture as proposed by Meftah et al. (2018), except that we have applied different sizes of stacked convolution layers. We have also used the same transfer settings across the domain for performing domain adaption the Hindi Treebank dataset.

Distributed word representation usually learns semantic and syntactic information about the word and ignores word size and morphological features. Part-of-speech tagging requires intra-word information when dealing with morphologically rich language.

Santos et al. (2014) have demonstrated that CNN is an effective approach for extracting morphological features and encoding it into neural representations. Singh et al. (2018) used CRF and LSTM Recurrent Neural Networks to model POS Tagging on Hindi-English Code Mixed dataset from Twitter and achieved a result of overall F_1-score of 90.20%. These works are related to our use of character level information in the models that we used.

7 Conclusion

The attention-based extended baseline model is a simple model for domain adaption to perform Part-of-Speech (POS) tagging if there is scarcity of annotated corpus. It is an extension of the LSTM-CNN-CRF model by replacing LSTM by GRU and appending attention mechanisms (self-attention and monotonic chunk-wise attention). This model was used to perform domain adaption on the Hindi Treebank dataset, where the Tourism domain was considered as the source domain and Disease as the target domain for the Transfer Learning scenario. The results show the improvement over the baseline model by the monotonic chunk-wise attention mechanism. The limitation of scarcity of annotated corpus of both of the domains can be overcome to some extent by using available pre-trained word embeddings or raw corpus to get better embeddings for this model as part of future work. In addition to this, additional linguistic information can be fused into the model to leverage the advantages of additional accessible annotations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November. Association for Computational Linguistics.

Chung-Cheng Chiu and Colin Raffel. 2017. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382*.

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.

Cicero Dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning*, pages 1818–1826.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2386–2395, Florence, Italy, July. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August. Association for Computational Linguistics.

Sara Meftah and Nasredine Semmar. 2018. A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pruthwik Mishra, Vandan Mujadia, and Dipti Misra Sharma. 2017. POS tagging for resource poor languages through feature projection. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 50–55, Kolkata, India, December. NLP Association of India.

Rudra Murthy, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2), December.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. 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)*, pages 412–418, Berlin, Germany, August. Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A twitter corpus for hindi-english code mixed pos tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 12–17.

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.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks.

Juntao Yu, Mohab Elkaref, and Bernd Bohnet. 2015. Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 1–10, Bilbao, Spain, July. Association for Computational Linguistics.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. 2019. A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. *Natural Language Engineering*, 25(1):43–67.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-supervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 588–597, Gothenburg, Sweden, April. Association for Computational Linguistics.

Understanding Transformers for Information Extraction with Limited Data

Minh-Tien Nguyen
CINNAMON LAB, 10th floor,
Geleximco building, 36 Hoang Cau,
Dong Da, Hanoi, Vietnam.
Hung Yen University of Technology and
Education Hung Yen, Vietnam.
tiennm@utehy.edu.vn

Dung Tien Le, Nguyen Hong Son,
Bui Cong Minh, Do Hoang Thai Duong,
Le Thai Linh
CINNAMON LAB, 10th floor,
Geleximco building, 36 Hoang Cau,
Dong Da, Hanoi, Vietnam.
{nathan,levi,matt,howard}@cinnamon.is
linh.le@uq.edu.au

Abstract

Transformers have recently achieved promising results in many natural language processing tasks; however, the understanding of transformers for information extraction in business scenarios is still an open question. This paper bridges the gap by introducing an investigation to understand the behavior of transformers in extracting information from domain-specific business documents. To do that, we employ transformers for taking advantage of these architectures trained on a huge amount of general data and fine-tune transformers to our down-stream IE task by using transfer learning. Experimental results on three Japanese datasets show that there are small margins among transformers in terms of F-scores but some models can achieve high accuracy with a small number of training data.

1 Introduction

The significant growth of data provides a chance for humans to approach information from many sources. Yet, it also makes an obstacle for distilling useful knowledge. To address this issue, information extraction (IE) can be considered as an appropriate solution for converting unstructured to structured data. From the research side, due to its large impact, IE has received attention from the research community with many studies (Corro and Gemulla, 2013; Angeli et al., 2015; Nguyen et al., 2019). From the business site, IE is a crucial step for digital transformation (Inmon and Nesavich, 2007; Herbert, 2017; Lin et al., 2019). The outputs of IE systems can be used in many natural language processing (NLP) applications, e.g. question answering, information retrieval (Shimaoka et al., 2016), or the automatic generation of ontology (Fleischman and Hovy, 2002).

The recent success of transformers draws a new direction for many NLP tasks. For example, BERT (Devlin et al., 2019) pioneers to creating a contextual language model for language understanding. As a result, BERT has achieved promising results on many NLP tasks, including IE. Following the success of BERT, a lot of transformer architecture has developed such as ALBERT (Lan et al., 2020), DistillBERT (Sanh et al., 2019), or ELECTRA (Clark et al., 2019). It leverages the adaptation of transformers for IE. For example, (Nguyen et al., 2019) adapted BERT to extract information from business documents. These studies achieved promising results; however, we argue that there exist gaps that limit the understanding of transformers for IE from domain-specific business documents. The first gap is that previous work only investigates the IE task with one transformer model, e.g. BERT. The second gap is that several important aspects of transformers were not studied well, e.g. the relationship between the number of training samples and performance.

This paper bridges the two gaps by investigating the behavior of transformers for extracting information from business documents, in actual scenarios. To do that, we empower IE models by using transformers in the form of transfer learning. Precisely, transformers are used to utilize the power of these models trained on the huge amount of general data. Then the transformer-based IE models are fine-tuned in the downstream IE task. By using transform-

ers for transfer learning, we simulate actual business cases that have a small number of training data. This paper makes three main contributions:

- It analyzes the behavior of transformers for IE in the context of business scenarios. The analysis examines the transformers in three aspects: performance comparison, the relationship between performance and the number of training samples, and training time. To the best of our knowledge, we are the first conducting the comprehensive investigation for IE from domain-specific business documents in a low-resource language, i.e. Japanese.
- It introduces a public dataset[1] for the IE task of business documents. The dataset mimics actual business cases in which IE models are trained with a small number of training data.
- It releases a pre-trained model[2] based on ELECTRA (Clark et al., 2019), which facilitates studies of NLP tasks on Japanese.

2 Related Work

Information extraction is an important task of NLP and has investigated in a long time with many studies. There are two main approaches for IE, using dictionaries (Watanabe et al., 2007) and machine learning (Corro and Gemulla, 2013; Angeli et al., 2015; Lample et al., 2016). The first approach usually defines a dictionary for extracting information. Input documents are parsed to tokens which are matched to each item in the dictionary for extraction. The second approach usually uses training data to train a classifier that can distinguish extracted or non-extracted information (Corro and Gemulla, 2013; Angeli et al., 2015; Lample et al., 2016). Using a dictionary-based method can achieve high accuracy, but it is time-consuming and labor-expensive for dictionary preparation. In contrast, machine learning models exploit linguistic features (Angeli et al., 2015) or hidden features learned from LSTM for classification (Lample et al., 2016). As a result, it can reduce the cost of dictionary maintenance and easy to adapt to other domains. In practice, several research projects focus on the nested named entities and have great progress so far (Finkel and Manning, 2009; Lample et al., 2016).

NER is a specific task of IE in which high-level concepts such as people, places, organizations usually need to extract. For example, CoNLL 2003 defined four types of entities, including locations, mixed entities, organizations, and persons (Sang et al., 2003). However, for document analysis in practical business cases, entity types should be at a more detailed level (Corro et al., 2015; Nguyen et al., 2019). To address this problem, fine-grained entity extraction was introduced and applied to several NLP applications such as question answering, information retrieval (Lee et al., 2006; Shimaoka et al., 2016), or the automatic generation of ontology (Fleischman and Hovy, 2002). The recent success of transformers draws a new method for NER. BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), DistillBERT (Sanh et al., 2019), and ELECTRA (Clark et al., 2019) are four pre-trained transformers which achieve promising results many NLP tasks. This paper employs the power of those transformers as transfer learning for our IE problem. This employment allows us to simulate our business cases which only have a small number of training samples.

The work of (Nguyen et al., 2019) is perhaps the most relevant to our task. In this paper, the authors adapted BERT for extracting information from domain-specific documents. However, understanding the behavior of BERT in terms of extracting from business documents is still an open question. We dig a deeper level to observe IE models by comparing four transformers. We believe that this comparison provides a comprehensive analysis of transformers for such IE task in actual business cases.

3 Task Definition and Data Preparation

3.1 Task definition

As mentioned, we deal with the task of IE with limited data for business documents. Given a document and pre-defined tags (keywords), IE models need to extract corresponding information to the tags. Formally, the task can be formally defined as follows.

- **Input**: a document and a set of tags.
- **Output**: extracted information corresponding to the tags.

[1] https://github.com/DungLe13/bidding-dataset

[2] https://github.com/thaiduongx26/electra_japanese

Our IE task is quite different from the common NER task in which we need to extract a large number of entity types, e.g. 24 (Table 1) while the common NER task extracts a small number of entities, e.g. four types of CoNLL. Also, due to the restriction of actual business cases, we use a small number of training samples instead of using a large number of training examples e.g. around 15,000 samples in CoNLL (Sang et al., 2003).

3.2 Data preparation

It is hard to use published datasets, e.g. CoNLL (Sang et al., 2003) for comparison due to our different purpose with common NER tasks. We, therefore, prepared three datasets, for testing IE models.

3.2.1 CinData

Because there are gaps in using common IE datasets to our task, we created a new corpus named CinData. To do that, we collected 124 public Japanese bidding documents from the Japan Oil, Gas and Metals National Corporation (JOGMEC).[3] Each document is a public notice, which outlines the information about the bidding process, including the dates of the contract, the deadlines for submission, and the contacts of the department or person in charge. These documents are raw texts, so we need to define a set of tags for the annotation process. To do that, we consulted our legal team for the definition. The discussion and definition were internally conducted. Finally, we defined 19 names that represent the categories of extracted information, which we formally refer to as ``tags". The list of tags covers common important information of a bidding document. The list is unique and remains unchanged in all three train/dev/test sets. Please refer to the Appendix for the description of tags.

The collected documents are `PDF` files, so they were converted to the text format for easy use. To do that, we used `pdfplumber`,[4] as a parser, combined with heuristic rules: bullets, numberings, indentication, title, table for keeping the structure of documents. After parsing, our QAs (quality assurance - people who have at least the N3 Japanese-Language Proficiency Test certificate, with N1 is the highest level) checked and corrected errors of outputs.

[3]http://www.jogmec.go.jp/news/bid/search.php

[4]https://github.com/jsvine/pdfplumber

The annotation was internally conducted with two annotators in two steps. In the first step, each annotator was assigned a set of documents. With each document, the annotator read predefined tags and assigned start and end positions for corresponding segments. The second step is cross-validation, in which documents were cross-checked and corrected based on the negotiation of the annotators. The agreement computed by Cohen Kappa[5] of two annotators is 0.8275 (before correction), showing that the annotators have a high agreement in annotating data.

3.2.2 Bidding and sale documents

To have a better assessment of IE models, we prepared two other datasets used internally in our company. The first contains bidding documents in different domains compared to the CinData. The second includes sale documents of hardware devices. Due to the policy, we can not disclose these datasets.

3.2.3 Data observation

Table 1 shows statistics of the three datasets. As observed, the number of training samples is small. It supports the point that in business cases, having a large number of training data is a big obstacle. In this sense, we also simulated our dataset with limited training samples. The documents are quite long, with a quite large number of sentences and characters per sample. A large number of entity types, e.g. 19 or 24 also challenges IE models.

Statistics	CinData	Bidding docs	Sale docs
#training docs	82	78	300
#dev docs	22	-	-
#testing	20	22	165
#chars/doc	3,030	22,537	2,083
#sentss/doc	120	616	56
# of tags	19	24	8

Table 1: Data observation on three datasets.

4 Extraction with Transformers

This section introduces the IE models based on transformers. We first describe transformers and then show transfer learning, information extraction, and the training process of the models.

[5]http://graphpad.com/quickcalcs/kappa1.cfm

4.1 Transformers

The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution (quoted from (Vaswani et al., 2017)). The Transformer complies with the overall architecture of encoder-decoder using stacked self-attention and point-wise, fully connected layers. The attention function of the transformer is computed by mapping a query and a set of key-value pairs to an output. Then, the output is computed as a weighted sum of the values, where the weight of each value is computed by a compatibility function of the query with the correlated key.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where dimension d_k of keys, and dimension d_v of values Moreover, Transformer performs the attention function in parallel, resulting d_v-dimensional output values using "multi-head attention" as following: $MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$ where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

This paper investigates the IE task with four transformer-based models: BERT, ALBERT, DistilBERT, and ELECTRA. We selected BERT because it pioneers the transformer direction (Devlin et al., 2019), after that, its variation also achieves promising results. For ELECTRA, it is up-to-date architecture that obtains improvements compared to the BERT family (Clark et al., 2019).

BERT BERT, introduced by (Devlin et al., 2019), was the state-of-the-art model for many benchmark datasets in multiple NLP tasks. It utilized the bidirectional pre-training to represent a language as dense and low-dimensional vectors. The model is pre-trained using two unsupervised tasks, namely masked language modeling and next sentence prediction. In Masked Language Modeling, 15% of all WorkPiece tokens in a sequence are either (i) replaced with a [MASK] token, or (ii) replaced with a random token, or (iii) remained the same. By learning to predict the masked tokens, the model learns the representation of tokens in association with the context surrounding it. In Next Sentence Prediction, the model learns the relationships between two sentences by predicting whether sentence B follows sentence A in a sequence.

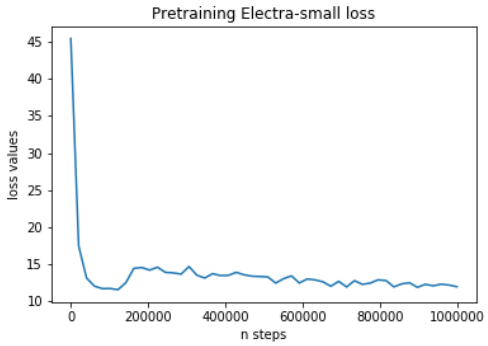
DistilBERT DistilBERT leverages knowledge distillation, in which a compact model - the student - is trained to reproduce the performance of a large model - the teacher (Sanh et al., 2019). Following this setting, the student - DistilBERT, which has the same architecture as BERT but fewer layers learn to perform pre-trained tasks by mimicking the output distribution of the teacher network - the original BERT model. The model uses the triple loss, which combines the losses of the masked language model, distillation, and cosine-distance. It also follows the practice of previous variations of BERT-based models by using dynamic masking and omitting the next sentence prediction objective.

ALBERT ALBERT is a lighter version of the original BERT, which incorporates two important techniques to reduce the number of parameters used in the model (Lan et al., 2020). The first one is a factorized embedding parameterization. Instead of projecting the one-hot vectors into a high-dimensional hidden space of size `H`, the model decomposes this step into two smaller steps. It first projects these vectors into a lower-dimensional embedding space size `E`, and then projects it into the hidden space. This reduces the embedding parameters significantly when `E << H`. The second technique is cross-layer parameter sharing, where all parameters are shared across multiple layers. This prevents the number of parameters from growing as the number of layers increases. In addition to the aforementioned techniques, ALBERT also employs an inter-sentence coherent loss in the replacement of the next sentence prediction task during the pre-training process.

ELECTRA ELECTRA is a replaced token detection method that trains a discriminative model predicting whether each token in the corrupted input could be replaced by a generator sample (Clark et al., 2019). Compared to BERT and its variations, ELECTRA makes two important differences. First, instead of training a `[MASK]` language model trained on the small subset that was masked, ELECTRA trains a language model on all input tokens. Second, ELECTRA was trained in a discriminative fashion to predict whether each token in the cor-

rupted input was replaced by a generator sample or not rather than predicting the original identities of the corrupted tokens. In addition, ELECTRA makes an important consideration for pre-training methods that should be efficiently computed without large amounts of data.

We employed the success of ELECTRA (Clark et al., 2019) to our IE task. Since ELECTRA is only for non-Japanese languages, we trained the Japanese ELECTRA model for our purpose. To do that, we collected Japanese Wiki-data then used the code of ELECTRA[6] for training an ELECTRA-small model. The difference compared to the original model is that we used SentencePiece instead of WordPiece because it is hard to apply word segmentation to Japanese. The idea of SentencePiece[7] bases on subword units and unigram language model, which help us to train our ELECTRA without any language-specific pre- and post-processing. More importantly, SentencePiece allows our ELECTRA to extend the vocabulary which is benificial for the training process. The size of our vocabulary for Japanese-wiki is 32,000. The pre-training task of Electra-small took 6 days with 1M steps by using a single GPU Radeon VII 16GB. The following figure shows the loss during the training process.



After training, we applied the model to the datasets. The idea is similar to BERT-QA, in which we fed hidden representation from ELECTRA to an MLP for classification.

4.2 Transfer learning

Transformers provide an appropriate solution for data representation by using contextual embeddings learned from a large amount of data. However, they should be adapted to downstream tasks by using training data in specific domains. To do that, we fine-tuned the models to the downstream IE task by using the samples data of each dataset. The pre-trained weights of transformers were first reused and then adjusted in the fine-tuning process.

4.3 Information extraction

Output vectors from the transfer learning layer were put into the extraction layer for extracting information. To do that, the extraction was formulated as a question answering (QA) task, thanks to the suggestion of BERT (Devlin et al., 2019). A question (tag) and corresponding segment were fed into transformers to learn hidden representation. The extraction predicts start and positions based on the probability of the word i in this span. The final score of a potential answer spanned from position i to position j defined as $\max_{i,j}(S\Delta T_i + E\Delta T_j)$ with $j \geq i$.

$$P_{start_i} = \frac{e^{S.T_i}}{\sum e^{S.T_j}}; \qquad P_{end_i} = \frac{e^{E.T_i}}{\sum e^{E.T_j}} \quad (2)$$

The extraction uses the positions $start$ and end to extract information corresponding to input tags.

4.4 Training

We used a multilingual BERT-base model trained for 102 languages (including Japanese) on a huge amount of texts from Wikipedia (Devlin et al., 2019). The BERT model has 12 layers, a hidden layer of 430, 768 neurons, 12 heads. For DistillBERT, we used a multilingual model pretrained with the supervision of BERT-base-multilingual-cased on the concatenation of Wikipedia in 104 different languages. The model has 6 layers, 768 dimensions, and 12 heads. For ALBERT,[8] we used the pre-trained Japanese model with 12 layers, the hidden size of 768, and the embedding size of 128. For ELECTRA, we used our pre-trained model trained on Japanese Wiki data. The ELECTRA-small has 12 layers, with a hidden size of 256.

Thanks to the suggestion of BERT, we formulated the training process as a QA task. Tags and corresponding segments were fed into the models for

[6]https://github.com/google-research/electra

[7]https://github.com/google/sentencepiece#comparisons-with-other-implementations

[8]https://huggingface.co/ALINEAR/albert-japanese-v2

learning. The training was done in two steps: pre-training and fine-tuning. For the first step, the pre-trained weights of transformers were reused, while the weights of the rest layers were generated with a truncated normal distribution. All models were fine-tuned in 20 epochs by using the cross-entropy loss function between predicted and correct information. The training process was done with a single GPU.

5 Settings and Evaluation Metrics

Settings We used training samples in Table 1 for training IE models and applied the model on the test sets. Due to our investigation purpose, we did not fine-tune IE models by using the development set of CinData. Instead of doing that, we report the performance on this set. For transformers, Table 2 summarizes its information. All models were trained by using the same data segmentation, settings, and GPU.

Model	Layers	Parameters
BERT-base	12	110M
DistillBERT	6	66M
ALBERT	12	12M
ELECTRA	12	14M

Table 2: Information of transformers.

As observed, BERT and DistillBERT have a large number of parameters while ALBERT and ELECTRA are significantly compressed.

Evaluation metrics Extracted information was matched with correct answers for computing F-scores based on precision and recall metrics. The F-score of a model on a dataset is the average of F-scores on all tags computed by fields.

6 Results and Discussion

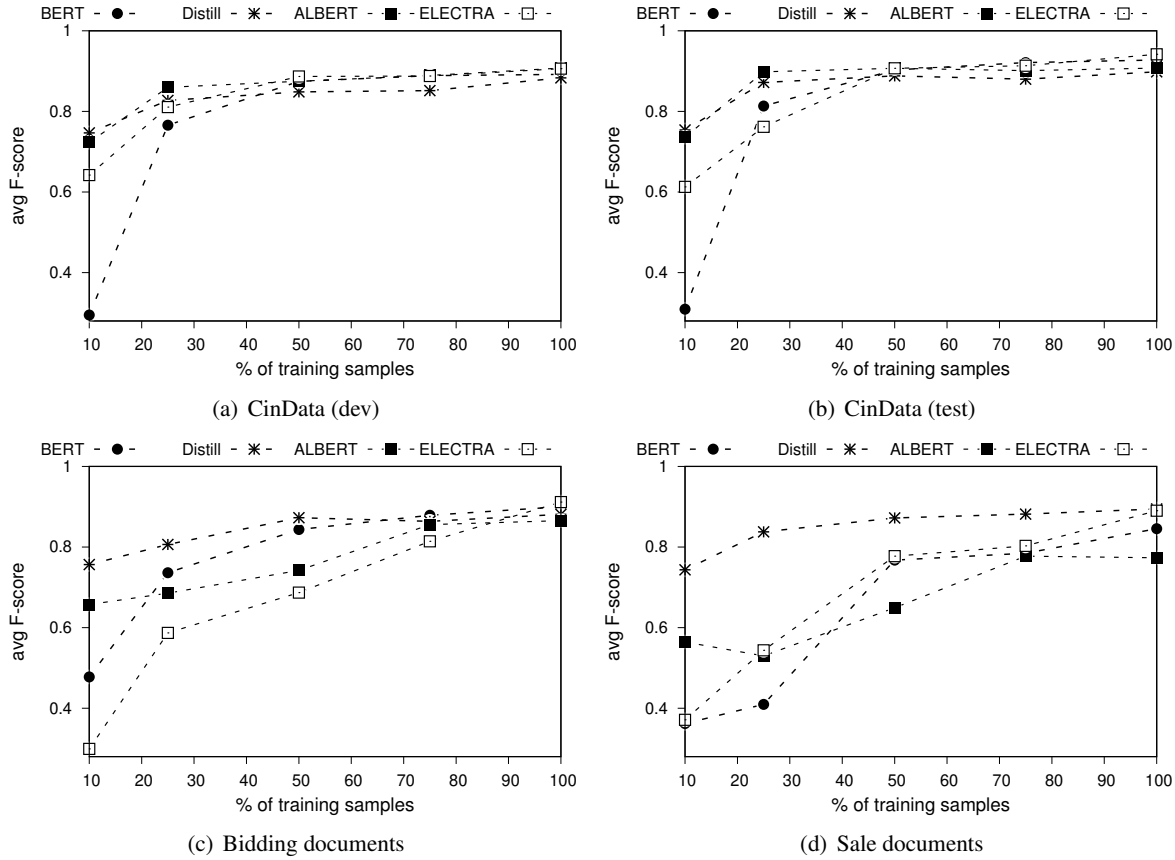
6.1 F-scores Comparison

Table 3 summarizes the comparison of transformer-based IE models on four datasets. As we can observe that the IE models based on BERT and ELECTRA achieve promising results. For example, the model of BERT is the best in two cases (CinData (dev) and CinData(test)) and ELECTRA obtains the highest F-score on the bidding dataset. For BERT, it is understandable that it has the largest model which includes 110M parameters. This enables BERT to capture the context of words from the input (the relationship between a tag-segment pair). As a result, the IE model using BERT achieves promising results. An interesting point comes from ELECTRA. It is a small model with 14M parameters, compared to BERT (110M) and DistilBERT (66M); however, the ELECTRA-based IE model outputs competitive F-scores on four datasets. For example, the IE model using ELECTRA is better than BERT of 1.15 F-score on bidding documents (0.9115 vs. 0.9000), which is the most challenging dataset with very long documents. The possible reason comes from the training process of ELECTRA that can contribute to the ELECTRA-based IE model. As mentioned in Section 4.1, we used SentencePiece instead of WordPiece due to the word segmentation of Japanese. This is different from BERT, DistilBERT, and ALBERT which used WordPiece for Japanese. The promising F-scores of ELECTRA with a small pre-trained model draw a new direction for adapting transformers to our IE task and confirm the results of ELECTRA (Clark et al., 2019).

Method	CinData (dev)	CinData (test)
BERT (QA)	**0.8887**	**0.9175**
DistilBERT	0.8831	0.8983
ALBERT	0.8585	0.8926
ELECTRA	*0.8879*	*0.9133*
Method	**Bidding docs**	**Sale docs**
BERT (QA)	0.9000	0.8456
DistilBERT	0.8811	**0.8944**
ALBERT	0.8655	0.7734
ELECTRA	**0.9115**	*0.8901*

Table 3: Comparison of methods according the average of F1-score. **Bold** is the best and *italic* is the second best.

The extension of BERT does not show the best performance on four datasets. For example, DistilBERT is only the best on sale documents with tiny margins compared to other models, even it is the second larger model (66M parameters). It is understandable that DistilBERT tries to compress the model size while approximating the performance with BERT. In other cases, DistilBERT and ALBERT output lower F-scores than BERT and ELECTRA. A possible reason comes from the size of the model. For instance, ALBERT obtains the lowest F-



(a) CinData (dev)

(b) CinData (test)

(c) Bidding documents

(d) Sale documents

Figure 1: F-scores and the number of training examples.

scores in all cases, e.g. 0.7734 on sale documents because it only has 12M parameters, which are hard to cover all the semantic aspects of the datasets.

6.2 F-scores and Training Samples

We observed the behavior of transformers regarding the number of training examples. This is because we would like to understand when the transformers can achieve good results. To do that, we randomly segmented data into several parts, ranging from 10%, 20%, 50%, 75%, and 100% and observed the F-scores at each data segment. Figure 1 visualizes the observation of BERT, ALBERT, DistillBERT, and ELECTRA wit different data segments.

As we can observe, the number of training samples affects the quality of transformer-based IE models. The general trend shows that adding more training examples increases F-scores. However, the behavior of transformers is different. For example, on CinData, F-scores significantly raise from 10% to 25% of training data and reach the top at 50% of training data. After that, the F-scores slightly grow. This indicates that for CinData, transformers only need 50% of data to obtain stable performance. For biddings and sales, the trend is quite different. For biddings in Figure 1(c), two strong models (BERT and ELECTRA) share the similar behavior, in which its F-scores dramatically increase from 10% to 75%. After that, the F-scores are stable. It is explainable that adding more data helps to improve the quality of BERT and ELECTRA-based IE models. In contrast, DistilBERT and ALBERT have the same trend, in which these models obtain quite high results at 10% and steadily raise until 75%. The trend on sale documents in Figure 1(d) is quite diverse, in which the behavior of DistilBERT is the same on bidding and sale documents. BERT and ELECTRA have significant improvements from 10% to 50% while ALBERT reaches the top at 75%. It is interesting to observe that by using a small number of data, Distil-

BERT seems to be better than others on bidding and sale documents in Figures 1(c) and 1(d). This suggests two use cases: (i) if we only have some dozens of data, e.g. 50-100 samples, DistilBERT can be a good option and (ii) otherwise, BERT and ELECTRA are appropriate the selection.

6.3 Training Time

We observed the training time of transformers with the same data segmentation of Section 6.2. Figures 2, 3, and 4 plots the observation.

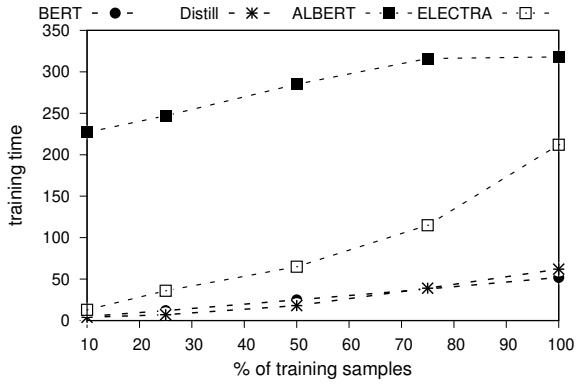


Figure 2: Training time (minutes) on CinData.

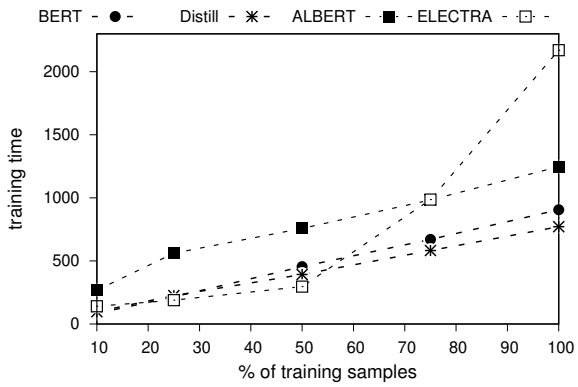


Figure 3: Training time (minutes) on bidding documents.

It is interesting to observe that BERT and DistilBERT are the fastest even they have the largest models with a huge of parameters. A possible reason is that with a large number of parameters, these models do not need to learn so much from the data of new domains. As a result, they are quick to be covered in the training process. In contrast, ALBERT and ELECTRA take a long time to complete the training process. For example, ALBERT needs 300 minutes

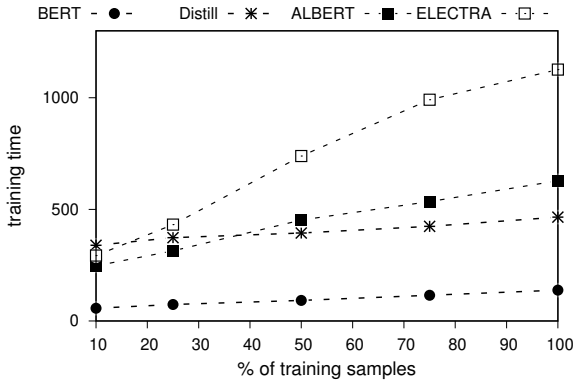


Figure 4: Training time (minutes) on sale documents.

for 100% data on CinData. Also, the small number of parameters seems to be efficient for inference only. For training, the computation operation is not so much different among the four transformers. As a result, ALBERT and ELECTRA took a longer time than BERT and DistilBERT for training.

7 Conclusion

This paper introduces an investigation of transformers for information extraction with limited data. The investigation simulates business scenarios that have small numbers of training data to build IE models. To do that, we employ four well-known transformers for taking advantage of the contextual aspect learned on huge data and fine-tune to our down-stream IE tasks by using transfer learning. Experimental results on three domain-specific business datasets confirm the efficiency of BERT and ELECTRA, that can be applied to actual business cases. The observation of training samples indicates that in some cases, transformers can achieve good results with 50% of training data. The training time shows that BERT is potential while ALBERT and ELECTRA take a long time when training with all data.

For future direction, we encourage to deeply investigate sophisticated models for the IE task, e.g. stacking transformers with refined architecture.

Acknowledgments

We would like to thank Gaku Fujii, Shahab Sabahi, Akira Shojiguchi, and reviewers for useful comments and discussion. This research is funded by Hung Yen University of Technology and Education under the grant number UTEHY.L.2020.04.

References

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 344-354*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *Proceedings of International Conference on Learning Representations*.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web, pp. 355-366*.

Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. Finet: Context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 868-878*.

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*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Volume 1-Volume 1, pp. 141-150. Association for Computational Linguistics.*

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistic, Volume 1, pp. 1-7. Association for Computational Linguistics.*

Lindsay Herbert. 2017. Digital transformation: Build your organization's future for the innovation age. Technical report, Bloomsbury Publishing.

Bill Inmon and Anthony Nesavich. 2007. *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence*. Pearson Education.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of The International Conference on Learning Representations*.

Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang. 2006. Fine-grained named entity recognition using conditional random fields for question answering. In *Asia Information Retrieval Symposium, pp. 581-587. Springer, Berlin, Heidelberg*.

Jerry Chun-Wei Lin, Yinan Shao, Yujie Zhou, Matin Pirouz, and Hsing-Chung Chen. 2019. A bi-lstm mention hypergraph model with encoding schema for mention extraction. *Engineering Applications of Artificial Intelligence 85: 175-181*.

Minh-Tien Nguyen, Viet-Anh Phan, Le Thai Linh, Nguyen Hong Son, Le Tien Dung, Miku Hirano, and Hajime Hotta. 2019. Transfer learning for information extraction with limited data. In *Proceedings of 16th International Conference of the Pacific Association for Computational Linguistics, pp. 469-482*.

Erik Sang, Tjong Kim, 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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *arXiv preprint arXiv:1910.01108*.

Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2016. An attentive neural architecture for fine-grained entity type classification. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, pp. 69-74*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems, pp. 6000–6010*.

Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. 2007. A graph-based approach to named entity categorization in wikipedia using conditional random fields. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 649-657*.

Appendix A. The Name of Tags

English Tag	Type
Year of Procurement	datetime (year only)
Prefecture	text
Bid Subject	text
Facility Name	text
Address for Demand	text
Start Date of Procurement	datetime
End Date of Procurement	datetime
Public Announcement Date	datetime
Deadline for Questionnaire	datetime
Deadline for Applying Qualification	datetime
Deadline for Bidding	datetime
Opening Application Date	datetime
PIC for Inquiry of Questions	text
TEL/FAX for Inquiry of Questions	text
Address for Submitting Application	text
Department/PIC for Submitting Application	text
Address for Submitting Bid	text
Department/PIC for Submitting Bid	text
Place of Opening Bid	text

Table 4: Extracted information of CinData.

English Tag	Type
Model code	mixed
Model name	mixed
Start of sales	date
End of sales (planed)	date
End of sales (fixed)	date
End of sales (special)	date
End of support	date
Revision	mixed

Table 5: Extracted information of sale documents.

English Tag	Type
Year of procurement	datetime (year only)
Prefecture	text
Title of bidding	text
Name of institution	text
Address for demand	text
Start date of procurement	datetime
End date of procurement	datetime
Contract value	number
Amount of value	number
Class of reserved value	number
Amount of reserved value	number
Public Announcement Date	datetime
Deadline for delivery specification	date
Deadline for questionnaire	datetime
Deadline for applying qualification	datetime
Deadline for bidding	datetime
Opening application date	datetime
PIC for inquiry of questions	text
TEL/FAX foriInquiry of questions	tel/fax
Address for submitting application of qualification	address
Address of submitting of bidding applications	address
Department/PIC for submitting application	name
Place of Opening Bid	text

Table 6: Extracted information of bidding documents.

A Study on Seq2seq for Sentence Compression in Vietnamese

Thi-Trang Nguyen[1,2], **Huu-Hoang Nguyen**[2], **Kiem-Hieu Nguyen**[1]

[1]Hanoi Uni. of Science and Technology, Hanoi, Vietnam
[2]VCCorp, Hanoi, Vietnam
{nguyenthitrang03,hoangnguyenhuu}@admicro.vn, hieunk@soict.hust.edu.vn

Abstract

Text summarization is an important yet challenging task in natural language processing. In this paper, we investigate Pointer Generator Networks for sentence compression. Using Vietnamese as a case study, our model could yield sentence summaries with high quality of syntax, factual correctness and completeness. Interestingly, we demonstrate that only a simple filtering technique is required to generate training data of sentence-summary pairs without any human annotation

1 Introduction

Text summarization is an important and challenging aspect of natural language processing. The ultimate goal is to generate summary that retains essential information of the original text. Extractive approaches attempts to identify salient parts of the source text and assembles them into a summary. In contrast, abstractive approaches uses language modelling technique which is conditioned on original text to generate summary that might have different sentence structure and novel words. This paper focuses on sentence compression, which could be used in both abstractive and extractive summarization systems or as a stand-alone application. Sentence compression methods could be broadly classified into two categories, deletion-based and abstractive models. Deletion-based approach depends upon efforts to find and delete unimportant words or phrases in the original sentence. A shorter sentence is then produced by stitching together remaining fragments. An abstractive sentence compressor, on the other hand, is essentially similar to a document or multi-document summarizer except that it only takes the original sentence as context and it produces one sentence instead of a multi-sentence summary.

Recent success of sequence-to-sequence (seq2seq) framework in machine translation paves the way for the emergence of neural abstractive summarization [Chopra et al.2016, Wubben et al.2016, Nallapati et al.2016, See et al.2017, Rush et al.2015]. Although there have been several studies in Vietnamese text summarization [Nguyen and Nguyen2011, Dac et al.2017], we haven't been aware of any neural networks-based methods, partially due to the lack of a large volume of high-quality training data.

In a newswire article, we observe that sapo[1] is typically a longer version of title. Based on this observation, we could collect a large amount of sentence - summary pairs from online newswire. It then requires a simple filtering step based on textual similarity to remove unwanted title - sapo pairs. We rely on a particular model in seq2seq family, i.e. Pointer Generation Networks [See et al.2017] to learn summarizing from a large amount of training pairs.

As suggested in [Kryscinski et al.2019], we evaluate generated summaries using human judgement alongside ROUGE metric [Lin2004]. Despite inevitable noise in training data, our model could yield summaries with high quality of syntax, factual correctness and completeness.

[1]Sapo is a paragraph, usually containing one sentence, that follows the title and precedes the first sentence of an article.

2 Related work

There are two main approaches to sentence compression: deletion-based and abstractive approach. In deletion-based approach, one has to decide whether to keep or remove each token in the original sentence [Jing2000, Clarke and Lapata2008, Fevry and Phang2018, Wang et al.2017, Galanis and Androutsopoulos2010]. An early work of Jing [Jing2000] applied linguistic rules and lexicon to parse tree to remove non-salient phrases from the original sentence. Recently, Filippova [Filippova et al.2015] adopted an LSTM network to resolve the problem as sequence labeling. Following this direction, Wang et al [Wang et al.2017] show that syntax could be useful for LSTM-based sentence compression.

Abstractive approach to sentence compression is more powerful in that it considers all the operations including deletion, reordering, substitution and insertion. Cohn and Lapata [Cohn and Lapata2008] learn a set of parse tree transduction rules from a training dataset of sentence-summary pairs. Recently, seq2seq framework has been studied for this task [Chopra et al.2016, Nallapati et al.2016, See et al.2017]. Seq2Seq realizes encoder-decoder paradigm where the encoder encodes an input sequence into hidden states, from which the decoder then generates the output sequence. Taken together, the attention mechanism automatically align input and output sequences, that boosts the performance of seq2Seq significantly.

Besides just that, there are researches in unsupervised sentence compression. In [Fevry and Phang2018], Fevry and Phang train a denoising auto-encoder to recover the original, constructing an end-to-end training regime. In [Baziotis et al.2019], Baziotis et al present a seq2seq autoencoder which consists of two chained encoder-decoder pairs. It learns to restore the original sentence while forcing the middle hidden sequence to generate important information in the sentence, thus generate its summary without parallel training data.

There are several studies on sentence compression in Vietnamese. Nguyen [Nguyen and Horiguchi2003] learns transition rules from example pairs to generate summaries in English and Vietnamese from an original English sentence. In another work, the author [Nguyen et al.2004] utilized Hidden Markov Models for deletion-based sentence compression. In [Nguyen and Nguyen2011], Nguyen et al applied Viterbi decoding to find the most likelihood substrings and then concatenate them to generate compression. In [Tran et al.2015], Tran proposed Conditional Random Fields using information on meaningful chunks as feature. In [Tran and Nguyen2018], Tran proposed a three-phase method for summarizing paragraph. It first builds a graph to represent input content with coreference resolution. The graph is then transformed into abstract semantic representation. New sentence is finally generated from this representation.

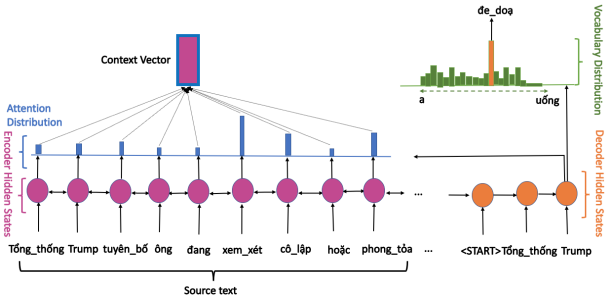


Figure 1: Vanilla Seq2seq with attention

3 Seq2seq models

In this section, we briefly show (1) baseline Seq2seq model, (2) pointer-generator model, and (3) coverage mechanism that can be added to either of the first two models. The original paper contains far more details and in-depth specifications [See et al.2017].

3.1 Vanilla seq2seq

A vanilla seq2seq framework for abstractive summarization is composed of an encoder and a decoder with attention mechanism. The encoder is a single-layer bidirectional LSTM. In an attention-based encoder-decoder architecture (shown in Figure 1), the decoder (a single-layer unidirectional LSTM) not only takes the encoded representations of the source sequence as input, but also selectively focuses on parts of the sequence at each decoding step. On the other hand, that tells the decoder where to look up to produce the next word. The attention distribution is calculated as in below:

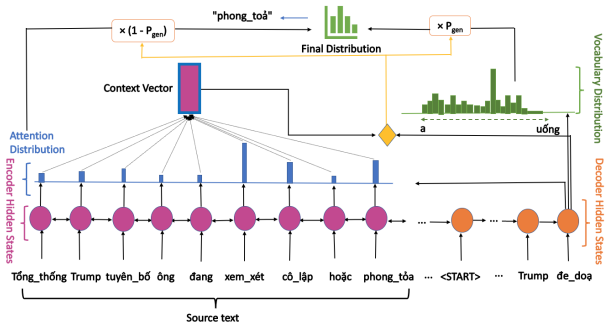


Figure 2: Pointer-generator networks.

$$e_i^t = \boldsymbol{\nu}^T \tanh\left(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + b_{attn}\right) \quad (1)$$

$$\boldsymbol{a}^t = softmax(\boldsymbol{e}^t) \quad (2)$$

where $\boldsymbol{h_i}$ is encoder hidden state after feeding the token w_i to the encoder network, $\boldsymbol{s_t}$ is generated by the decoder when receiving the previous word representation at each step t. $\boldsymbol{\nu}$, $\boldsymbol{W_h}$, $\boldsymbol{W_s}$ and b_{attn} are learnable parameters.

$\boldsymbol{P_{vocab}}$, probability distribution over all words in the decoder vocabulary, is produced by concatenating the context vector $\boldsymbol{h_i^*}$ with the decoder state $\boldsymbol{s_t}$ and feeding through two linear layers. Details are described in the following formulas:

$$\boldsymbol{h}_t^* = \sum_i \boldsymbol{a}_i^t \boldsymbol{h}_i \quad (3)$$

$$\boldsymbol{P}_{vocab} = softmax\left(\boldsymbol{V}'\left(\boldsymbol{V}[\boldsymbol{s}_t, \boldsymbol{h}_t^*] + b\right) + b'\right) \quad (4)$$

where $\boldsymbol{V}$, $\boldsymbol{V}'$, b and b' are learnable parameters.

The overall loss for whole sequence is shown below:

$$loss = \frac{1}{T}\sum_{t=0}^{T} -\log \boldsymbol{P}_{vocab}(\boldsymbol{w}_t^*) \quad (5)$$

where the loss for each timestep t is the negative log likelihood of the target word $\boldsymbol{w}_t^*$ for that time step

3.2 Pointer-generator network

Because the baseline model are restricted to their pre-set vocabulary, the ability to generate out-of-vocabulary words (OOV) is one of the primary advantages of pointer-generator models.The pointer-generator network allows the model to generate tokens by copying from the input sequence. The architecture of pointer-generator is described on figure 2. It is equipped with a "soft-switch", decoder vocabulary or point to one in the source article at each decoding step. The soft-switch is explicitly modeled by

$$p_{gen} = \boldsymbol{\sigma}\left(\boldsymbol{w}_{h^*}^T \boldsymbol{h}_t^* + \boldsymbol{w}_s^T \boldsymbol{s}_t + \boldsymbol{w}_x^T x_t + b_{ptr}\right) \quad (6)$$

where vectors $\boldsymbol{w}_{h^*}$,$\boldsymbol{w}_s$,$\boldsymbol{w}_x$ and scalar b_{ptr} are learnable parameter and $\boldsymbol{\sigma}$ is the sigmoid function. For each sequence let the *extended vocabulary* denote the union of the vocabulary, and all words appearing in the source sequence. The vocabulary distribution over an extended vocabulary is calculated by:

$$\boldsymbol{P}(w) = p_{gen}\boldsymbol{P}_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} \boldsymbol{a}_i^t \quad (7)$$

Note that if w is OOV, $P_{vocab}(w)$ is zero; if w does not appear in the source sequence, then $\sum_{i:w_i=w} \boldsymbol{a}_i^t$ is zero. The loss function is described similarly on the seq2seq attention models as equations (5) and (6), but with respect to probability distribution $\boldsymbol{P}(w)$ given in equation (8).

3.3 Coverage mechanism

Coverage model was created to solve repetition problem for seq2seq models. In this model, they first defined a *coverage vector* $\boldsymbol{c}^t$ as the sum of attention distributions of the previous decoding steps:

$$\boldsymbol{c}^t = \sum_{t'=0}^{t-1} \boldsymbol{a}^{t'} \quad (8)$$

Thus, it contains the accumulated attention information on each token in the source sequence during the previous decoding steps. Note that c^0 is zero vector, because none of the source sequence has been covered on the first timestep. The coverage vector

will then be used as an extra input to the attention mechanism, changing (1) to:

$$e_i^t = \boldsymbol{\nu}^T \tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{w}_c \boldsymbol{c}_i^t + b_{attn}) \quad (9)$$

where $\boldsymbol{w}_c$ is a learnable parameter vector of same length as $\boldsymbol{\nu}$

Finally, the *coverage loss* $covloss_t$ is reweighted by some hyperparameters λ , is added to primary loss function (5) to yield a new loss function:

$$covloss_t = \sum_i min(\boldsymbol{a}_i^t, \boldsymbol{c}_i^t) \quad (10)$$

$$loss_t = -\log P(\boldsymbol{w}_t^*) + \lambda \sum_i min(\boldsymbol{a}_i^t, \boldsymbol{c}_i^t) \quad (11)$$

4 Experiments

In all experiments, we used pytorch implementation of [See et al.2017][2].

4.1 Data preparation

Manually building a large dataset really requires a lot of commitment from annotators and a great deal of time and effort. This would appear to be infeasible. In this section, we explain how we obtained a corpus of sentences and their compressions.

The underlying idea is to harvest news articles from the Internet where the sapo appears to be an extension of the title of an article, and vice versa, the tile appears to be a compression of the sapo. Using a news crawler, we collected a large number of news items in Vietnamese. Word segmentation was processed by UETSegmenter[3]. From every article, we examined the correlation between sapo (S) and title (T). For each pair of (S, T), we used some filters and a simple scoring function to decide which pairs should be retained:

- The number of words in S must be greater than 25 (excluding punctuation)
- The number of words in T is within (8, 15] (excluding punctuation)
- The word duplicate rate - d(S, T) is greater than a threshold α.

We obtained 1M from 3M pairs of (S, T) with $\alpha = 0.25$. We use 900K pairs as training set and 100K pairs as development set. To test our model, we asked annotators with good language skill and good knowledge in news summarization to manually select from a large candidate pool the pairs in which title is indeed a summary of sapo sentence. The test-set, namely PegaTest, contains 9K of such sentence pairs.

4.2 Hyper-parameters

In this subsection, we present hyper-parameters which were tuned on the development set. For all experiments, our models have 256-dimensional hidden states and 128-dimensional word embeddings. We used Adagrad optimizer with mini-batch size 32. Gradient clipping with a maximum gradient norm of 2 was used, but we didn't use any form of regularization. We chose an learning rate of η_0 = 0.15 and an initial accumulator value of 0.1 when training *vanilla seq2seq* and *Pointer* models. Learning rate was reduced to 0.1 on *Pointer+coverage* model.

When it comes to the *Pointer+coverage* model, we experimented in two versions. In the first version, we followed [See et al.2017] to first learn a pointer-generator network and then add the coverage mechanism into loss function and continue to train for 40000 steps. In the second version, we simply trained with coverage from the first iteration with the weight of coverage loss λ = 0.5.

In all our models, we used a vocabulary of 50k words for both source and summary sentences. During training and test time, we truncated source sentence to 50 tokens. In decoding, the length of the summary were limited to 25 tokens. On the other hand, our summaries were produced using beam search with beam size 4. We also used early stopping based on development set in which convergence is reached after around 5 epochs.

4.3 Evaluation on Rouge

Rouge metric is used for evaluation using PegaTest. In this experiment, we compare the following models:

[2] https://github.com/atulkum/pointer_summarizer

[3] https://github.com/phongnt570/UETsegmenter

Table 1: Evaluation on ROUGE

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
Vanilla seq2seq (no filter)	52.25	64.23	56.71	32.96	40.16	35.60	43.42	53.16	47.03
Vanilla seq2seq	69.29	63,63	65.59	49.34	45.51	46.77	58.97	54.33	55.91
Pointer	**69.80**	65.69	66.91	**50.28**	47.38	48.23	**59.43**	55.98	57.00
Pointer+coverage	65.23	69.06	66.36	46.70	49.27	47.38	55.16	58.25	56.00
Pointer+coverage [See et al.2017]	67.09	**69.65**	**67.65**	48.83	**50.50**	**49.11**	57.21	**59.25**	**57.62**

Table 2: Evaluation by humans

Model	Syntax	Factual correctness	Completeness
Vanilla seq2seq	84.67	60.62	68.70
Pointer	91.00	79.48	75.89
Pointer+Coverage	91.30	**80.29**	**80.77**
Pointer+Coverage [See et al.2017]	**92.00**	75.00	73.96

- ***Vanilla seq2seq***: Encoder-decoder with attention mechanism.
- ***Vanilla seq2seq (no filter)***: We want to investigate the effect of filtering noisy data as described in Section 4.1.
- ***Pointer***: Pointer-generator network.
- ***Pointer+coverage***: Pointer-generator network with coverage mechanism.

Firstly, as shown in Table 1, it is clear that filtering noisy data gains a substantial enhancement in quality of summaries. In our experiments, we noticed that, in terms of efficiency, not only training data was reduced, but the models also converged faster than when training on full data.

Secondly, pointer-network and the coverage mechanism bring a significant improvement over vanilla seq2seq model. As already mentioned in their paper, our version of Pointer-network resulted in a slightly worse Rouge score than the original version in [See et al.2017].

As pointed out in [Kryscinski et al.2019] and several works, Rouge metric is insufficient for evaluating summarization. In our experiments, we decided to further assess summary quality by human (Section 4.4).

4.4 Human evaluation

We evaluated summary on three criteria, namely syntax, factual correctness and completeness. We randomly selected 300 sapos and evaluated summaries generated by our models. The evaluation will comply with the following rule: Firstly, a summary that has correct syntax will be further considered for factual correctness. A summary that describes a correct fact as in the original sentence will then be taken into account for completeness. Syntax and factual correctness will have score as 0 (false) or 1 (true). Score of completeness is in the range [0,10], which shows the amount of important information preserved, in comparison with golden summaries.

Table 2 shows the evaluation results. The percentage of correct syntax sentences of vanilla seq2seq model is appreciably lower than the other two models. It's just in as few as 84.67% of samples. It can be observed that the model seq2seq with only attention, unsuccessfully learned Vietnamese grammar.

Pointer is on par with Pointer+coverage in factual correctness. Pointer+coverage [See et al.2017] lags behind with 5% below. Pointer+coverage performs the best on completeness. Two conclusions could be drawn from these results: Coverage mechanism proves to be useful for generating high quality compressions; and most importantly, ROUGE met-

1. (**S**): Ngày 15/8 , Công_ty VinSmart đã có phản_hồi với báo_giới về video so_sánh thiết_kế và linh_kiện smartphone Vsmart_Live với model Meizu 16X tới từ Trung_Quốc .
(**T**): VinSmart nói gì về video so_sánh điện_thoại Vsmart_Live với Meizu
(**Vanilla ses2seq**): [UNK] [UNK] - Tin_tức mới nhất về thiết_kế và linh_kiện smartphone [UNK]
(**Pointer**): Công_ty VinSmart phản_hồi về video so_sánh thiết_kế và linh_kiện smartphone
(**Pointer+coverage**): VinSmart phản_hồi về video so_sánh thiết_kế và linh_kiện smartphone Vsmart_Live

2. (**S**):Bệnh_viện Dã_chiến cấp 2 số 2 được điều_chuyển nguyên_trạng từ Học_viện Quân_y về Cục Gìn_giữ hoà_bình Việt_Nam , bao_gồm chức_năng , nhiệm_vụ , tổ_chức , biên_chế quân_số , trang_thiết_bị .
(**Vanilla ses2seq**):Bệnh_viện Dã_chiến cấp 2 cấp 2 số 2 được điều_chuyển
(**Pointer**): Bệnh_viện Dã_chiến cấp 2 được điều_chuyển nguyên_trạng từ Học_viện Quân_y
(**Pointer+coverage**): Bệnh_viện Dã_chiến cấp 2 được điều_chuyển nguyên_trạng từ Học_viện Quân_y

3. (**S**): Do Trung_Quốc gây khó_dễ và kiểm_soát chặt cửa_khẩu , hàng_loạt trái_cây Việt_Nam như : dưa_hấu , dừa_xiêm và thanh_long ... đều mất_giá từ 50 % .
(**T**): Hàng_loạt trái_cây Việt mất_giá vì Trung_Quốc siết đầu_vào .
(**Vanilla ses2seq**): Hàng_loạt trái_cây Việt_Nam ' bốc_hơi ' từ 50 %
(**Pointer**): hàng_loạt trái_cây Việt_Nam mất_giá từ 50 % từ 50 %
(**Pointer+coverage**): hàng_loạt trái_cây Việt_Nam đều mất_giá từ 50 %

4. (**S**): Một nguồn_tin từ Chính_phủ Nhật_Bản ngày 16/8 cho_biết , nước này vừa đề_xuất với Mỹ rằng Tokyo sẵn_sàng cung_cấp người_máy tự_động sử_dụng cho việc dỡ bỏ các cơ_sở hạt_nhân của Triều_Tiên .
(**T**): Nhật_Bản đề_xuất dùng robot dỡ bỏ các cơ_sở hạt_nhân của Triều_Tiên .
(**Vanilla ses2seq**): Nhật_Bản đề_xuất cho Mỹ dỡ bỏ cơ_sở hạt_nhân Triều_Tiên
(**Pointer**): Nhật_Bản sẵn_sàng cung_cấp người_máy tự_động sử_dụng cho Triều_Tiên .
(**Pointer+coverage**): Nhật_Bản đề_xuất dỡ bỏ các cơ_sở hạt_nhân của Triều_Tiên.

5. (**S**): Virgil van Dijk đã xuất_sắc vượt_qua Leo_Messi và Cristiano_Ronaldo để giành giải_thưởng " Cầu_thủ xuất_sắc nhất mùa_giải 2018/2019 " của UEFA với số điểm 305 .
(**T**): Virgil van Dijk giành danh_hiệu " Cầu_thủ xuất_sắc nhất mùa_giải 2018/2019 " của UEFA .
(**Vanilla ses2seq**): [UNK] giành giải Cầu_thủ xuất_sắc nhất mùa_giải 2018/2019
(**Pointer**): Cristiano_Ronaldo giành giải_thưởng " Cầu_thủ xuất_sắc nhất mùa_giải 2018/2019 " .
(**Pointer+coverage**): Cristiano_Ronaldo giành giải Cầu_thủ xuất_sắc nhất mùa_giải 2018/2019 .

Figure 3: Example of compresses sentences.

ric doesn't always correlate with human evaluation.

Furthermore, we provide a detailed analysis of grammatical and semantic issues of seq2seq models through the examples in figure 3.

Vanilla seq2seq usually generates [UNK] tokens or uses novel words that change the meaning of the original sentence. Proper names and numbers are usually replaced by [UNK]. In Example 3, the verb 'mất_giá' (devalue) was replaced by the verb 'bốc_hơi' (evaporate) making the summary sentence semantically incorrect. Even more catastrophically, the condensed sentences sometimes evolve into repetitive nonsense, such as the second sample.

The other two models show a significant improvement when considering these issues. The verb 'mất_giá' (devalue) was captured exactly in Pointer but the repetition problem has not yet been resolved. When the coverage mechanism is added, accurate condensed information is generated. Moreover, both proper names 'VinSmart' and 'Vsmart_live' in Example 1 have been correctly presented by the last model.

In Example 4, all three models generate sentences with correct grammar but none of them is capable of retaining the main idea of the original sentence. In Example 5, vanilla seq2seq unsuccessfully deals with OOV words ('Virgil van Dijk'). Meanwhile, both Pointer and Pointer+coverage select 'Cristiano_Ronaldo' as misleading subject of the sentence.

The models also tend to learn writing style of news title such as using a location with ':', or subject ellipsis. This suggests that further filtering could be applied to enhance data quality.

5 Conclusions and Future Work

In this work, we have presented Pointer-Generator Networks for Vietnamese sentence compression. Results are significant given no requirement on manual annotation. However, the generated summaries are far from perfect.

As a next step, we would like to further improve the summary quality, in a data-driven way, as well as scale this system to generate paragraph-level summaries. Both pose additional challenges in terms of efficient alignment and consistency in generation. Another direction is delving into seq2seq

framework, including recent transformer-based alternative, to increase model interpretability.

References

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. 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 673–681, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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, San Diego, California, June. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *J. Artif. Intell. Res. (JAIR)*, 31:399–429, 01.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 137–144, Manchester, UK, August. Coling 2008 Organizing Committee.

Viet Lai Dac, Truong Son Nguyen, and Le Minh Nguyen. 2017. Deletion-based sentence compression using bi-enc-dec LSTM. In *Computational Linguistics - 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16-18, 2017, Revised Selected Papers*, pages 249–260.

Thibault Fevry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium, October. Association for Computational Linguistics.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal, September. Association for Computational Linguistics.

Dimitrios Galanis and Ion Androutsopoulos. 2010. An extractive supervised two-stage method for sentence compression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 885–893, Los Angeles, California, June. Association for Computational Linguistics.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *In Proceedings of the 6th Applied Natural Language Processing Conference*, pages 310–315.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China, November. Association for Computational Linguistics.

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.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Guilcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Le Minh Nguyen and Susumu Horiguchi. 2003. A sentence reduction using syntax control. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*, pages 146–152, Sapporo, Japan, July. Association for Computational Linguistics.

Thi Thu Ha Nguyen and Huu Quynh Nguyen. 2011. Concatenate the most likelihood substring for generating vietnamese sentence reduction. *International journal of engineering and technology*, 3:203–207.

Le Minh Nguyen, Susumu Horiguchi, Akira Shimazu, and Tu Bao Ho. 2004. Example-based sentence reduction using the hidden markov model. In *ACM Transactions on Asian Language Information Processing (TALIP*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-

generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Trung Tran and Dang Tuan Nguyen. 2018. Text generation from abstract semantic representation for summarizing vietnamese paragraphs having co-references. *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 93–98.

Nhi Thao Tran, Van Giau Ung, An Vinh Luong, Minh Quoc Nghiem, and Luu Thuy Ngan Nguyen. 2015. Improving vietnamese sentence compression by segmenting meaning chunks. *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pages 320–323.

Liangguo Wang, Jing Jiang, Hai Leong Chieu, Chen Hui Ong, Dandan Song, and Lejian Liao. 2017. Can syntax help? improving an LSTM-based sentence compression model for new domains. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1385–1393, Vancouver, Canada, July. Association for Computational Linguistics.

Sander Wubben, Emiel Krahmer, Antal van den Bosch, and Suzan Verberne. 2016. Abstractive compression of captions with attentive recurrent neural networks. In *Proceedings of the 9th International Natural Language Generation conference*, pages 41–50, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Indirectly Determined Comparison and Difference:

The Case of Japanese

Toshiko Oda
Tokyo Keizai University
1-7-34 Minami-cho, Kokubunji
185-8502 Tokyo, Japan
toda@tku.ac.jp

Abstract

When making comparisons, Japanese allows somewhat sloppy comparisons. In Japanese, 'Taro's grade is better than Hanako' means Taro's grade is better than Hanako's grade. Also, when referring to a difference, 'Taro's opinion is different from Hanako' in Japanese means Taro's opinion is different from Hanako's opinion. Such collocations are widely observed in the language. This paper argues that comparison constructions and 'different' constructions are calculated in context-dependent manners in Japanese. In doing so, I will apply Hohaus's (2015) framework of the 'indirect strategy' of degree comparison to phrasal comparatives and 'different' constructions in Japanese.

1 Introduction

This paper is concerned with the phrasal *yorimo*-comparatives and 'different' constructions in Japanese such as the following. Intuitively, (1) compares Taro's grade and Hanako's grade, and (2) means Taro's opinion is different from Hanako's opinion. However, what appears on the surface in both cases is only *Hanako*.[1] 'Hanako's grade' or 'Hanako's opinion' never directly appears.

Yorimo-comparative in Japanese[2]

(1) Taro-no seiseki-wa Hanako-yorimo
Taro-Gen grade-Top Hanako-YORIMO
ii.[3]
good.Nonpast
Lit. 'Taro's grade is better than Hanako.'
'Taro's grade is better than Hanako's grade.'

'Different' in Japanese[4]

(2) Taro-no iken-wa Hanako-to
Taro-Gen opinion-Top Hanako-with
tigau.
different.Nonpast
Lit. 'Taro's opinion is different from Hanako.'
'Taro's opinion is different from Hanako's opinion.'

These examples are somewhat unexpected, as their English equivalents never mean what the Japanese sentences mean.

[1] I will use the following abbreviations in this paper. Gen = genitive case marker; Top = topic marker; Acc = accusative case marker; Nonpast = nonpast tense; RC = relative clause; Lit = literal translation

[2] *Yorimo* adopted in this paper is interchangeable with *yori*. Both are normally translated as 'than'. Nevertheless, there are some exceptions where only *yori* is allowed. See Sawada (2013) for details.

[3] *Yorimo* is normally glossed 'than'. However, I will simply gloss it as 'YORIMO' throughout the paper. Later in Section 3 I will argue that it is an equivalent of *compared to* in English.

[4] *Tigau* 'different' in Japanese is a verb.

(3) #John's grade is better than Mary.

(4) #John's opinion is different from Mary.

One may assume syntactic deletions of '-'s grade' or '-'s opinion' in the Japanese examples.

(5) Taro-no seiseki-wa Hanako-~~no seiseki~~
Taro-Gen grade-Top Hanako-~~Gen grade~~
yorimo ii.
YORIMO good.Nonpast
'Taro's grade is better than Hanako~~'s grade~~.'

(6) Taro-no iken-wa Hanako-~~no ikien~~
Taro-Gen opinion-Top Hanako-~~Gen opinion~~
to tigau.
with different.Nonpast
'Taro's opinion is different from Hanako~~'s opinion~~.'

However, these are unlikely options. In (7), *no ronbun* '-'s paper' is deleted in the second sentence, and it is ungrammatical for the intended interpretation, in sharp contrast to the commonly observed NP deletions exemplified in (8).

Deletion of '-'s paper'
(7) Taro-wa Jiro-no ronbun-o yonda.
Taro-Top Jiro-Gen paper-Acc read
*Hanako-wa Maki-~~no ronbun~~-o yonda.
Hanako-Top Maki-~~Gen paper~~-Acc read
'Taro read Jiro's paper. Hanako read Maki~~'s paper.~~'

NP deletion
(8) Taro-wa Jiro-no ronbun-o yonda.
Taro-Top Jiro-Gen paper-Acc read
Hanako-wa Maki-no ~~ronbun~~-o yonda.
Hanako-Top Maki-Gen ~~paper~~-Acc read
'Taro read Jiro's paper. Hanako read Maki's ~~paper~~.'

Therefore, (1) and (2) should not be analyzed as the deletion of '-'s grade' or '-'s opinion'.

How then do the Japanese examples mean what they mean? In this paper, I will argue that the somewhat sloppy *yorimo*-comparative and the 'different' construction in Japanese given in (1) and (2) are made possible in a context-dependent manner. I will adopt Hohaus's (2015) framework of the indirect strategy of degree comparison. Unlike standard compositional comparison, the value of the standard is determined less compositionally and in a more context-dependent manner in the indirect strategy. It roughly means as follows. In the case of (1), for instance, *Hanako yorimo* 'Hanako YORIMO' adds information to the presupposition of the sentence instead of directly being part of the assertion. The interpretation of 'Hanako's grade' is indirectly provided from the utterance context, which is enriched by *Hanako yorimo* 'Hanako YORIMO'.

I will further argue that practically the same mechanism takes place for the 'different' construction given in (2). In other words, (2) employs a non-degree version of the indirect strategy. Thus, *Hanako to* 'with Hanako' in (2) simply adds information to the presupposition of the sentence, and the interpretation of 'Hanako's opinion' is indirectly provided from the utterance context enriched by *Hanako to* 'with Hanako'.

The organization of this paper is as follows. Section 2 reviews basic analyses of comparatives and 'different' constructions as well as Hohaus's (2015) framework of the indirect strategy of degree comparison. Section 3 provides an analysis of *yorimo*-comparatives and 'different' constructions in Japanese under the indirect strategy. Section 4 has concluding remarks and a list of related topics for further research.

2 Previous studies

2.1 Phrasal comparatives and 'different' constructions

Comparatives have always been at the center of research on degree constructions. Furthermore, Heim (1985) notes that there are more basic types of comparison than degree comparisons, where "the issue is simply they (items in comparisons) are the same or different." (Heim 1985, p. 21) In other words, 'different' constructions are also regarded as a type of comparative construction, and this notion has been widely shared (Carlson 1987, Moltmann 1992, Beck 2000, among others).

In what follows, I will briefly review the semantics of phrasal comparatives as well as the semantics of 'different' constructions in English.

Given in (9) is a prototypical example of phrasal comparative in English. For the purposes of discussion, let us call the mechanism employed in (9) 'standard comparison'. One of the most

commonly assumed comparative operators for phrasal comparatives is presented in (10). The LF structure of (9) is given in (11), where the subject and the comparative operator undergo movement. This creates a degree predicate that is shared by *John* and *Mary*. The semantics of the sentence are as shown in (12). Note that *than* is considered semantically null.

Standard comparison
(9) John is taller than Mary.

(10) ⟦*-er*⟧ (x, y)($D_{<d, <e,t>>}$) =1 iff
∃d[D(d)(x) ∧ d>MAX(λd'. D(d')(y))]

(11) [John [$_{DegP}$ -er than Mary][2[1[t_1 is t_2-tall]]]

(12) ⟦(9)⟧ =1 iff MAX(λd.tall(d)(John))>
MAX(λd.tall(d)(Mary))

As for the semantics of 'different' constructions, let us first examine a very simple example in (13). The semantics of *different* is given in (14), ignoring tense, intensionality, etc. Simply put, *different* is a two-place predicate as shown in (15). The semantics of the sentence is given in (16), which means that 'our last car' and 'this car' are not the same. Note that *from* is considered semantically null.

'Different'
(13) Our last car was different from this one.
(Beck 2000)

(14) ⟦*different*⟧ (a, b)=1 iff (i) or (ii):
(i) a≠b
(ii) a and b belong to kinds a' and b', and a'≠b' (Beck 2000)

(15) ⟦*different*⟧ =λyλx[Different(x, y)]

(16) Different(our_last_car', this_one')
(Beck 2000)

Parallelism between the comparative sentence and the 'different' construction is obvious. The comparative operator *-er* defines the relationship of two degrees, whereas *different* determines the relationship of two non-degree individuals. The standard markers *than/from* are semantically null, and they simply introduce a standard of comparison. Most importantly, individual *x* and *y*, i.e., *John* and *Mary* in our cases, are directly involved in the compositional calculations.

2.2 The indirect strategy of degree comparison

Hohaus (2015) proposes a mechanism of degree comparison that is less compositional and more context-dependent than the mechanism of standard comparison. She calls this mechanism 'the indirect strategy'. In English, an example of comparison made by the indirect strategy is given in (17).

Comparison by the indirect strategy
(17) Compared to Mary, John is taller.[5]

This sentence induces intuitively the same meaning as (9). However, its semantics is quite different from (9) and consists of two parts. The main clause *John is taller* provides an assertion, and the adjunct phrase *Compared to Mary* contributes to the presuppositions.

The LF of (17) is given in (18). In the main clause *John is taller*, the standard degree comes from a free variable of degree d_7, whose value is determined by an assignment function *g*.

(18) [[$_{FrameP}$ FRAME [Compared to Mary]][3 [[$_{DegP}$ –er d_7][1[John is t_1-tall in s_3]]]]]

Such comparison with a free variable of degree is quite common. In English, for instance, *John is taller* means that John's height is compared with a degree that is salient in the utterance context. The comparative morpheme given in (19) makes a comparison with a free variable of degree. Note that d_n stands for a free variable of degree with an index *n*.

(19) ⟦*-er*⟧ ^{g}(d, d_n)($D_{<d,t>}$)=1 iff ∃d[D(d)(x) ∧ d>d_n]

[5] It should be noted that the adjective of the sentences is in the comparative form *taller*. When a positive form is used as in (i), it has a 'vague' semantics that is very different from the semantics of (17). (17) is a case of 'explicit comparison', whereas (i) is a case of 'implicit comparison'. See Sawada (2009) for the analysis of (i).
(i) Compared to Mary, John is tall.

The unique proposal of the indirect strategy by Hohaus concerns the semantics of FrameP in (18). First, *Compared to Mary* introduces a comparison with Mary regarding some kind of measurement μ. This is shown in (20). Then *Compared to Mary* is an argument of FRAME, a phonologically null operator defined as in (21). Hohaus assumes that FRAME introduces a presupposition that a proposition *p* holds in a minimal situation. As a result, FrameP brings in a presupposition that the relevant comparison introduced by *compared to Mary* takes place in a minimal situation that is small enough for the relevant comparison with Mary but nothing else. In other words, the presupposition introduced by FrameP defines a very narrow situation where the assertion *q* holds.

(20) ⟦*Compared to Mary*⟧ = $\lambda s_{<s>}.\exists x_{<e>,} \exists \mu_{<s,<e,d>>}$ [μ(s)(x)≥ μ(s)(Mary)]

(21) ⟦FRAME⟧ =$\lambda p_{<s,t>}.\lambda q_{<s,t>}.\lambda s$:MIN(p)(s). q(s)

(22) ⟦FrameP⟧ =$\lambda q_{<s,t>}.\lambda s$:s∈MIN($\lambda s^*$. $\exists x_{<e>,}\exists \mu_{<s,<e,d>>}$[$\mu$(s*)(x)≥ μ(s*)(Mary)]). q(s)

The truth conditions of (17) are given in (23). They are defined if a relevant comparison involves Mary in a minimal situation *s*. When defined, the sentence is true iff John's height is greater than a contextually provided degree in *s*. This assertion needs to satisfy the presupposition. Thus, the value of the free variable of degree *g(7)* is naturally understood as the height of Mary.

(23) λs: s∈MIN($\lambda s^*.\exists x_{<e>,}\exists \mu_{<s,<e,d>>}$[$\mu$(s*)(x)≥ μ(s*)(Mary)]). MAX(λd. John is d-tall in s)>g(7)

The secret of flexible semantics of (23) lies in the assignment function *g*. *g(7)* denotes a degree of height, and it takes Mary as its argument, as is required by the presupposition. The formation of *g* is not subject to syntactic constraint.

(24) g(7)= [λx. max(λd. tall(d)(x)) in s](Mary)
= max(λd. tall(d)(Mary)) in s

In summary, the semantics of (17) is achieved by a combination of complicated presuppositions and an assertion that involves a free variable of degree. Importantly, the semantics of *John is taller* and *Compared to Mary* are not directly combined via compositional calculation. They are rather indirectly combined via the utterance context.

What is the motivation for Hohaus (2015) to propose the indirect strategy? One of the motivations comes from the fact that some comparative constructions such as *compared to*-constructions induce meanings that are not syntactically possible under standard comparison.

Let us see an example. Intuitively, (25) is a comparison between a paper written by John and another paper written by Bill, and the former is longer than the latter. This interpretation is not available in the phrasal *more-than* comparative given in (26). In order to derive the intended reading, *Bill* would need be an argument of a predicate of 'wrote_a_paper_that_was_d-long'. This requires the comparative morpheme *-er* to move out of the relative clause, which is syntactically ruled out. When *-er* moves within the relative clause island without violating island constraints, it only derives an odd comparison between the length of a paper and Bill himself. In fact, that is the only reading available for (26).

(25) Compared to Bill, John wrote a paper [$_{RC}$ that was longer d_5].

(26) #John wrote a paper [$_{RC}$ that was longer than Bill].
(Beck et al. 2012)

On the other hand, comparison by the indirect strategy in (25) does not involve such syntactic constraints. The comparative morpheme *-er* moves within the relative clause, and it takes a free degree variable, say d_5, as its argument. The free degree variable d_5 is understood as the length of the paper that Bill wrote thanks to the presupposition enriched by *compared to Bill*. In other words, the assignment function has the meaning described in (27), and it takes Bill as required by the presupposition. The meaning of *g* is contextually determined, and it is free from syntactic constraints.

(27) g(5)= [λx. max($\exists$y. paper(y) $\wedge$ wrote(y)(x) $\wedge$ long(d)(y)](Bill)
= max(λd. $\exists$y. paper(y) $\wedge$ wrote(y)(Bill) $\wedge$ long(d)(y))

In summary, the comparison by the indirect strategy given in (17) achieves an interpretation that is intuitively very similar in meaning to that of the standard comparison in (9). However, its mechanism is quite different from standard comparison. Comparisons using the indirect strategy have more context-dependent semantics, and they sometime achieve interpretations that are not possible under standard comparison.

3 Japanese data

In this section I will apply Hohaus's (2015) indirect strategy of degree comparison to the phrasal *yorimo* comparative that we saw in (1). Its unexpected reading is accounted for under the indirect strategy. I will further argue that practically the same analysis applies to the 'different' construction in (2).

3.1 Context-dependent comparison

Let us first consider the case of the phrasal *yorimo*-comparative sentence in (1), repeated as (28) below. Note that following Bhatt and Takahashi (2011), Kubota (2011), Matsui and Kubota (2012), and others, I assume that phrasal *yorimo*-comparatives are underlyingly phrasal. In other words, they are not derived from underlying clausal comparatives.

Yorimo-comparative in Japanese

(28) Taro-no seiseki-wa Hanako-yorimo
Taro-Gen grade-Top Hanako-YORIMO
ii.
good.Nonpast
Lit. 'Taro's grade is better than Hanako.'
'Taro's grade is better than Hanako's grade.'

In (28), what precedes *yorimo* is just *Hanako*. Nevertheless, the sentence produces an interpretation of a comparison between Taro's grade and Hanako's grade.

The following contrast observed in English supports our assumption of (28) as a comparison by the indirect strategy. (29) is an instance of standard comparison, and it is ungrammatical for the intended comparison. However, a comparison with the indirect strategy given in (30) intuitively means a comparison between Taro's grade and Hanako's grade. Therefore, the Japanese sentence (28) should be an equivalent of (30) rather than that of (29).

(29) *Taro's grade is better than Hanako.
(Intended: Taro's grade is better than Hanako's grade.)

(30) Compared to Hanako, Taro's grade is better.

Let us see how Hohaus's framework captures the Japanese data. Given in (31) is the LF of (28). *Hanako yorimo* is part of FrameP. *Hanako yorimo* is an equivalent of *compared to Hanako*, and its semantics is given in (32). It means that Hanako is involved in a degree comparison relation with another individual *x*. The FRAME operator is repeated in (33) from (21). As a result, FrameP introduces a presupposition that a relevant degree comparison in the context is made with Hanako in a minimal situation.

(31) $[[_{\text{FrameP}}$ FRAME [Hanako-yorimo]][3 $[[_{\text{DegP}}$ $\varnothing_{\text{-er}}$ d_9][1[Taro-no seiseki-wa t_1-ii s_3]]]]]

(32) $[\![\text{Hanako-yorimo}]\!] = \lambda s_{<s>}.\exists x_{<e>}.\exists \mu_{<s,<e,d>>}$ $[\mu(s)(x) \geq \mu(s)(\text{Hanako})]$

(33) $[\![\text{FRAME}]\!] = \lambda p_{<s,t>}.\lambda q_{<s,t>}.\lambda s{:}\text{MIN}(p)(s).\ q(s)$

(34) $[\![\text{FrameP}]\!] = \lambda q_{<s,t>}.\lambda s{:}s \in \text{MIN}(\lambda s^*.\ \exists x_{<e>},$ $\exists \mu_{<s,<e,d>>}[\mu(s^*)(x) \geq \mu(s^*)(\text{Hanako})]).\ q(s)$

As for the main clause, I assume that Japanese has a phonologically null comparative operator $\varnothing_{\text{-er}}$ as defined in (35), whose semantics is the same as that of the comparative operator for phrasal comparatives in English that we adopted in (19).

(35) $[\![\varnothing_{\text{-er}}]\!]^g(d, d_n)(D_{<d,t>})=1$ iff $\exists d[D(x)(d) \wedge d>d_n]$

The truth conditions of (28) are given in (36). They are defined if a relevant degree comparison involves Hanako in a minimal situation. When defined, the sentence is true iff the degree of Taro's grade is better than a contextually provided degree in a minimal situation.

(36) $\lambda s{:}\ s \in \text{MIN}(\lambda s^*.\exists x_{<e>},\exists \mu_{<s,<e,d>>}[\mu(s^*)(x) \geq$

μ(s*)(Hanako)]). MAX(λd. Taro's grade is d-good in s)>g(9)

The secret of (36) in deriving a comparison with Hanako's grade lies in the flexibility of the assignment function *g*. It takes Hanako as required by the presupposition and derives the degree of the grade that Hanako possesses.

(37) g(9)= [λx. max(λd. $\exists y_{<e>}$.grade(y) ∧ good(d)(y) ∧ possess(y)(x))](Hanako)
= max(λd. $\exists y_{<e>}$.grade(y) ∧ good(d)(y) ∧ posses(y)(Hanako))

In this subsection, we have seen that somewhat puzzling data of Japanese comparatives can be captured by the indirect strategy of degree comparison.[6] As the indirect strategy is more context-dependent than the standard strategy of comparison, it sometimes derives interpretations that are not possible for standard comparison.

3.2 Context-dependent 'different'

In this subsection, I will apply the indirect strategy of degree comparatives to 'different' constructions in Japanese. Applying Hohaus's (2015) indirect strategy to 'different' construction is a novel approach. Nevertheless, it should be a natural consequence if 'different' constructions are a type of comparative constructions, as previous studies have argued.

Let us first recall the relevant example in (2), repeated as (38) below.

(38) Taro-no iken-wa Hanako-to
Taro-Gen opinion-Top Hanako-with
tigau.
different.Nonpast
Lit. 'Taro's opinion is different from Hanako.'
'Taro's opinion is different from Hanako's opinion.'

When (38) is analyzed in a standard manner, it produces an incorrect result. Assuming that the semantics of *tigau* 'different' is the same as that of *different* as given in (39) and (40), the resulting semantics will be (41). This is incorrect, as it directly compares Taro's opinion and Hanako herself.

(39) ⟦*tigau*⟧ (a, b)=1 iff (i) or (ii):
(i) a≠b
(ii) a and b belong to kinds a' and b', and a'≠b'

(40) ⟦*tigau*⟧ =λyλx[Different(x, y)]

(41) Different(Taro's_opinion, Hanako)

Put differently, the semantics given in (41) is for a sentence such as (42) in English. This is not what (38) means. Instead, the interpretation of (38) can be paraphrased in English with *compared to* as shown in (43). Note that (43) may not be the most natural sentence in English, but it somehow carries an interpretation that Taro's opinion and Hanako's opinion are different.

(42) #Taro's opinion is different from Hanako.

(43) Compared to Hanako, Taro's opinion is different.

What we need is a non-degree version of the indirect strategy. I assume the LF structure in (44) for (38). *Hanako to* 'with Hanako' is part of FrameP. This is a crucial assumption in capturing (38). Another crucial assumption is that the assertion part has the free variable of individual e_4.

(44) [[$_{\text{FrameP}}$ FRAME [Hanako-to]][3 [Taro-no iken-wa e_4 tigau s_3]]]

It is quite normal for 'different' constructions to have such free variables. (45) is minimally different from (38) in that it does not have *Hanako to* 'with Hanako'. A compared item is given in the context, and *tigau* 'different' takes a free variable e_4, indicated in (45) for convenience. The same phenomenon is observed in an equivalent example in English in (46). In both cases, there is already a salient opinion in the utterance context that is different from Taro's. The exact value of e_4 is determined by an assignment function *g* for the utterance context.

[6] In this paper I only discuss phrasal *yorimo*-comparatives by the indirect strategy. However, I discussed in Oda (2020a) that phrasal *yorimo*-comparatives are ambiguous. Some *yorimo*-comparatives are made by the indirect strategy and others are standard comparison. Whether or not similar ambiguity exists in 'different' construction in Japanese is left for further research.

(45) Taro-no iken-wa tigau e_4.
Taro-Gen opinion-Top different.
'Taro's opinion is different.'

(46) Taro's opinion is different e_4.

Now let us see how the semantics of (38) is calculated. The semantics of FrameP is composed as follows. *Hanako to* 'with Hanako' means that Hanako is in a relation of *r* along with another individual *x*. FRAME operator is the same as adopted from Hohaus in (21). As a result, FrameP brings a presupposition that Hanako and another individual *x* are in a certain relationship *r* in a minimal situation. *r* could be any two-place relation such as same-relation, different-relation, or something else. In any case, the assertion *q* needs to satisfy the presupposition.

(47) ⟦*Hanako to*⟧ = $\lambda s_{<s>}.\exists x_{<e>},\exists r_{<s,<e,<e,t>>>}$ [*r*(s)(x)(Hanako)]

(48) ⟦FRAME⟧ =$\lambda p_{<s,t>}.\lambda q_{<s,t>}.\lambda s$:MIN(p)(s). q(s)

(49) ⟦FrameP⟧ =$\lambda q_{<s,t>}.\lambda s$:s∈MIN($\lambda s^*.\ \exists x_{<e>},\ \exists\ r_{<s,<e,<e,t>>>}$ [*r*(s)(x)(Hanako)]). q(s)

The assertion part is composed in the normal manner. One of the arguments of *tigau* 'different' is a free variable e_4.

(50) Different(Taro's_opinion, g(4))

The truth conditions of (38) are given in (51). They are defined if an individual *x* and Hanako are in a certain relation in a minimal situation. When defined, the sentence is true iff Taro's opinion and a contextually given item are not the same. The item in comparison is understood as Hanako's opinion due to the presupposition enriched by *Hanako to* 'with Hanako'.

(51) λs: s∈MIN($\lambda s^*.\exists x_{<e>},\exists\ r_{<s,<e,<e,t>>>}$ [*r*(s)(x)(Hanako)]). Different(Taro's_opinion, g(4)) in s

The secret of (51) in deriving an intuitive comparison with Hanako's opinion lies in the assignment function *g*, which takes *Hanako* as its argument because of the requirement by the presupposition. This can be described as in (52) below. The maximality operator brings the effect of a definite determiner. Intuitively speaking, the value of *g(4)* is 'the opinion of Hanako.'

(52) g(4)= [λx. MAX($\lambda y_{<e>}$.opinion(y) ∧ possess(y)(x))](Hanako)
= MAX($\lambda y_{<e>}$.opinion(y) ∧ possess(y)(Hanako))

The summary of this section is as follows. Somewhat unexpected interpretations of phrasal *yorimo*-comparatives and 'different' constructions in Japanese are captured by Hohaus's (2015) framework of the indirect strategy of degree comparison. 'Different' constructions are treated as a non-degree version of comparison. To my knowledge, this is the first attempt to apply Hohaus's indirect strategy to non-degree constructions.

4 Conclusion

In this paper I pointed out that phrasal *yorimo*-comparatives and 'different'-constructions present data that are normally not possible for the corresponding phrasal comparatives or 'different' constructions in English. Standard analyses of comparatives and 'different' constructions fail to capture these data. I argue that the Japanese examples are made possible by a mechanism that is different from what is normally assumed. I adopted Hohaus's (2015) framework of indirect strategy, where comparisons are made in a more context-dependent manner via value assignment of free variables.

This paper makes the following theoretical contributions. First, it provides cross-linguistic support for the parallelism between degree comparatives and 'different' constructions. Second, it provides cross-linguistic as well as cross-structural support for Hohaus's (2015) indirect strategy. In particular, this paper marks the first attempt to apply such a context-dependent mechanism to 'different' constructions in any language.

This paper provides many interesting topics for further research. I would like to point out four of them. First, the analysis made in this paper is very likely to apply to 'same' constructions.

‘Same’ constructions in Japanese allow somewhat ‘sloppy’ sentences that are quite similar to (2).
‘Same’ in Japanese

(53) Taro-no iken-wa Hanako-to
Taro-Gen opinion-Top Hanako-with
onaji-da.
same-Copula.Nonpast
Lit. ‘Taro’s opinion is the same as Hanako.’
‘Taro’s opinion is the same as Hanako’s opinion.’

Second, the analysis of phrasal comparatives presented in this paper can be applied to phrasal comparatives in other languages. Oda (2020b) points out that some *bi*-comparatives in Mandarin Chinese are better captured by the indirect strategy than the standard manner of comparison. Phrasal *pota*-comparatives Korean are another candidate, as Park (2016) and An (2020) present data very similar to (1) in Japanese.

Third, the conclusion of this paper raises an interesting issue regarding the semantics of comparatives in Japanese. Clausal *yorimo*-comparatives exhibit certain unique behaviors that are not observed in clausal comparatives in English and other languages. Beck et al. (2004) proposed a very context-dependent semantics to capture them. However, many researchers have argued against such a context-dependent analysis (Shimoyama 2012, Sudo 2015, among others). It should be kept in mind that clausal comparatives and phrasal comparatives can be quite different within a language. Nevertheless, a context-dependent mechanism may be worth considering again for clausal *yorimo*-comparatives if the analysis of this paper of phrasal *yorimo*-comparatives turns out to be on the right track.

Finally, the semantics-based analysis provided in this paper may need to be compared with a syntax-based analysis. I denied a syntactic analysis by providing the ungrammatical example with the deletion of ‘-’s paper’ in (7). If we pursued such deletion in *yorimo*-comparatives, we would need to assume that *yorimo*-comparatives exceptionally allow deletions that are normally banned in Japanese. Such assumption may not be promising, but it could still be defensible. In fact, this is the direction that Park (2018) and An (2019) pursue regarding phrasal *pota*-comparatives in Korean. A comparison between semantics vs. syntactic analysis could be an interesting topic for further research.

Acknowledgments

I thank the two anonymous reviewers of PACLIC34 for their comments and suggestions. All errors are my own. This study was supported by JSPS KAKENHI Grant Number JP20K00582.

References

An, Duk-Ho. 2020. Reduced NP Comparatives in Korean. *Journal of East Asian Linguistics* 29: 337-364.

Beck, Sigrid. 2000. The Semantics of *Different*: Comparison Operator and Relational Adjective. Linguistics and Philosophy 23: 101-139.

Beck, Sigrid, Toshiko Oda, and Koji Sugisaki. 2004. Parametric Variation in the Semantics of Comparison: Japanese vs. English. *Journal of East Asian Linguistics* 13: 289-344.

Beck, Sigrid, Vera Hohaus and Sonja Tiemann. 2012. A Note on Phrasal Comparatives. *Proceedings of Semantics and Linguistic Theory* 22: 146-165.

Bhatt, Rajesh and Shoichi Takahashi. 2011. Reduced and Unreduced Phrasal Comparatives. *Natural Language and Linguistic Theory* 29: 581-620.

Carlson, Greg N. 1987. Same and Different: Some Consequences for Syntax and Semantics. Linguistics and Philosophy 10: 531-566.

Heim, Irene. 1985. Notes on Comparatives and Related Matters. Manuscript, University of Texas at Austin.

Hohaus, Vera. 2015. Context and Composition: How Presuppositions Restrict the Interpretation of Free Variables. Doctoral dissertation, University of Tübingen, Germany.

Kubota, Yusuke. 2011. Phrasal Comparatives in Japanese: A Measure Funtion-Based Analysis. *Empirical Issues in Syntax and Semantics* 8: 267-286.

Matsui, Ai and Yusuke Kubota. 2012. Comparatives and Contrastiveness: Semantics and Pragmatics of Japanese *Hoo* Comparatives. *Formal Approaches Japanese Linguistics* 5: 125-138.

Moltmann, Frederike. 1992. Reciprocals and *Same/ Different*: Towards a Semantic Analysis. Linguistics and Philosophy, 15: 411-462.

Oda, Toshiko. 2020a. Contextual Comparison as a Last Resort by Interpretive Economy. *Paper presented at*

Japanese and Korean Linguistics 28. University of Central Lancashire.

Oda, Toshiko. 2020b. Pragmatic Phrasal Comparison in Mandarin Chinese. *Proceedings of the 19th Meeting of the Texas Linguistics Society*, 43-61.

Park, So-Yong. 2018. Arguments for NP-Ellipsis in Korean. *Korean Journal of Linguistics* 41: 289-311.

Sawada, Osamu. 2009. Pragmatic Aspects of Implicit Comparison: An Economy-Based Approach. *Journal of Pragmatics* 41: 1079-1103.

Sawada, Osamu. 2013. The Comparative Morpheme in Modern Japanese: Looking at the Core from 'Outside'. *Journal or East Asian Linguistics* 22: 217-260.

Shimoyama, Junko. 2012. Reassessing Crosslinguistic Variation in Clausal Comparatives. *Natural Language Semantics* 20: 83-113.

Sudo, Yasutada. 2015. Hidden Nominal Structures in Japanese Clausal Comparatives. *Journal of East Asian Linguistics* 24: 1-51.

Park, So-Yong. 2018. Arguments for NP-Ellipsis in Korean. *Korean Journal of Linguistics* 41: 289-311.

Extraction of Novel Character Information from Synopses of Fantasy Novels in Japanese using Sequence Labeling

Yuji Oka
Kagawa University
Takamatsu, Kagawa, JAPAN
s20g460@stu.kagawa-u.ac.jp

Kazuaki Ando
Kagawa University
Takamatsu, Kagawa, JAPAN
ando.kazuaki@kagawa-u.ac.jp

Abstract

In this paper, we propose a method of extracting novel character information such as characters' names, gender, age, occupations, and a part of relationships between characters from synopses of fantasy novels in Japanese using sequence labeling by comparing sequence labeling models based on CRF (Conditional Random Fields) and deep learning with CRF. From the experimental results, we confirmed that the BiLSTM-CRF model with the information of part-of-speech of words has achieved the best performance, the precision of 85.40%, the recall of 91.47%, and F1-measure of 88.30% for extracting characters' names. The BiLSTM-CRF model has achieved the best overall performance for extracting all tags.

1 Introduction

In recent years, the spread of electronic books has created an environment in which we can read novels easily. Moreover, websites or applications to post and publish new user-generated novels have also received a lot of attention all over the world. On the other hand, as the total number of published and posted novels increases, it will be difficult to find novels that suit a reader's taste. We can search for novels by authors' names, genres of novels, and keywords set independently by the author, and choose novels from the popularity ranking. However, we cannot search for the contents of novels based on personal preferences.

In order to solve such problems, the following methods can be useful: (1) Search function using personal preferences for novels, (2) Synopsis generation and presentation based on personal preferences, (3) Generation and presentation of a correlation diagram of the characters in the novel.

Personal preferences for novels can be divided into preferences for a story such as "surprise ending" and "happy ending", and preferences for a novel character such as "handsome butler" and "dark hero". As the first step to achieve three methods, we focus on the preferences for the novel characters. In this paper, we consider a method to extract information about novel characters. By extracting and organizing the information about the characters from the novel text, it can be used not only for search targets, but also for generating synopses and correlation diagrams.

Although some of text data such as novel posting sites are available free of charge, most text data of novels published for commercial use must be used for a fee. In addition, when extracting information about characters from text of novels by supervised machine learning, it is very costly to construct training data by annotating the whole text of novels. Therefore, we use "synopses" of novels as text for training data in this paper. A synopsis is attached to almost all novels regardless of whether it is published for commercial use, the length of text is short, and the cost of constructing learning data is low. In addition, we confirmed that many synopses of fantasy novels include information about main characters in our preliminary investigation.

In this paper, we propose a method of extracting novel character information such as characters' names, gender, age, occupations, and a part of rela-

tionships between characters from synopses of fantasy novels in Japanese using sequence labeling by comparing sequence labeling models based on CRF (Conditional Random Fields) and deep learning with CRF. Since the proposed method is based on a sentence-by-sentence basis, it can be also applied to the text of novels.

2 Related Work

There are several studies on extracting information about characters from novel text and constructing diagrams or tables of relationships between them. Baba et al. (Baba, 2007) have proposed a method for extracting information about characters from novel text in Japanese by matching with a dictionary and rules. Characters' names are extracted by rules based on part-of-speech information, and attributes of the character are extracted by matching with a dictionary and rules for extraction. A diagram of relationships between characters is constructed based on whether the character exists in a specific scene and the frequency of co-occurrence of characters. From the experimental results, they obtained that the precision of extracting characters' names from novel text was 42.4%, and the recall was about 67.0%.

Yoneda et al. (Yoneda, 2012) have proposed a method for extracting character's name from novel text in Japanese using the local frequency of subjects and information of predicates that co-occur with the subject in a sentence. Based on the hypothesis that a character appears as a subject in a novel, the candidates of the character's name are extracted using particles of Japanese from sentences. Characters' names are identified based on the relationships between each candidate of subjects and predicates that co-occur with the subject in sentences. They showed that the precision of extracting characters' names was 60.3%, the recall was 91.9%, and the F-measure was 71.5%.

Next, we describe three studies to generate diagrams or tables using the extracted character information. Zhang et al. (Zhang, 2017) have proposed methods for inferring salient attributes to generate the description of main characters by extracting attributes from the source story by ranking candidates or classifying using a list of attributes abstractively. They showed that the abstractive model works better than the extractive model, and both model outperform a SVM-based baseline.

Vani et al. (K, 2019) have proposed a method for producing visual summaries by machine learning. Characters and their aliases are detected by standard natural language processing tools for clustering algorithms and named entity recognition. To generate relationship diagrams the most relevant ones and their relations are evaluated based on simple statistical analysis. The color of characters' nodes and undirected edges between characters are determined by a special sentiment analysis method based on sentence embedding.

Iyyer et al. (Iyyer, 2016) have proposed a method for generating a trajectories of temporal changes in the relationship between two characters by unsupervised neural network. The model jointly learns a set of relationship descriptors as well as a trajectory over these descriptors for each relationship in the raw novel text dataset.

Other studies include speaker identification (Iosif, 2014; Ek, 2018) and personality prediction of characters (Flekova, 2015).

3 Proposed Method

In this paper, we compare methods of extracting novel character information using sequence labeling by CRF (Conditional Random Fields) and by deep learning and CRF. We extract characters' names, gender, age, attributes, occupations, position, and organizations in this paper.

3.1 Collection of Novels' Synopses

"Synopses" of novels in Japanese are used as text for training data. To collect synopses of novels, we use Webcat Plus [1], which is an information service provided by the National Institute of Informatics (NII). Webcat Plus provides various information related to paper books and electronic books. We collect synopses of fantasy novels in BOOK database on Webcat Plus.

Finally, we randomly collected 1,008 synopses of fantasy novels written by novelists extracted from "List of Japanese Fantasy Novelists" in Japanese Wikipedia. The synopses collected consists of two or more sentences.

[1] http://webcatplus.nii.ac.jp/

3.2 Construction of Training Data

Training data used for sequence labeling is constructed by the following steps.

1. Tokenize each sentence into words by a Japanese morphological analysis tool.
2. Tag each word based on the following rules. We use IOB2 labeling scheme.

- Tag a sequence related to a character's name with "NAME"
  Example: Nishio, Nobunaga, Charlemagne
- Tag a sequence related to a gender with "MF"
  Example: man, he, beautiful woman, she
- Tag a sequence related to age with "AGE"
  Example: 16 years old, boy, old man, young, high school student
- Tag a sequence related to appearance and characteristics with "STATE"
  Example: white hair, fineness, domineering, genius, craftsmanship
- Tag a sequence related to occupations and position with "PRO"
  Example: hermit, supreme authority, member, king
- Tag a sequence related to organizations and races with "AFF"
  Example: art team, Japanese government, subjugation army, elf
- Tag a sequence related to other information about characters with "OTHER"
  Example: alien, god, demon, penguin
- Tag a sequence related to place names and architectural structures with "PLACE"
  Example: Mu, Japan, Paris, chapel, wizard school
- Tag a sequence related to relationships between characters with "REL"
  Example: brother, parent, enemy, partner, marriage
- Tag the others with O

Although place names and architectural structures are not information about novel characters, we extract them at the same time, because they are named entities useful for considering the stage of the novel (real world, parallel world, different world, etc.). In addition, we try to extract a part of relationships between characters in order to use it for the label of a correlation diagram of the characters. We will extract the rest of the relationships using information such as dependency relationships, conversation sentences in the future work.

3.3 CRF Model

A Conditional Random Fields (CRF) based model is one of the conventional sequence labeling models. We use the CRF model based on word by word labeling to extract character information from synopses of novels. CRFsuite [2] is used to implement the CRF model. The default values are used for hyperparameters of the CRF model. The window size of the CRF model is set to two.

The features used for the CRF model are as follows: (1) Notation of a word, (2) Character types (Seven types), (3) Part-of-speech, (4) Character 1-gram, (5) Character 2-gram, (6) Tag flag : If any characters in the attention word are included in each list of the ten most frequently used Chinese characters within each tag in the training data, tag flags are set based on the types of tags.

The CRF model based on the notation, character types, and part-of-speech is used for the baseline model (general CRF model). The proposed features, namely, character 1-gram, character 2-gram and tag flags, are combined to the baseline in order to verify their effectiveness.

3.4 Deep Learning Model

Recently, a sequence labeling model that combines CRF with deep learning has been proposed. In this paper, in order to extract character information, we use four models, BiLSTM-CRF proposed by Huang et al. (Huang, 2015), BiLSTM-CNN-CRF proposed by Ma et al. (Ma, 2016), BiLSTM-CRF proposed by Lample et al. (Lample, 2016), and Char-BiLSTM-CRF proposed by Misawa et al. (Misawa, 2017). Then we compare the best performance model of them with the CRF model.

Figure 1 shows BiLSTM-CRF (Huang model (Huang, 2015)). The Huang model gives word vectors, which are obtained by inputting word embedding of each word in a sentence into Bidirectional LSTM, as inputs of CRF.

Figure 2, and Figure 3 show BiLSTM-CNN-CRF

[2] http://www.chokkan.org/software/crfsuite/

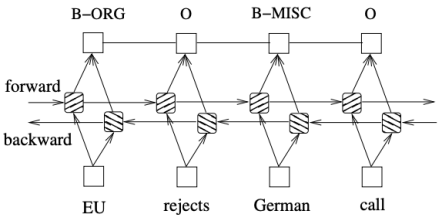


Figure 1: Main Architecture of Huang Model (cited from (Huang, 2015))

(Ma model (Ma, 2016)) and BiLSTM-CRF (Lample model (Lample, 2016)), respectively. Two models improved the performance of the Huang model by using character information included in each attention word. The Ma model inputs word embeddings and character representations computed by CNN into BiLSTM-CRF. The Lample model uses Char-BiLSTM instead of CNN in the Ma model.

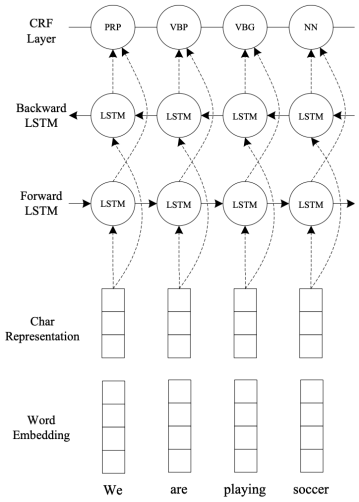


Figure 2: Main Architecture of Ma Model (cited from (Ma, 2016))

Figure 4 shows Char-BiLSTM-CRF (Misawa model (Misawa, 2017)). The Misawa model performs labeling on a character-by-character basis. Labeling on a character-by-character basis has the advantage of not being affected by errors of morphological analysis. The Misawa model inputs each character embedding and word embedding of the attention word containing the characters into BiLSTM-CRF. The Misawa model has achieved the state-of-the-art performance in Japanese NER.

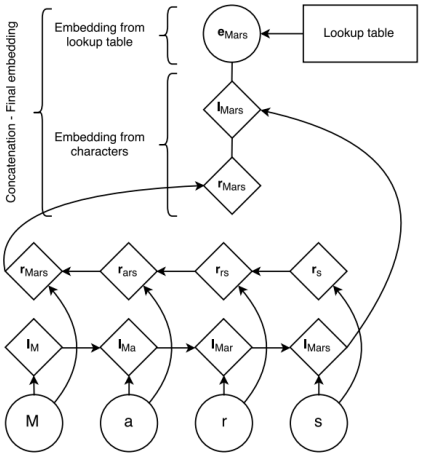


Figure 3: Main Architecture of Lample Model (cited from (Lample, 2016))

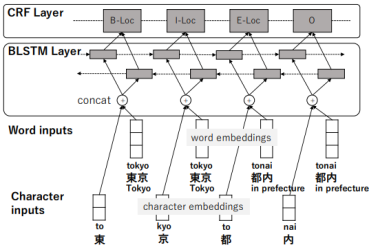


Figure 4: Main Architecture of Misawa Model (cited from (Misawa, 2017))

The parameters used in the deep learning models are shown in Table1. Dropout was adapted before and after input to BiLSTM. The parameters used in the Ma model and Lample model are shown at the bottom of Table1. The dimension of character embeddings used in deep learning models except for the Huang model is set to 50, and the dimension of the hidden layers of CNN of the Ma model and Char-BiLSTM of the Lample model is also set to 50. Each character embedding was initialized within the range of $[-\sqrt{\frac{3}{dim}}, \sqrt{\frac{3}{dim}}]$ with reference to the research (Ma, 2016). Here, dim is the number of dimensions of the character embeddings, and is set to 50 in this experiment. Since the number of characters per word in Japanese is smaller than that in English, the window size of CNN used in the Ma model is set to two. As for the word vectors, we use Wikipedia Entity Vectors [3] which were pre-trained

[3] https://github.com/singletongue/WikiEntVec

in the full text of Japanese Wikipedia. The parameters used in the pre-training are shown in the table 2. The values of word and character embeddings are updated as the model learns.

In addition to word and character embeddings, we also consider the use of part-of-speech vectors as the input of deep learning models. Aguilar et al. state that inputting part-of-speech (pos) information to the model improved the extraction performance for the WNUT2017 dataset constructed from texts in social media (Aguilar, 2017). In this paper, we also consider whether the extraction performance changes by inputting the part-of-speech information into the deep learning models. In order to use the part-of-speech information in the deep learning models, we prepare a part-of-speech vector that is initialized randomly by part-of-speech. Each part-of-speech vector is initialized in the same way as the character embeddings, and the number of dimensions of the vectors is set to zero , five and ten, respectively in the experiment in this paper. The part-of-speech vectors and word embeddings or character embeddings are input to BiLSTM at the same time, and each value of the vectors is updated as the model learns.

Table 1: Hyperparameters of Deep Learning Models

Common parameters	
Demension of BiLSTM hidden layers	128
Number of BiLSTM layers	1
Maximum number of epochs	50
Batch size	32
Learning rate	0.001
Dropout rate	0.5
Gradient clipping	5.0
Optimizer	Adam
Early stopping patience	20
Parameters of Ma model	
Number of CNN filters	50
Window size of CNN	2
Parameters of Lample model	
Dimension of character BiLSTM hidden layer	50
Number of character BiLSTM layers	1

4 Evaluation Experiments

4.1 Evaluation Methods

We evaluate the extraction performance by comparing the results of the combinations of features for the CRF model with the results of four deep learning models. The extraction performance is evaluated by 10-fold cross-validation with precision, recall, and F-measure (F1) as the evaluation criteria. The entities tagged by hand and the results tagged by the machine learning models are compared, and the correctness is judged only when they match perfectly.

Table 2: Hyperparameters of Pre-trained Distributed Representation of Words

Model	CBOW
Number of dimensions	200
Window size	5
Negative sampling	5
Down sampling	0.001

4.2 Dataset

As mentioned in Subsection 3.1, we use 3,679 sentences in 1,008 synopses collected by Webcat Plus for a dataset. In order to construct a dataset, each sentence was segmented by Japanese morphological analyzer MeCab (Kudo, 2004), and character information was tagged in each sentence manually by one person. We are going to tag each sentence by multiple person to ensure accuracy of the dataset in the future work.

The total number of characters in the dataset is 127,486. The average, minimum and maximum number of words in a sentence are 20.60, 2 and 83, respectively. The average, minimum and maximum number of characters in a sentence are 34.65, 3 and 125, respectively. The average, minimum and maximum number of characters in a word are 1.68, 1 and 22, respectively. Figure3 show the information of each tag in the data set.

Table 3: Information about Word and Character in Each Tag

Tag Types	tag	word			character		
	Num	Max	Min	Ave	Max	Min	Ave
NAME	2703	8	1	1.3	18	1	3.5
MF	363	2	1	1.0	3	1	1.5
AGE	436	4	1	1.2	6	1	2.4
STATE	518	8	1	1.8	13	1	3.5
PRO	1358	8	1	1.7	13	1	3.1
AFF	543	8	1	2.0	25	1	4.3
OTHER	825	9	1	1.5	14	1	3.0
REL	1405	8	1	1.6	17	1	3.7
PLACE	722	5	1	1.2	9	1	2.3

4.3 Results

We focus on the results of extracting "NAME" and all tags in this paper, and discuss the extraction performance of each model.

First, we focus on extraction performance of each model for "NAME". Table 4 shows the baseline (general CRF model), the CRF model with combination of the proposed features, and the four deep learning models. The upper part of Table 4 shows the results related to the CRF models. "u", "b", and "f" of the name of the CRF models represent the use of 1-gram, 2-gram, and a tag flag, respectively. "All" of the model's name represents the use of the features obtained from all words in the window size, and "One" represents the use of the features obtained from only the target word. The bottom part of Table 4 shows the results of the deep learning models. BiLSTM-CRF and BiLSTM-CRF-L represent Huang model and Lample model, respectively. "pos5" and "pos10" of the last part of the models' names represent the number of dimensions of the part-of-speech vectors. The underlined values indicate the maximum value of each item in the CRF models. The best performance in all models is highlighted by boldface type.

From the results of uOne and uAll models in the upper part of Table 4, it can be confirmed that when the feature of character 1-gram was added to the baseline, the value of recall improved by 12 points and the value of F1-measure increased by about 7 points. From the results of bOne and bAll models, the model added Character 2-gram to the baseline has a higher precision than the model with character 1-gram, however, it could not improve recall because information could not be obtained from a word consisting of a single character and the comprehensiveness of 2-gram is lower than 1-gram. From the result of fOne and fALL models, it turns out that the extraction performance of the model added tag flags to the baseline was not much improved since the use of tag flags are limited.

As for the deep learning model in the bottom part of Table 4, the extraction performance of all models was improved by the use of part-of-speech vectors. The Char-BiLSTM-CRF-pos5 model had the highest precision, and the BiLSTM-CRF-pos5 model had the highest recall and F1-measure. As for extracting "NAME", the part-of-speech information was effective to improve the extraction performance.

From all the results, we confirmed that the BiLSTM-CRF-pos5 modle has the best extraction performance for "NAME".

Next, we focus on extraction performance of each model for all tags. As for extracting all tags, the uOne-fAll model was the best performance in the CRF models, and the BiLSTM-CRF model without part-of-speech information was the best performance in the deep learning models. Therefore, Table 5 shows performance of the baseline model (general CRF model) and these models.

From Table 5, we can see that the BiLSTM-CRF model achieved the best performance, however, there is no big difference between the extraction performance for "MF". Comparing the extraction performance for all tags of the BiLSTM-CRF model and the BiLSTM-CRF-pos5 model, the extraction performance of BiLSTM-CRF model is higher than that of the BiLSTM-CRF-pos5 for "AGE", "PRO", and "STATE" tags.

Table 4: Extraction Performance of Each Model for NAME

Models	Precision	Recall	F1-measure
baseline	77.79	60.62	68.07
uOne	77.73	71.94	74.68
uAll	78.29	73.65	75.87
bOne	79.23	67.51	72.87
bAll	78.93	67.46	72.71
fOne	79.10	61.80	69.34
fAll	78.14	62.58	69.43
uAll-bAll	78.95	74.50	76.64
uAll-bAll-fOne	78.54	74.53	76.46
uAll-bOne-fOne	78.59	74.79	76.61
BiLSTM-CRF	84.24	90.99	87.48
BiLSTM-CRF-pos5	85.40	**91.47**	**88.30**
BiLSTM-CNN-CRF	85.57	88.16	86.82
BiLSTM-CNN-CRF-pos5	85.93	89.50	87.66
BiLSTM-CRF-L	85.79	88.36	87.04
BiLSTM-CRF-L-pos5	85.84	88.60	87.19
Char-BiLSTM-CRF	85.81	89.75	87.72
Char-BiLSTM-CRF-pos5	**86.12**	90.61	88.29

5 Discussion

We analyze the features of extraction errors. Figure 5 shows the percentage of each error by the baseline model, the best CRF model (uOne-fAll model), and the best deep learning model (BiLSTM-CRF

Table 5: Extraction Performance of the Three Models for Each Tag

Tag	Baseline			uOne-fAll			BiLSTM-CRF		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure	Precision	Recall	F1-measure
NAME	77.79	60.62	68.07	78.23	72.96	75.47	84.24	90.99	**87.48**
MF	94.41	93.33	93.72	93.19	96.56	94.79	96.06	96.03	**95.95**
AGE	92.36	84.76	88.21	91.37	89.47	90.25	92.56	92.83	**92.62**
STATE	53.67	19.59	28.22	59.45	29.64	39.10	58.85	57.72	**57.98**
PRO	68.22	47.93	56.19	71.51	59.55	64.92	80.39	79.92	**80.14**
AFF	69.42	42.48	52.42	72.07	50.16	58.89	72.45	71.51	**71.86**
OTHER	58.78	28.53	38.19	63.72	38.63	47.86	63.62	62.58	**63.02**
PLACE	66.50	43.80	52.72	67.64	53.99	59.99	73.30	78.78	**75.89**
REL	84.35	59.53	69.71	81.80	69.48	75.10	82.99	83.93	**83.33**

model). The extraction errors are classified into the following five types.

- An error of labeling the character information as "O" (ne2oMiss)
- An error of labeling character information tags to a sequence except for character information (o2neMiss)
- A range of labeling is correct, however, a type of the character information tag is wrong (classMiss)
- A type of the character information tag is correct, however, a range of labeling is wrong (rangeMiss)
- An error of labeling both a tag type and a range (r&cMiss)

From "ne2oMiss" of the baseline and the uOne-fAll models in Figure5, it can be seen that the baseline model often tags "O" even if a sentence contains character information. On the other hand, the BiLSTM-CRF model has a low percentage of "ne2oMiss" and a high percentage of other error classes as compared to the CRF models. We consider that errors other than ne2oMiss are possible to modify using other information comparatively. From these results, we found that the CRF models and the deep learning model differ in the tendency of extraction errors, and the deep learning model has more possibility for labeling the character information correctly than the CRF models.

From the results of error analysis for each tag, we found that there is a tendency to mistake "NAME", "AFF", and "PLACE" for two different tags of them. After analyzing the details of the extraction errors, we confirmed that sequences related to three tags often contains Katakana in Japanese. Table 6 shows the percentage of the extraction errors containing Katakana by tag. From Table 6, it can be seen that about 30 to 40% in extraction errors for "NAME", "AFF", and "PLACE" contain Katakana. In Japan, a lot of fantasy novels are produced in the motif of the West, and Katakana tends to be used in characters' names and names of places in the novels. There is a possibility that the extraction performance can be improved by using the information of Katakana.

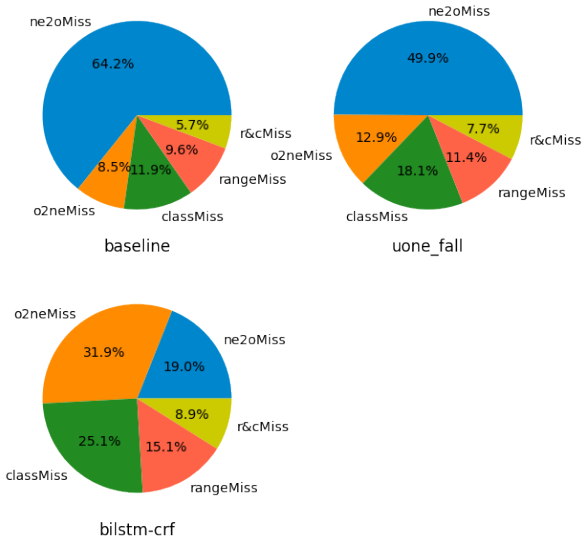


Figure 5: Extraction Errors by Three Models

6 Additional Experiment

Since synopses of fantasy novels in Japanese are targeted for extraction in this paper, names such as characters, organizations, and places tend to be written in Katakana. Focusing on names containing Katakana, We confirmed that there are names consisting of only Katakana and names consisting

Table 6: Extraction Errors Containing Katakana

Tag Types	Percentage of Katakana
NAME	**44.68**
MF	0.0
AGE	4.35
STATE	10.38
PRO	21.02
AFF	**37.10**
OTHER	20.42
PLACE	**33.33**
REL	7.47

of Katakana and other character types such as Hiragana and Kanji. Therefore, we construct a kChar-BiLSTM-CRF-pos5 model with the new character embeddings generated by compressing a sequence of Katakana to a single character and examine the performance of the model by comparing the best deep learning model for "NAME" (BiLSTM-CRF-pos5 model), the Char-BiLSTM-CRF-pos5 model and the kChar-BiLSTM-CRF-pos5 model.

Table 7 shows the results of experiments conducted with the same settings as in Section 3. The kChar-BiLSTM-CRF-pos5 model obtained all value was lower than that of the Char-BiLSTM-CRF-pos5 model. Therefore, it can be said that the extraction performance for NAME is not improved by the new character embeddings focusing on Katakana.

Comparing the extraction errors of the Char-BiLSTM-CRF-pos5 model with the kChar-BiLSTM-CRF-pos5 model, the latter can predict suitable tags for the sequences which appear more than once in the training data. However, we found a tendency that it was difficult to predict suitable tags for sequences never appeared in the training data. From the results, it is thought that the generalization ability of the model was degraded because the amount of information used for prediction was reduced by compressing a sequence of Katakana to a single character.

Table 8 shows the values of F1-measure of three models. By comparing the Char-BiLSTM-CRF-pos5 model with the kChar-BiLSTM-CRF-pos5 model, the value of F1-measure for "AFF" by the model with the new character embeddings is about 0.5 point higher than the normal model, however, the values of F1-measure for "NAME " and "PLACE" of the the model with the new character embeddings are about 2 and 0.5 point lower, respectively. The extraction performance for all tags by the kChar-BiLSTM-CRF-pos5 model is lower than that of the best deep learning model for "NAME", BiLSTM-CRF-pos5. From the results, the new character embeddings focusing Katakana did not much contribute to extraction performance. We will continue to consider the effective use of Katakana information.

Table 7: Extraction Performance for NAME by the Best Deep Learning Model with New Character Embeddings

Model	Precision	Recall	F1-measure
BiLSTM-CRF-pos5	85.40	**91.47**	**88.30**
Char-BiLSTM-CRF-pos5	**86.12**	90.61	88.29
kChar-BiLSTM-CRF-pos5	85.25	86.70	85.95

Table 8: F1-measure of Three Models on Extracting "NAME", "AFF", and "PLACE"

Model	NAME	AFF	PLACE
BiLSTM-CRF-pos5	**88.30**	**71.73**	**76.97**
Char-BiLSTM-CRF-pos5	88.29	69.73	72.98
kChar-BiLSTM-CRF-pos5	85.95	70.21	72.49

7 Conclusion

In this paper, we have compared methods for extracting novel character information using sequence labeling by CRF and by deep learning with CRF. The model with the part-of-speech vectors added to the input of the BiLSTM-CRF model has achieved the best performance for extracting "NAME", and the BiLSTM-CRF model has achieved the best overall performance for extracting all tags. Focusing on that Katakana tends to be used in "NAME", "AFF", and "PLACE" in the fantasy novels, we have considered the extraction performance of the kChar-BiLSTM-CRF-pos5 model with the character embeddings generated by compressing a sequence of Katakana to a single character, however, it could not confirm the beneficial effect.

In the future, we are going to consider the effective use of Katakana information, expand the data set, and consider a method of linking a character's name to other character information.

References

Takaaki Yoneda, Takahiro Shinozaki, Yasuo Horiuti, and Shingo Kuroiwa. 2012. Extraction of Novel Characters using Predicate Information (述語情報を利用した小説の登場人物の抽出). In *Proceedings of the Annual Meeting of the Association for Natural Language Processing*. pp.855-858. (in Japanese)

Kozue Baba and Atsushi Fujii. 2007. Extraction and Systematization of Character Information from Novel Text (小説テキストを対象とした人物情報の抽出と体系化). *In Proceedings of the Annual Meeting of the Association for Natural Language Processing*. pp.574-577. (in Japanese)

Zhiheng Huang, Wei Xu, and Kai Yu, 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv: 1508.01991*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Dyer Chris. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. pp.97-102.

Gustavo Aguilar, Adrian Pastor López Monroy, Fabio González, and Thamar Solorio. 2018. Modeling Noisiness to Recognize Named Entities Using Multitask Neural Networks on Social Media. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. volume 1. pp. 1401–1412.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the Empirical Methods in Natural Language Processing*. pp.230–237.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating Character Descriptions for Automatic Summarization of Fiction. In *Proceedings of the ThirtyThird AAAI Conference on Artificial Intelligence, 2019*. pp. 7476–7483.

K Vani and Alessandro Antonucci. 2019. NOVEL2GRAPH: Visual Summaries of Narrative Text Enhanced by Machine Learning. In *Proceedings of Text2Story - Second international Workshop on Narrative Extraction from Text co-located with 41th European Conference on Information Retrieval*. pp.29-37.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp.1534-1544.

Elias Iosif and Taniya Mishra. 2014. From Speaker Identification to Affective Analysis: A Multi-Step System for Analyzing Children ' s Stories. In *Proceedings of the Third Workshop on Computational Linguistics for Literature*. pp.40-49.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pp.1805-1806.

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam, Gintarė Grigonytė and Sofia Gustafson Capková. 2018. Identifying Speakers and Addressees in Dialogues Extracted from Literary Fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Redefining Verbal Nouns in Japanese: From the Perspective of Polycategoriality

David Y. Oshima
Nagoya University
Furo-cho, Chikusa-ku
Nagoya, Japan 464-8601
davidyo@nagoya-u.jp

Midori Hayashi
Chukyo University
101 Tokodachi, Kaizu-cho
Toyota, Japan 470-0393
mhayashi@lets.chukyo-u.ac.jp

Abstract

Japanese has a grammatical class commonly referred to as "verbal nouns (VN)", whose members form a phrasal verb in combination with the light verb SURU. While most VNs can be used as the head of a complement nominal (subject, object, etc.) of a predicate, some lack this use and are used only as part of a phrasal verb. Also, some lexemes that may function as a VN also may function as an adjectival noun, an adverb, etc. We report the results of a corpus-based survey on what patterns of polycategoriality are exhibited by those high-frequency lexemes that may function as a VN, and put forth some proposals as to how to classify and taxonomize "noun-like" categories in Japanese.

1 Introduction

In studies on the Japanese grammar, those lexemes (words) that form a phrasal verb in combination with the light verb SURU[1] 'do' have been referred to as verbal nouns (VNs) (Martin 1988, Kageyama 1993, Uchida and Nakayama 1993).

(Lexemes that may function as) a VN typically may function as a regular noun and serve as a complement nominal, but there are some exceptional items, which may be called "pure" or monocategorial VNs. Also, some VNs allow a use as an adjectival noun, an adnominal (nominal modifier), or an adverb. This work reports the results of a corpus-based survey inquiring the patterns of (mono- and) polycategoriality exhibited by high-frequency VNs, and puts forth some proposals as to how to classify and taxonomize "noun-like" categories in Japanese.

[1]Expressions in small capitals refer to lexemes.

2 What are verbal nouns?

Lexemes like CHOOSA (調査)[2] 'investigation', JANPU (< *jump*) 'jump', and UKETSUKE 'acceptance', can be used either as a component of a phrasal verb with SURU, as in (1a), or as the head of a complement nominal (often accompanied by a case particle such as nominative GA and accusative O), as in (1b,c).[3]

(1) a. Iseki o **choosa** shita.
ruins Acc investigation do.Pst
'(They) investigated the ruins.'
b. Iseki no **choosa** ga hajimatta.
ruins Gen investigation Nom begin.Pst
'The investigation of the ruins began.'
c. Iseki no **choosa** o yameta.
ruins Gen investigation Acc stop.Pst
'(They) canceled the investigation of the ruins.'

There are, on the other hand, lexemes that form a phrasal verb with SURU but are not used—or are used only marginally—as the head of a complement nominal.[4] Such words include: KYOOTSUU

[2]Japanese scripts are provided for Sino-Japanese lexical items (but not for native and Western ones).

[3]The abbreviations in glosses are: Acc = accusative, Attr = attributive, Cop = copula, Dat = dative, Neg = negation, Nom = nominative, Prs = present, Pst = past, Th = thematic *wa* (topic/ground-marker)

[4]To our knowledge, this type of lexemes has attracted rather scarce attention in the literature. Mizutani and Hoshino (1994), Mizutani (2001), and Nonaka (2009), however, make some rel-

(共通) 'commonality', NETCHUU (熱中) 'enthusiasm', DOOTEN (動転) 'perturbation', IPPEN (一変) 'drastic change', UROURO 'strolling', ENJOI (< *enjoy*) 'enjoying', and FITTO (< *fit*) 'fitting'.

(2) a. **Kyootsuu** suru ten ga sonzai
common do.Prs point Nom exist
shinai.
do.Neg.Prs
'There are no common features.'
b. ??**Kyootsuu** {ga/o} ...
common Nom/Acc

For convenience, we will say that "Lexeme α has a complement-nominal use" to mean that α has the potential to head a complement nominal of a wide range of predicates. The qualification with "a wide range of" (which admittedly is somewhat fuzzy and slippery) is needed to exclude items that serve as the head of a complement nominal only (i) with some exceptional predicates that select for clearly non-nominal words/phrases as case-marked complements, such as YOI 'good', as in (3), or (ii) with some predicates that form idioms with them, as in (4).

(3) {Nonbiri/shinsen/pikapika/chikara o
laid-back/fresh/spick-and-span/force Acc
awasete} ga ichiban **yoi**.
join.Ger Nom the.most good.Prs
'{(Doing it) in a laid-back way/(its being) fresh/(its being) spick-and-span/(doing it) joining (our) forces} is the best.'

(4) a. **Monogokoro** ga tsuku.
thing.sensation(lit.) Nom adhere.Prs
'(They) will reach the age of discretion.'
b. **Shinchoo** o kishita.
cautious Acc determine.Pst
'(They) were cautious.'

Whether a given word has a complement-nominal use undoubtedly has to do with its semantic nature (e.g., whether it refers to a concrete entity); at the same time, it is to a good extent a feature conventionalized on the lexeme-by-lexeme basis. The contrasts presented in (5b) and (6b), for example, are hard to explain in terms of semantic factors alone.

(5) a. Watashi wa {**konran/dooten**} shita.
I Th confusion/perturbation do.Pst
'I got {confused/upset}.'
b. {**Konran**/??**dooten**} ga osamaranai.
confusion/perturbation Nom settle.Neg.Prs
'(My) confusion won't settle down.' / ((My) perturbation won't settle down.')

(6) a. Jookyoo ga {(hageshiku)
situation Nom intense.Inf
henka/ippen} shita.
change/drastic.change do.Pst
'The situation {changed (drastically)/changed drastically}.'
b. {(Hageshii) **henka**/??**ippen**} o
intense.Prs change/drastic.change Acc
hikiokoshita.
cause.Pst
'(It) caused a (drastic) change.' / ((It) caused a drastic change.')

It is debabtable whether lexemes like KYOOTSUU (共通) and NETCHUU (熱中), which lack a complement-nominal use, are to be subsumed under the noun category, though we believe that it is by and large a matter of terminology. A similar issue has been raised with respect to so-called adjectival nouns (ANs), which come in two major varieties: (i) the *na*-type (or "AN-*na*" for short), such as ZENRYOO (善良) 'good', and (ii) the *no*-type (or "AN-*no*"), such as MURYOO (無料) 'free of charge' (Oshima *et al.* 2019). The *na*- and *no*-types have almost identical grammatical distributions, but are combined with different attributive copula forms, *na* and *no*, when occrurring in some environments including relative clauses.

(7) a. **zenryoo na** hito
good Cop.Attr person
'(a) good person'
b. **muryoo no** hon
free Cop.Attr book
'(a) free book'

ANs are noun-like in forming stative predicates combined with a copula, but are not noun-like in lacking a complement-nominal use. In line with Martin (1988), we adopt a broad definition of Japanese nouns, and take them to be those content

evant remarks.

words that form a predicate with the aid of a copula or a light verb.

VNs generally are understood as those content words that can be used to form a phrasal verb *in addition to* being used as the head of a complement nominal. This criterion, however, fails to cover words like KYOOTSUU (共通), NETCHUU (熱中), and DOOTEN (動転).

We propose to treat regular nouns (RNs) and verbal nouns as distinct categories, defining them as in the following:

(8) **Regular Nouns**: Those lexemes that have a complement-nominal use (i.e., can be used as the head of a complement nominal of a wide range of predicates).

(9) **Verbal Nouns**: Those lexemes that form a phrasal verb combined with the light verb SURU.

Lexemes like CHOOSA and JANPU, accordingly, are regarded as being polycategorial, having the potential to be used either as an RN or a VN. KYOOTSUU, NETCHUU, etc., on the other hand, have a use as a VN but lack one as an RN.

Many other types of polycategoriality (that involve "nouns" broadly understood) can be found in the Japanese lexicon. TOOMEI (透明) 'transparent' and DOKUTOKU (独特) 'peculiar' for example, can be combined with either of the copula forms *na* and *no*, and can be taken to be polycategorial across the *na*-type and *no*-type ANs, for which we assume the following definitions.[5]

(10) **Adjectival Nouns**: Those lexemes that lack a complement-nominal use but can form a predicate combined with the copula DA.
   a. ***Na*-Type**: Those adjectival nouns that select for the attributive copula form *na*.
   b. ***No*-Type**: Those adjectival nouns that select for the attributive copula form *no*.

To give some other examples, (i) KENKOO (健康) 'health(y)' is polycategorial across the RN and the AN-*na*, (ii) CHOKKAKU (直角) 'right angle, perpendicular' is polycategorial across the RN and the AN-*no*, and (iii) SHINPAI (心配) 'worry, worrisome' is polycategorial across the RN, the AN-*na*, and the VN.

The major features of the four nominal categories discussed so far are summarized in Table 1.

3 A lexical survey

We probed the patterns of polycategoriality exhibited by high-frequency lexemes that may function as a VN, utilizing the Tsukuba Web Corpus[6] (TWC; approximately 1.1 billion words) and the frequency list of words associated with the Balanced Corpus of Contemporary Written Japanese[7] (BCCWJ; approximately 100 million words).

The target of the survey was the 1,820 lexemes that (i) are in or higher than the 10,000th place in the aforementioned frequency list, or in other words are among the 10,008 lexical items occurring at least 555 times in the BCCWJ (henceforth, these 10,008 items will be referred to as the "top tier"), and (ii) meet the criterion specified in (11).

(11) **The working criterion for identifying VNs**: Those lexemes α such that the form

[5] Under some limited conditions, SURU in a VN phrasal verb allows alternation with DA, as in (i) (Sato 2014, Kubota 2018).

(i) a. Ano hito wa ni-ji ni shuppatsu {a. **suru** / b. **da** }.
that person Th two.o'clock Dat depature do.Prs / Cop.Prs
'That person will leave at two o'clock.'

b. ni-ji ni shuppatsu {a. **suru** / b. **no** } hito
two.o'clock Dat depature do.Prs / Cop.Attr person
'the person who will leave at two o'clock'

We take the potential to participate in the construction instantiated by (ib) and (iib)— the "VN+DA" construction, to give it a tentative label—not to count as the potential to "form a predicate combined with the copula DA", part of the qualifying conditions of ANs. That is, lexemes like NETCHUU, DOOTEN, and IPPEN may participate in the "VN+DA" construction, but this does not entail that they may function as ANs.

[6] International Student Center, Tsukuba University (2013–2019); http://nlt.tsukuba.lagoinst.info/. NINJAL-LWP (National Institute for Japanese Language and Linguistics & Lago Institute of Language 2012–2019) was used as the search tool.

[7] National Institute for Japanese Language and Linguistics (2011–); http://www.ninjal.ac.jp/corpus_center/bccwj/. The utilized frequency list, available at https://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html, is based on "short unit words (SUWs)" (one of the units of expressions in the corpus) and is associated with Version 1.1 of the corpus.

	RN	VN	AN-*no*	AN-*na*
has a complement-nominal use	+	−	−	−
forms a phrasal verb with SURU	−	+	−	−
forms a predicate with DA	+	−*	+	+
co-occurs with the attributive copula form:	*no*	(*no*)	*no*	*na*

*But see Note 5.

Table 1: The distributional properties of RNs, VNs, and the two types of ANs

"α+SURU" occurs more than 50 times in the TWC.

Most of the 1,820 lexemes ("the target VNs") belong to one of the two BCCWJ tag groups: (A) [名詞-普通名詞-サ変可能] (common nouns that may function as a VN) and (B) [名詞-普通名詞-サ変形状詞可能] (common nouns that may function as a VN and as an AN-*na*); however, two of them, IRAIRA 'irritat{ion/edly}' and UKKARI 'careless{ness/ly}', belong to (C) [名詞-普通名詞-副詞可能] (common nouns that may function as an adverb), and 25, including YUKKURI 'rest, slowly' and IKIIKI 'vivid{ness/ly}", belong to (D) [副詞] (adverbs).

Some items, including ZEHI (是非) 'right-and-wrong, no matter what', SOOTOO (相当) 'considrabl{e/y}', and IKKEN (一見) '(at first) glance', are listed twice in the frequency list, assigned either (i) tags (A) and (C) or (ii) tags (B) and (C), as if they each corresponded to two homophonous lexemes. In our survey, such items were treated as single lexemes, their instances lumped together.

Among the top-tier lexemes tagged either as (A) or (B), 50 do not meet criterion (11) and were excluded from the target VNs; they include CHOKUSETSU (直接) 'direct(ly)', TEKITOO (適当) 'suitable', KENSATSU (検察) 'prosecute(r)', TONNERU (< *tunnel*) '(in baseball, etc.) failure to field a grounder', and DATOO (妥当) 'appropriate'.

Also, the items listed in (12) were not identified as a lexeme that may function as a VN, despite meeting criterion (11).

(12) a. (tag group (A))
(i) HANASHI 'speech, story'[8]
(ii) MONO 'thing', KOKORO 'sensation', NAMIDA 'tear', ATAI 'value', EKI (益) 'benefit', KOI 'love'[9]
b. (tag group (D))
CHANTO 'properly', KICHINTO 'neatly', JITTO 'fixedly', HOTTO 'with a feeling of relief', HYOTTO 'by chance', HATTO 'gaspingly', BOOTTO 'dimly', KATTO 'ill-temperedly'[10]

4 Polycategoriality that involve the VN category

Most lexemes that may function as a VN also may function as a RN. Also, some lexemes are polycategorial across the VN and such other categories as the AN and the adverb. In this section, we will examine the patterns and tendencies as to polycategoriality involving the VN, referencing the results of the lexical survey.

4.1 VN/RN polycategoriality

Oftentimes, whether a given lexeme has a use as an RN (as defined in (8)), cannot be determined with crisp judgements. Accordingly, identifying lexemes that are used as a VN {as well as/but not} as a RN in a consistent and objective manner is not a straightforward task.

As a tentative measure of how easily a lexeme that has a use as a VN can be used as a complement

[8]Many hits of "HANASHI+SURU" in the TWC are actually misparses of forms of the verb HANASU 'speak' (e.g., *hanashite* (話して) being misparsed as *hanashi shite*). Also, HANASHI is not regarded as having a use as a VN in most contemporary monolingual dictionaries.

[9]We take MONOSURU, KOKOROSURU, etc. to be compound lexemes, rather than consisting of a VN and a light verb, in line with the treatment in most contemporary monolingual dictionaries.

[10]We consider a structure like "CHANTO+SURU" or "KICHINTO+SURU" to consist of an adverb, rather than a VN, and a light verb. One piece of evidence for this is that SURU participating in this structure never undergoes alternation with DA (a phenomenon explained in Note 5).

nominal, we introduce the "R/V index" defined as follows (α is a lexeme, and # is to be read as "the number of occurrences"; GA and O are nominative and accusative case markers, respectively):

$$\alpha\text{'s R/V index} = \frac{(\#\text{ of "}\alpha\text{+GA"}) + (\#\text{ of "}\alpha\text{+O"})}{\#\text{ of "}\alpha\text{+SURU"}}$$

An extremely low R/V index suggests that the RN use is not allowed, or at least is marginal in comparison to the VN use. An extremely high index, in contrast, indicates that the VN use is marginal, if existent at all. The R/N index admittedly is not an immaculate measure for detecting VN/RN polycategoriality; an obvious shortcoming is that it may be heavily affected by some frequently occurring idioms or set phrases. Nevertheless, we believe that it helps us obtain good ideas about how common it is for a lexeme that may function as a VN to lack a use as a RN.

We calclulated the R/V indices of the 1,820 target VNs, using the TWC as sample data.[11] To exemplify, the R/N indices of CHOOSA (調査) 'investigation, survey', JANPU 'jump', and UKETSUKE 'acceptance' were 2.032, .292, and 3.355, respectively, while those of KYOOTSUU (共通) 'commonality', NETCHUU (熱中) 'enthusiasm', IPPEN (一変) 'drastic change', and UROURO 'strolling', which were mentioned in §2 as examples of lexemes lacking a complement-nominal use, were .001, .003, .002, and .002, respectively.[12] Note that the adopted criterion (11) guarantees that evey target VN has at least 50 occurences in the TWC; the target VN with the fewest occurrences was ZENSHUTSU (前出) 'previous mentioning, aforementioned', with 69 occurrences in total and the R/V index 6.154.

The threshold value (of the R/V index) for acknowledging/dismissing the VN/RN polycategoriality of a given lexeme cannot be set without a certain degree of arbitrariness. The target VNs with a value smaller than .01 (Tier A) are listed in (13), and the ones with a value in the range of: $.01 \leq x < .03$ (Tier B) are listed in (14), in the ascending order of the index.

(13) (**Tier A**; $n = 52$)
SHIMIJIMI 'keen{ness/ly}', MANKITSU (満喫) 'satisfaction', KUSHI (駆使) 'full use', UNZARI 'boredom', ATTOO (圧倒) 'overpowering', HAKKIRI 'cl{arity/early}', SHIKKARI 'secur{ity/ely}', KYOOTSUU (共通) 'common(ality)', KYOOSHUKU (恐縮) 'embarrassment', KITCHIRI 'precis{ion/ely}', TSUUYOO (通用) 'validity', RINSETSU (隣接) 'adjacency', BIKKURI 'astonishment', IPPEN (一変) 'drastic change', ZENJUTSU (前述) 'previous mentioning, aforementioned', UROURO 'strolling, aimlessly', TANNOO (堪能) 'satisfaction, skilled', CHOKKETSU (直結) 'direct connection', HANMEI (判明) 'ascertainment', IKKEN (一見) '(at first) glance', NETCHUU (熱中) 'enthusiasm', HEIKOO (並行) 'simultane{ity/ous}', GAITOO (該当) 'correspond{ence/ing}', GAKKARI 'disappointment', KOOJUTSU (後述) 'subsequent mentioning, to be mentioned later', IKKAN (一貫) 'consistency', NONBIRI 'rest, slowly', KI'IN (起因) 'cause', BATABATA 'bustling, noisily', BON'YARI 'vague{ness/ly}', KANSHIN (感心) 'admira{tion/ble}', KAN'AN (勘案) 'consideration', CHOKUMEN (直面) 'confrontation', BURABURA 'idl{ing/y}', SUKKIRI 'cl{arity/early}', JUUJI 'engagement', YUTTARI 'rest, slowly', IKIKI 'vivid{ness/ly}', SAPPARI 'plain{ness/ly}', HAIKEN (拝見) 'look', HISSORI 'silen{ce/tly}', AIYOO (愛

[11] In the TWC, VN phrasal verbs are often treated as words distinct from the corresponding (regular) nouns. CHOOSA (調査) 'investigation', for example, corresponds to two lemmas (lexical entries) in the TWC: (i) the "verb" ⟨*choosa suru*⟩ and (ii) the "noun" ⟨*choosa*⟩. (For convenience, angle brackets are used to refer to TWC lemmas.) The occurrences of "α+{GA/O}" correspond to the cases where ⟨α⟩ is regarded as a "noun" or as an "adverb" and immediately precedes GA/O. The occurrences of "α+SURU" consist of (i) the occurrences (including various conjugated forms) of ⟨α *suru*⟩ treated as a "verb", and (ii) the cases where ⟨α⟩ is regarded as a "noun" or as an "adverb" and immediately precedes ⟨*suru*⟩. More generally, when (what we consider to be) a single lexeme corresponds to multiple TWC lemmas, their occurrences were lumped for the purpose of the calculation of R/V indices. The following six pairs of lemmas too were treated as single lexemes: (i) ⟨*kooyoo* (紅葉) *suru*⟩/⟨*momiji suru*⟩, (ii) ⟨*yakedo suru*⟩/⟨*kashoo* (火傷) *suru*⟩, (iii) ⟨*tannoo* (堪能)⟩/⟨*kannoo* (堪能)⟩, (iv) ⟨*meeku* (メーク)⟩/⟨*meiku* (メイク)⟩, (v) ⟨*meeku suru* (メークする)⟩/⟨*meiku* (メイクする)⟩, and (vi) ⟨*ofu suru* (オフする)⟩/⟨*OFF suru* (OFFする)⟩. The first three pairs are considered by the TWC to be homographic but heterophonic, but the pronunciation intended in the sources are in many cases unclear.

[12] The other examples like DOOTEN (動転) 'perturbation' were not part of the top tier.

用) ‘regular use, regularly used’, IKKATSU (一括) ‘consolidat{ion/ed}’, SHIHAN (市販) ‘(on) public sale’, TSUUKAN (痛感) ‘acute realization’, KYOOCHOO (強調) ‘emphasis’, HIREI (比例) ‘proportion’, JUN’YOO (準用) ‘mutatis mutandis application’, ASSARI ‘plain{ness/ly}’, JUUSHI (重視) ‘serious consideration’, CHOODAI (頂戴) ‘receiving’, GOROGORO ‘loafing around, with rumbling’

(14) (**Tier B**; $n = 54$)
SENNEN (専念) ‘devotion’, MOKUGEKI (目撃) ‘witnessing’, KAGOO (化合) ‘combination’, GATCHI (合致) ‘consistency’, SENZAI (潜在) ‘latency’, ZAISEKI (在籍) ‘regist{ration/ered}’, FUZOKU (付属) ‘attach{ment/ed}’, HOOCHI (放置) ‘neglect’, YUUSEN (優先) ‘priority’, TEISHOO (提唱) ‘advocacy’, KURIKKU (< *click*) ‘clicking’, HAKKI (発揮) ‘manifestation’, GENSON (現存) ‘existence’, SHUTSUDO (出土) ‘unearthing’, HOOKATSU (包括) ‘inclusion’, CHAKUMOKU (着目) ‘attention’, TOOMEN (当面) ‘immedia{cy/te}, for the time being’, SHITTORI ‘moisture, mellowly’, SOOGUU (遭遇) ‘encounter’, TAIKOO (対抗) ‘competition’, SOOTOO (相当) ‘correspondence, considerable’, KOORYO (考慮) ‘consideration’, ZENKI (前記) ‘previous note, previously noted)’, HEIKOO (平行) ‘parallel(ism)’, SENKOO (先行) ‘precedence’, SOO’OO (相応) ‘correspondence, suitable’, KEIYU (経由) ‘passage’, HAN’EI (反映) ‘reflection’, TOOJOO (登場) ‘appearance’, ISSHO (一緒) ‘company, same’, JISAN (持参) ‘bringing’, GETTO (< *get*) ‘acquisition’, SHUSAI (主催) ‘hosting (of an event)’, TOOSAI (搭載) ‘loading’, MEIKI (明記) ‘specification’, KYOYOO (許容) ‘allowance’, TOOTATSU (到達) ‘attainment’, TEKIGOO (適合) ‘conformance’, HIROO (披露) ‘announcement’, RYUUI (留意) ‘attention’, DANGEN (断言) ‘assertion’, SEISHI (静止) ‘stillness’, SENREN (洗練) ‘elaboration’, ZAIGAKU (在学) ‘enrollment in school’, SANSHUTSU (算出) ‘calculation’, MITCHAKU (密着) ‘adherence’, DANNEN (断念) ‘abandonment’, AIKOO (愛好) ‘love’, HAKKAKU (発覚) ‘revelation’, FUKA (付加) ‘addition’, KOOAN (考案) ‘invention’, FUJOO (浮上) ‘surfacing’, TAHATSU (多発) ‘frequent occurrence’, KOOFU (公布) ‘proclamation’

We suggest that the items in these two tiers can safely be taken to lack the use as an RN or allow it only marginally.

4.2 VN/AN-*na* polycategoriality

Among the 1,820 target VNs, the 34 listed in (15) are noted to have a use as an AN-*na* in Nishio *et al.* (eds.) (2019), an acclaimed monolingual dictionary of Japanese with approximately 67,000 entries.[13]

(15) SHINPAI (心配) ‘worr{y/isome}’, ANTEI (安定) ‘stab{ility/le}’, HANTAI (反対) ‘opposi{tion/ng}’, ANSHIN (安心) ‘relie{f/ving}’, KYOOTSUU (共通) ‘common(ality)’, FUSOKU (不足) ‘insufficien{cy/t}’, MANZOKU (満足) ‘satisfact{ion/ory}’, SAIWAI ‘fortun{e/ate(ly)}’, KUROO (苦労) ‘trouble(some)’, OOPUN (< *open*) ‘open(ness)’, MEIWAKU (迷惑) ‘annoy{ance/ing}’, JAMA (邪魔) ‘obst{acle/ructive}’, SHITSUREI (失礼) ‘rude(ness)’, HETA ‘unskilled(ness)’, PITTARI ‘tight(ness/ly)’, OSHARE ‘stylish(ness)’, KURIA (< *clear*) ‘clear(ing)’, KANSHIN (感心) ‘admira{tion/ble}’, ZEITAKU (贅沢) ‘luxur{y/ious}’, ETCHI (< *H*) ‘sex, obscene’, UWAKI (浮気) ‘(prone to) adultery’, BINBOO (貧乏) ‘po{verty/or}’, OOBAA (< *over*) ‘exceed, exaggerated’, BOODAI (膨大) ‘swell up, huge’, GOODOO (合同) ‘combin{ation/ed}’, ITAZURA ‘mischie{f/vous}’, FURIN (不倫) ‘immoral(ity)’, TAIKUTSU (退屈) ‘bor{edom/ing}’, RANBOO (乱暴) ‘violen{ce/t}’, HEIKOO (平行) ‘parallel(ness)’, HEIKOO (並行) ‘simultane{ity/ous}’, POPPU (< *pop*) ‘popping, popular’, SOO’OO (相応) ‘suitab{ility/le}’, TANNOO[14] (堪能) ‘satisfaction, skilled’

[13] The potential for a lexeme to be used as an AN-*na* is reflected in the BCCWJ tag information, but there are some cases of discrepancies between it and the treatment in Nishio *et al.* (eds.) (2019).

[14] It is said that TANNOO (堪能) in the sense of ‘satisfaction’ and TANNOO in the sense of ‘skilled’ used to be distinct lexemes, the latter being a variant form of KANNOO (堪能) (Nishio *et al.* (eds.), 2019:964).

PITTARI is also used as an adverb (see below).

4.3 Polycategoriality across verbal nouns and *no*-type adjectival nouns and adnominals

In the literature, a clear consensus is yet to established as to which lexemes are to be regarded as (having a use as) an AN-*no* (Oshima *et al.* 2019).

The same holds true for the category subsuming KISSUI (生粋) 'native, pure' and KAISHIN (会心) 'satisfactory', which has a distribution similar to that of the AN-*no* but occurs only in a noun-modifying construction (relative clause). We refer to this category as the *no*-type adnominal, or "Adn-*no*".[15]

(16) ***No*-Type Adnominals**: Those lexemes that form a noun-modifying clause being accompanied by the attributive copula form *no*, but cannot be accompanied by other copula forms such as *da* and *de*.

We do not attempt here to exhaustively identify which target VNs have a use as an AN-*no* or Adn-*no*, and will merely point out a few examples. ISSHO (一緒) 'company, same', KIRAKIRA 'glitter(ingly)', GIZOO (偽造) 'forge{ry/d}', KIN'EN (禁煙) 'smoking cessation, nonsmoking', and HIGAERI '(going and returning) in one day' can be regarded as having a use as an AN-*no*.[16]

AIYOO (愛用) 'regular use, regularly used', TOKUTEI (特定) 'specifi{cation/ed}', DAIYOO (代用) 'substitut{ion/e}', KYOODOO (共同) 'cooperat{ion/ed}', and TOOMEN (当面) 'immedia{cy/te}, for the time being' can be regarded as having a use as an Adn-*no*.[17]

4.4 VN/adverb polycategoriality

We adopt (17) as the definition of the Japanese adverb category.

(17) **Adverbs**: Those lexemes that meet at least one of conditions (a)–(c).

a. **Null Type**: Those lexemes that are used to modify a predicate or a clause by themselves (e.g., KANARI 'considerably', TOTEMO 'very').
b. ***To*-Type**: Those lexemes that are used to modify a predicate or a clause being accompanied by *to* (e.g., DOODOO (堂堂) 'majestic(ally)' and YUUZEN (悠然) 'calm(ly)').
c. ***Ni*-Type**: Those lexemes that (i) are used to modify a predicate or a clause being accompanied by *ni*, but (ii) do not meet the definition of ANs (e.g., OMOMURO 'slowly', TOMI 'suddenly').

Among the target VNs, we consider the 32 listed in (18) to meet the definition above (the judgments may well fluctuate to some extent among speakers). All of them may function as a null type adverb, and some may function as a *to*-type and/or a *ni*-type as well.

(18) HAKKIRI 'cl{arity/early}', SHIKKARI 'secur{ity/ely}', YUKKURI 'rest, slowly', ZEHI (是非) 'right-and-wrong, no matter what', SOOTOO (相当) 'correspondence, considerabl{e/y}', GOOKEI (合計) '(in) total', SAIWAI 'fortun{e/ate(ly)}', IKKEN (一見) '(at first) glance', PITTARI 'tight(ness/ly)', SUKKIRI 'cl{arity/early}', NONBIRI 'rest, slowly', SAPPARI 'plain{ness/ly}', BON'YARI 'vague{ness/ly}', IRAIRA 'irrita{tion/tedly}', TOOMEN (当面) 'immedia{cy/te}, for the time being', ASSARI 'plain{ness/ly}', IKIKI 'vivid{ness/ly}', YUTTARI 'rest, slowly', DOKIDOKI 'pit-a-pat', KITCHIRI 'precis{ion/ely}', KIRAKIRA 'glitter(ingly)', NIKKORI 'smil{e/ingly}', WAKUWAKU 'excite{ment/dly}', UKKARI 'careless{ness/ly}', KIPPARI 'decisive{ness/ly}', SHITTORI 'moisture, mellowly', GOROGORO 'loafing around, with rumbling', HISSORI 'silen{ce/tly}', UROURO 'strolling, aimlessly', BATABATA 'bustling, noisily', BURABURA 'idl{ing/y}', SHIMIJIMI 'keen{ness/ly}'

5 Discussions and conclusion

Tier A (§4.1) accounts for 2.86% (52/1,820) of the target VNs, and Tiers A and B together account for

[15] Other types of adnominals include (i) the *taru*-type, such as KENRAN (絢爛) 'gorgeous' and YUUZEN (悠然) 'calm(ly)', and (ii) the *naru*-type, such as SETSU (切) 'eager(ly)' and TAE 'exquisite(ly)'.

[16] KIRAKIRA additionally has a use as an adverb (§4.4), and GIZOO, KIN'EN and HIGAERI have a use as a RN.

[17] TOOMEN additionally has a use as an adverb (§4.4).

5.82% (106/1,820). It seems fair to say that lexemes that have a use as a VN but lack a use as an RN are not uncommon (though much less common than lexemes that have a use as an RN but lack a use as an VN).

Among the Tier A lexemes, the ones in (19) (and possibly some others) can be regarded as monocategorial or "pure" VNs.[18]

(19) MANKITSU (満喫) 'satisfaction', KUSHI (駆使) 'full use', UNZARI 'boredom', ATTOO (圧倒) 'overpowering', KYOOSHUKU (恐縮) 'embarrassment', TSUUYOO (通用) 'validity', BIKKURI 'astonishment', IPPEN (一変) 'drastic change', HANMEI (判明) 'ascertainment', NETCHUU (熱中) 'enthusiasm', GAKKARI 'disappointment', IKKAN (一貫) 'consistency', KIIN (起因) 'cause', KAN'AN (勘案) 'consideration', CHOKUMEN (直面) 'confrontation', JUUJI 'engagement', HAIKEN (拝見) 'look', TSUUKAN (痛感) 'acute realization', KYOOCHOO (強調) 'emphasis', HIREI (比例) 'proportion', JUN'YOO (準用) 'mutatis mutandis application', JUUSHI (重視) 'serious consideration'

Lexemes with a low R/V index include many ideophones, most of which may function as an adverb. The following items in Tiers A and B are ideophones that can be used as an adverb:

(20) (**Tier A**) SHIMIJIMI 'keen{ness/ly}', HAKKIRI 'cl{arity/early}', SHIKKARI 'secur{ity/ely}', KITCHIRI 'precis{ion/ely}', UROURO 'strolling, aimlessly', NONBIRI 'rest, slowly', BATABATA 'bustling, noisily', BURABURA 'idl{ing/y}', IKIKI 'vivid{ness/ly}', HISSORI 'silen{ce/tly}', ASSARI 'plain{ness/ly}', GOROGORO 'loafing around, with rumbling'; (**Tier B**) SHITTORI 'moisture, mellowly'

It is plausible that for many (if not all) such lexemes, the use as an adverb is basic, and the use as a VN was derived from it.

Phrasal verbs with a VN with a low R/V index value appear to tend to be stative.[19] The members of Tiers A and B listed in (21) form a stative verb with SURU.

(21) (**Tier A**) HAKKIRI 'cl{arity/early}', SHIKKARI 'secur{ity/ely}', RINSETSU (隣接) 'adjacency', CHOKKETSU (直結) 'direct connection', HEIKOO (並行) 'simultane{ity/ous}', GAITOO (該当) 'correspond{ence/ing}', IKKAN (一貫) 'consistency', KIIN (起因) 'cause', CHOKUMEN (直面) 'confrontation', SUKKIRI 'cl{arity/early}', YUTTARI 'rest, slowly', IKIKI 'vivid{ness/ly}', HISSORI 'silen{ce/tly}', HIREI (比例) 'proportion', ASSARI 'plain{ness/ly}', GOROGORO 'loafing around, with rumbling'; (**Tier B**) GATCHI (合致) 'coincidence', SENZAI (潜在) 'latency', ZAISEKI (在籍) 'regist{ration/ered}', FUZOKU (付属) 'attach{ment/ed}', GENSON (現存) 'existence', SHITTORI 'moisture, mellowly', HEIKOO (平行) 'parallel(ism)'

Lexemes that may function as a VN but not as an RN have attracted scarce attention in the literature, and the contrast between "VNs" like CHOOSA (調査) 'investigation' and SHUPPATSU (出発) 'departure' on the one hand and ones like MANKITSU (満喫) 'satisfaction' and KUSHI (駆使) 'full use' on the other—the information that only the former can be used as a complement nominal—have tended to be ignored in existing dictionaries, reference grammars, etc. This is unfortunate from perspectives of both theoretical research and language education; it should be recognized VNs and RNs are distinct categories, essentially in the same way as, say, adverbs and RNs are.

References

Taro Kageyama. 1993. *Bunpoo to gokeisei*. Hituzi Syobo, Tokyo.

Haruhiko Kindaichi. 1950. Kokugo dooshi no ichibunrui. *Gengo Kenkyu*, 15:48–53.

[18](19) exclude those lexemes that can be reasonably suspected to have a use as an AN-*no* or Adn-*no*, as well as CHOODAI (頂戴) 'receiving', which is used in the idiomatic construction: "X (*o*) *choodai*" 'Give me X'.

[19]Here, stative verbs are taken to subsume (i) Kindaichi's (1950) Type I verbs and Type IV verbs and (ii) verbs that may function either as a Type I or Typr IV verb (Kinsui's 1994 "Type V").

Satoshi Kinsui. 1994. Rentai shuushoku no "-ta" ni tsuite. In Yukinori Takubo, editor, *Nihongo no meishi shuushoku hyoogen*, pages 29–65. Kurosio Publisher, Tokyo.

Kazumitsu Kubota. 2018. Dekigoto no hassei o arawasu meishi jutsugo bun. *Bulletin of Aichi Shukutoku University, Faculty of Letters, Graduate School of Letters*, 43:129–148.

Samuel E. Martin. 1988. *A Reference Grammar of Japanese*, 1st Tuttle ed. Tuttle, Rutland, VT.

Shizuo Mizutani. 2001. Sokuron kara mita goruidate. *Mathematical Linguistics*, 23(3):135–156.

Shizuo Mizutani and Kazuko Hoshino. 1994. Meishi kara fukushi made: Gorui no atarashii ichizuke. *Mathematical Linguistics*, 19(7):331–340.

Minoru Nishio, Etsutaro Iwabuchi, Shizuo Mizutani, Wakako Kashino, Kazuko Hoshino, Naoko Maruyama, editors. 2019. *Iwanami kokugo jiten*, 8th ed. Iwanami Shoten, Tokyo.

Hiroo Nonaka. 2009. Nihongo "VN suru koobun", "VN o suru koobun" no eigo shokuyooogo no eigo hinshi to nihongo no kanren ni tsuite. *Bulletin of Kiryu University*, 20:23–31.

David Y. Oshima, Kimi Akita, and Shin-ichiro Sano. 2019. Gradability, scale structure, and the division of labor between nouns and verbs: The case of Japanese. *Glossa: A Journal of General Linguistics*, 4(1): Article 41:1–36.

Yutaka Sato. 2014. Japanese passives with verbal nouns. 2016. In Mikio Giriko, Naonori Nagaya, Akiko Takemura, and Timothy J. Vance, editors, *Japanese/Korean Linguistics*, volume 22, pages 207–322. CSLI Publications, Stanford, CA.

Yoshiko Uchida and Mineaharu Nakayama. 1993. Japanese verbal noun constructions. *Linguistics*, 31(4):623–666.

Speech Recognition for Endangered and Extinct Samoyedic languages

Niko Partanen
University of Helsinki
Finland
niko.partanen@helsinki.fi

Mika Hämäläinen
University of Helsinki
and Rootroo Ltd
Finland
mika@rootroo.com

Tiina Klooster
Luua Forestry School
Estonia
tiinaklooster@gmail.com

Abstract

Our study presents a series of experiments on speech recognition with endangered and extinct Samoyedic languages, spoken in Northern and Southern Siberia. To best of our knowledge, this is the first time a functional ASR system is built for an extinct language. We achieve with Kamas language a Label Error Rate of 15%, and conclude through careful error analysis that this quality is already very useful as a starting point for refined human transcriptions. Our results with related Nganasan language are more modest, with best model having the error rate of 33%. We show, however, through experiments where Kamas training data is enlarged incrementally, that Nganasan results are in line with what is expected under low-resource circumstances of the language. Based on this, we provide recommendations for scenarios in which further language documentation or archive processing activities could benefit from modern ASR technology. All training data and processing scripts haven been published on Zenodo with clear licences to ensure further work in this important topic.

1 Introduction

Samoyedic languages are spoken in the Western Siberia, and Tundra Nenets also extends far to the European part of Northern Russia. These languages belong to the Uralic language family. All currently spoken Samoyedic languages are endangered (see Moseley (2010). Tundra Nenets is the largest language in the family, while some, such as Kamas, are extinct. Even extinct Samoyedic languages have, however, been documented to various degrees.

Documentation of Samoyedic languages has reached a mature stage in last decades. Three languages in this group have a recent monograph long grammars in English. These are Tundra Nenets (Nikolaeva, 2014), Forest Enets (Siegl, 2013) and Nganasan (Wagner-Nagy, 2018). Similarly resources on these languages have been steadily becoming available for researchers, often in connection with major language documentation projects, such as 'The word stock of the Selkup language as the main source of cultural and historical information about a moribund endangered native ethnicum of Western Siberia'[1], 'Enets and Forest Nenets'[2], 'Corpus of Nganasan Folklore Texts'[3], 'Documentation of Enets: digitization and analysis of legacy field materials and fieldwork with last speakers'[4], 'Tundra Nenets Texts'[5], 'Comprehensive Documentation and Analysis of Two Endangered Siberian Languages: Eastern Khanty and Southern Selkup'[6], 'Selkup Language Corpus (SLC)' (Budzisch et al., 2019), 'Nganasan Spoken Language Corpus' (Brykina et al., 2018), 'INEL Selkup Corpus' (Brykina et al., 2020), 'INEL Kamas corpus' (Gusev et al., 2019). Also Online Dictionary for Uralic Languages (Rueter and Hämäläinen, 2017) includes Tundra Nenets lexical material.

[1] https://cordis.europa.eu/project/id/INTAS2005-1000006-8411
[2] https://dobes.mpi.nl/projects/nenets/
[3] https://iling-ran.ru/gusev/Nganasan/texts/
[4] https://elar.soas.ac.uk/Collection/MPI950079
[5] https://elar.soas.ac.uk/Collection/MPI120925
[6] https://elar.soas.ac.uk/Collection/MPI43298

Our study here focuses on two of these languages: Nganasan and Kamas. Our topic is ASR, but we anchor this work into a wider context of computational resources and language documentation. The work has two goals: to examine feasibility of developing an ASR system for an extinct language, in the case of Kamas, and to investigate the usability of such a system in a real on-going endangered language documentation scenario that is presented by Nganasan. These scenarios are not wide apart from one another. Worldwide extinction of linguistic diversity has been recognized for the last 30 years (Krauss, 1992), and many languages are in a very endangered situation.

The models trained in this paper have been released and made openly accessible on Zenodo with a permanent DOI[7]. As both corpora used in our study are distributed with restrictive non-commercial license (CC-BY-NC-SA), we have also published our training and testing materials with the same license. We think open practices such as these will be gaining importance in the future, and we want to contribute to this development as well by our example (Garellek et al., 2020).

2 Context and Related Work

Despite the wide documentation activities, we have not witnessed large improvements in language technology and computational linguistics on these languages. Thereby one of the goals in this paper is to encourage further work on these languages, also through our published new datasets. There is a rule based morphological analyser for Nganasan (Endrédy et al., 2010), which, however, appears to be available only through a web interface, and is not open access.

What it comes to other Samoyedic languages, a rule based Tundra Nenets morphological analyser exists in the GiellaLT infrastructure (Moshagen et al., 2014)[8] with availability through UralicNLP (Hämäläinen, 2019). There are also early Selkup[9] and Nganasan[10] analysers in the same infrastructure. Also OCR models have been developed to target early writing systems on these languages (Partanen and Rießler, 2019b), with associated data package (Partanen and Rießler, 2018). This responds well to OCR challenges identified earlier for these languages by Partanen (2017). The vast majority of these languages have virtually no language technology at the moment, but as there are increasingly larger and larger corpora, the possibilities for future work are many.

One challenge in working with these endangered languages is that very few researchers are able to transcribe them confidently and accurately. In the past few years, however, speech recognition in endangered language context has seen significant improvements, especially in scenarios where there is only one single speaker. Adams et. al. (2018) report a reasonable accuracy under these circumstances already with just a few hours of transcribed data, with rapid increase in accuracy when there is more training data. They also present a comparison of models trained on different amounts of training data using Na and Chatino data, which also inspired our own comparative experiments.

We have also seen very recently large improvements in such systems on related Uralic languages, for example Zyrian Komi (Hjortnaes et al., 2020b; Hjortnaes et al., 2020a). We have also seen experiments where ASR is being integrated to the language documentation workflows, for example, in Papuan context (Zahrer et al., 2020). Most widely applied speech recognition systems have been Persephone (Adams et al., 2018), Elpis (Foley et al., 2018) and DeepSpeech (Hannun et al., 2014). In this paper, we present and discuss several experiments we have done using Persephone system.

3 Languages and Data

Nganasan is an endangered Samoyedic language spoken by Nganasans, a small ethnic group in Taimyr Peninsula, Northern Siberia (Janhunen and Gruzdeva, 2020). According to official statistics there are 470 Nganasans, from who approximately 125 speak the Nganasan language (Wagner-Nagy, 2018, 3,17). Despite languages endangerment, plenty of documentation work has been conducted (Leisio, 2006; Wagner-Nagy, 2014; Kaheinen, 2020). Largest available Nganasan corpus was published in 2018 (Brykina et al., 2018), and it was used in our study.

[7] https://zenodo.org/record/4029494
[8] https://github.com/giellalt/lang-yrk
[9] https://github.com/giellalt/lang-sel
[10] https://github.com/giellalt/lang-nio

Kamas is another Samoyedic language, representing the southern group of this branch of Uralic languages. Kamas was spoken in the slopes of Sayan mountains in the Central Siberia. It is believed that by the 19th century Kamas tribe consisted of only 130 people (Matveev, 1964). Kamas were forced to abandon their nomadic lifestyle in the beginning of 20th century, which, in connection with large societal changes, increased the contact with Russian speakers and led to a cultural assimilation (Klooster, 2015, 9). The last Kamas speaker was Klavdiya Plotnikova, who was born in 1895 in a small village of Abalakovo in Central Siberia. She worked with several linguists since 1960s, and this results in a sizable collection of Kamas recordings that are available in various archives.

In 2019, a corpus containing transcribed versions of these materials was published (Gusev et al., 2019). We used Klavdiya Plotnikova's part of the corpus in our Kamas experiments, as she contributes the vast majority of all Kamas materials that exists. In the Nganasan experiments, we used the data from three prominent speakers in the Nganasan Spoken Language Corpus, who are also mentioned in the Nganasan grammar based largely to the same materials (Wagner-Nagy, 2018, 30).

One of the most important preprocessing steps was to exclude from training all sentences that are longer than 10 seconds. This is a condition set by Persephone system, and a convention followed also in other investigations (Wisniewski et al., 2020, 30). Similarly, Hjortnaes et al (2020b) filtered Zyrian Komi corpus by this limit. This choice leaves open an obvious possibility to improve the current results. As the filtered portion of the corpus is relatively large, either finding a way to include longest segments into the training process, or splitting those into smaller units, would easily increase the amount of training data.

Preprocessing conventions were very similar with both corpora, although independent inspection of particularities of individual datasets was done. It is customary that we work with speech corpora includes an intensive preprocessing step. With the case of Kamas the work was greatly aided by having a specialist of Kamas in our team. In the case of Nganasan we worked primarily with the light shed by the project documentation, which was also an useful and realistic scenario.

As Nganasan corpus was significantly larger than Kamas, also more preprocessing was needed, probably reflecting the longer time frame where it has been created. We excluded all segments that were shorted than 400 milliseconds, and removed all empty annotations or annotations that contained only punctuation characters. There were several invisible Unicode space characters that were removed. Also all annotations that contained number written as number characters were excluded. We also removed from Nganasan corpus all instances of utterances that contained Cyrillic characters.

For both corpora the annotations that contained unclear words marked with HIAT conventions (Ehlich and Rehbein, 1976) were removed. When the transcriptions contained annotations for nonverbal expressions, including coughing or laughter, we chose to remove those extra annotations, but keep those transcriptions in the training data. Self-corrections were kept, but the hyphens and brackets around them were removed.

It has been asked previously to what degree the preprocessing of language documentation data can be automatized (Wisniewski et al., 2020). Based to our experience with these corpora we can say that a good deal of manual inspection and examination is necessary to understand how the raw data has to be processed to make it usable in ASR training. In our experience the actual transformation of corpus XML files into the structure expected by Persephone was relatively easy. Much more time was consumed by analysing the annotation conventions used in the corpus, and processing some of the mistakes in transcriptions. To this vein we can strongly recommend for different projects the approach suggested by (Partanen and Riessler, 2019a), where a team working with endangered languages of Barents region have integrated automatic testing and validation very deeply into their corpus workflows, thus ensuring the systematic and orderly presentation of the corpus.

4 Method and Experiment Design

Our model is a bi-directional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997), which is trained to pre-

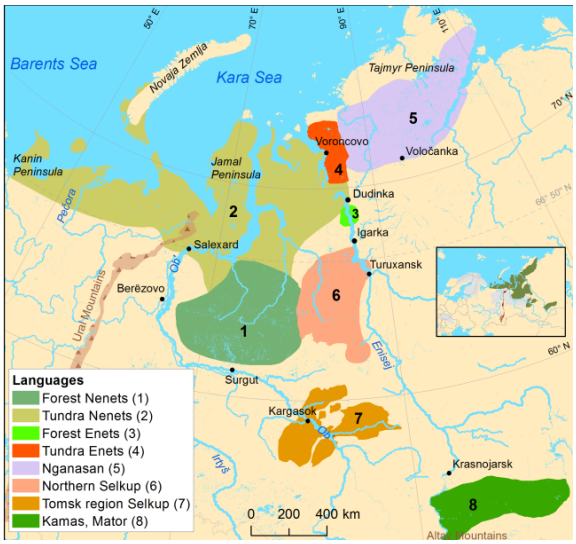


Figure 1: Distribution of the Samoyedic languages in the beginning of the 20 century (Timo Rantanen, BEDLAN)

dict the character sequences from audio input. The loss function used in the training is a connectionist temporal classification (CTC) in order to make it possible to train the model with only a coarse alignment between the audio and text (Graves et al., 2006). As suggested by (Wisniewski et al., 2020) and (Adams et al., 2018), we use 3 hidden layers with 250 hidden units. Using the same settings as other experiment have done maximises the comparative value of current work. We train our models by using Persephone framework (Adams et al., 2018).

4.1 Nganasan Tests

Although Nganasan corpus is fairly large, 28 hours according to corpus description, there are few individual speakers who are most prominent in the dataset. We selected all data from three such speakers, and trained individual models for all of them. To our knowledge experiments with Persephone system have not been carried out with tens of hours of data, and as earlier experiments have also focused to single speaker settings, we decided to continue along these lines. As the Nganasan speakers represent different amounts of transcribed data, the differences in accuracy can still give important information about this particularly low-resource setting.

4.2 Kamas Tests

After our preprocessing methods the Kamas corpus contains approximately 5 hours of utterances from a single speaker, above mentioned Klavdiya Plotnikova. Thereby the most obvious experiment to conduct is to train the model with all the speech we have from her, taking the original transcription as it is.

However, this also allows more varied experiments. One of the questions that are unclear in low-resource ASR is what to do with the word boundaries. In phoneme-level recognition usually practices with Persephone they often are omitted. Also in experiments done with other tools, such as DeepSpeech, specific language models have been used to insert spaces into correct places (Hjortnaes et al., 2020b). The second experiment with Kamas arises from this starting point: let's just leave the word boundaries as predicted labels.

One problem with the word boundaries must be their nature as a higher level construct. In real speech they often do not appear as pauses. As forced alignment tools have already developed relatively far, we opted for the MAUS system (Kisler et al., 2012) to align our data automatically. More specifically, we used the functionality provided in the emuR R package (Winkelmann et al., 2017). The language was left unspecified, and the alignment used grapheme to phoneme -mapping for the original transcriptions, and returned phoneme aligned SAMPA and IPA versions (Reichel, 2012; Reichel and Kisler, 2014). The mapping will be published openly with the other resources described here. It must be noted that there are minor differences between the original transcription and IPA representation. These are primarily about the how the long vowels are presented, as the original transcript was primarily split to characters, whereas the converted IPA has more phonemic units as individual labels.

This process gave us two more transcriptions versions: Plain IPA text, and an IPA version where only those word boundaries were retained which occurred within natural pauses. This work was highly experimental, and we did not correct the segmentations manually. Same Kamas data will be included in manually corrected DoReCo corpus (Paschen et al., 2020), which will allow better inspection of these

Experiment	Utterances	Minutes	LER
1	1152	108	0.334
2	512	57	0.930
3	704	43	0.892

Table 1: Nganasan experiments with three different speakers.

features. Our primary goal in this experiment was simply to investigate whether essentially very minor changes in transcriptions impact the result, and to see if different representation of word boundaries brings any benefits.

4.3 Gradual Data Augmentation Test

In order to evaluate the importance of the amount of transcribed minutes and hours, we designed additional tests. In these experiments we use the exactly same Kamas corpus as in the Experiment 6, but take only smaller portions of it, that are augmented gradually. As the maximum amount of data was close to 5 hours, we selected intervals that should represent realistically increasing corpus size, and thereby show where the most important thresholds lie.

These experiments are described in Table 3. While discussing the results we have also compared our error rates to those reported in other studies, in order to understand better how the variation we see connects to earlier studies with different languages.

5 Results

In this section, we present the results of the different models. These results are reported as a LER (label error rate) score. In practice, this is a measurement similar to CER (character error rate) that is widely used in studies focusing on text normalization and OCR correction (see (Tang et al., 2018; Veliz et al., 2019; Hämäläinen and Hengchen, 2019)).

5.1 Nganasan Results

In Nganasan experiment we selected utterances from three most prominent speakers. Table 1 shows the amount of data that we used and the accuracy reached.

We can easily conclude that the results were not successful in all experiments. In the cases where we had less than an hour of transcriptions, the quality was extremely low. When the label error rate is this high the model does not produce a useful result. However, there was a clear improvement with one speaker for who we had more training material. Brief example and discussion is provided in Section 6.

Exp.	Description	LER
4	Original transcript, no spaces	0.226
5	Original transcript, with spaces	0.195
6	IPA transcript, no spaces	0.149
7	IPA transcript, real pauses	0.243

Table 2: Kamas experiments with 4224 training samples, 266 minutes.

5.2 Kamas Results

Compared to the Nganasan experiment, the Kamas results are very different. Indeed, the results we achieve are very high, and on par with the best scores reported elsewhere for Persephone. We argue that the primary reason to this is sufficient amount of training data. Table 2 shows these results in detail.

In Experiment 4 we trained Persephone on the original Kamas transcriptions, without word boundaries separately marked, and with no modifications to the existing transcriptions. In the Experiment 5 the space characters were left to their original places. Surprisingly the result is significantly better with the word boundaries than without them.

Since Experiments 4 and 6 use extremely similar training data, just in different transcription system, we would had assumed the results to be very similar. We see, however, very large difference between the models. As we did not run the experiments multiple times, it is left open whether the difference can be caused from different random seeds. In our error analysis some possible reasons for these differences are discussed further.

The Experiment 5, however, was necessary in order to evaluate with more confidence the results of Experiment 7. Between these experiments the only difference was in detected pauses, instead of original word boundaries. The procedure was described in Section 4. As the Experiment 7 produces the worst results, we must conclude that this experiment was not successful. However, since the presence of word boundaries as their own tokens has small impact to the accuracy, and as they are useful information, this model may still be favoured in actual use.

Experiment	Utterances	Minutes	LER
6-1	448	28	0.612
6-2	896	57	0.254
6-3	1856	117	0.224
6-4	2816	177	0.176
6-5	3776	238	0.190

Table 3: Gradual data augmentation experiments

All Kamas models are relatively good, and the accuracy is inspected closer in Section 6. We see clearly in the Figure 2 that although there are minor differences, when the model has sufficient amount of training data the accuracy does not significantly change. We also cannot entirely exclude the possible impact of random run-time differences when the results are very close to one another.

5.3 Gradual Data Augmentation Test Result

The goal of this experiment was to investigate how the model's accuracy changes when the amount of training data is increased. In the past we have seen various tests with different corpora, often reaching very good results, as discussed in Section 2. The results of this experiment are presented in Table 3.

The major result we find here is that soon after containing two hours of training data, the models show extremely modest improvements. The largest improvement takes place between half an hour and a full hour. Especially when we compare the results to those reported for different languages by Wisniwski et. al. (2020), it appears that the amount of training data is the main denominator that impacts the models accuracy. Na language model is essentially as good as Kamas model, which reaches it's maximal accuracy after three hours of training data. There are possible exceptions, for example, Duoxu model is relatively good compared to its small size. Even then, it fits to the general curve very well.

Based on this comparison, more training data is not necessarily better, and the benefits decrease after certain level has been reached. We essentially have repeated the results of Adams et al. (2018) on Na and Chatino. We will discuss this further in Section 7, but this clearly gives us some guidelines of how much transcriptions are needed at the moment to achieve the best possible accuracy. This also contextualizes the Nganasan results, and explains why one of the models was much better than the others.

6 Error Analysis

Our error analysis focuses mainly on Kamas, since with this language we achieved a very high level of accuracy. In our error analysis the numbered lines in examples correspond to experiments described and numbered in Section 4. We can, however, state briefly that two Nganasan models with worst accuracy predict mainly short character sequences, essentially repeating the same fixed string. This prediction is, naturally, not useful. With the best Nganasan model, however, the result could already be useful as preliminary transcription stage. We see in Example (1) that the errors are primarily connected to vowel length, and most of the words and morphemes start to be recognizable.

(1) Mənə bəbəəd'əətəni̇ is'üðəm hüətə.
mənəbəbəd'ətəni̇s'ühuhətə
'I will be all the time at the old place.'

Next examples are all from the Kamas corpus. Some of the mistakes different models make seem to be systematic. Examples (2) and (3) show that especially consonant sequences are challenging for the model. Both /ll/ or /tt/ come out systematically as single consonants.

(2) Ujabə ajirbi mĭnzərzittə.
1: ujabajrbimĭnzərzitəo
2: ujabaj irbi mĭnzərzitə
3: ujabajirbimınzirzitə
4: ujabajirbimınzərzitəo
'He was reluctant to cook his meat.'

Especially in the second phoneme in Example (3) we see wide variation in the predicted vowel. This appears to be very common with reduced vowels.

(3) Mĭlle?bi, mĭlle?bi, ej ku?pi.
1: mĭle?timĭle?tiejku?piö
2: müle?pi mĭle?tə ej ku?pi
3: nule?bəmıle?təejku?pi
4: mıle?pimıle?piejku?pia
'He went, he went, he did not kill.'

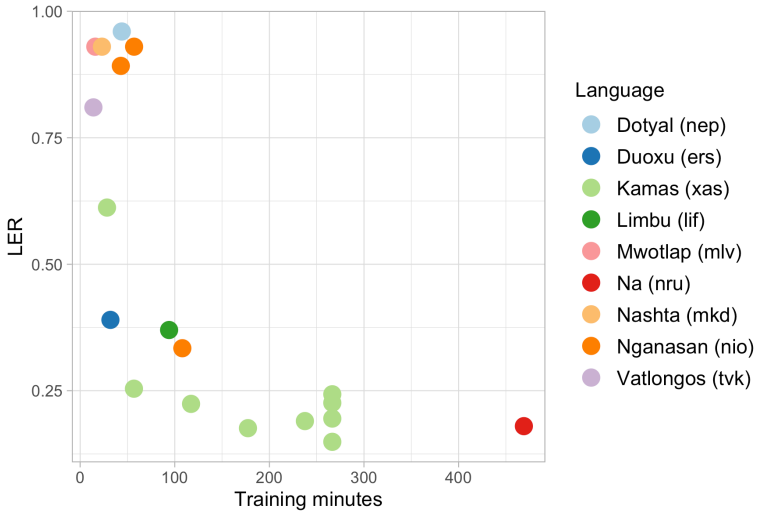


Figure 2: Results of our Nganasan and Kamas experiments compared with Wisniewski (2020)

In Example (4) we see how self-corrections have been treated in the original transcription and our various experiments. The model with original word boundaries is able to predict the spaces correctly, whereas the model with natural spaces only captures some of them. We see in this example also that none of the models recognize /b/ that is in the end of self correction. This plosive is not fully realized, which makes it acoustically very different from other instances of this phoneme. We can also notice that some of the models have a tendency to drop glottal stops, although some combinations show regularity.

(4) I dĭ poʔto (kuzab-) kuzazi mobi.
1: idĭpoʔtkuzakuzaziʔmobi
2: i dĭ poʔto kuza kuzaziʔ mobi
3: idıpoʔtokuzakuzazimobi
4: idı potəkuzakuzaziʔ mobia
‘This goat became man.’

Although not intended for ASR error analysis, we decided to use OCR tool Calamari’s error evaluation method (Wick et al., 2018). This gave us character level error information. Most frequent individual errors were related to letters /t/ and /l/. This seems to be related to them occurring as both short and long phonemes, and the models had particular problems to learn this distinction. Besides this we can see that many of the most prominent errors are related to vowels that share a very similar place of articulation or other properties: /e/ : /ə/, /o/ : /u/, /ɪ/ : /i/, /ø/ : /o/, /y/ : /u/, /ɐ/ : /a/. Within consonant system the errors between nasals /n/ : /m/ : /ŋ/ are frequent, and also glottal stop is often replaced with zero. We can also point that the error /t/ : /d/ is common, but other systematic errors related to voicing opposition cannot be found. These differences can be compared to the error analysis reported from Yongning language, where the confusions between characters also appeared to be related to acoustic realities (Michaud et al., 2020).

When our error analysis was repeated with the original transcription, which generally gave much worse results than IPA, very curious picture starts to emerge. Especially vowels containing diacritics were often misrecognized or omitted from prediction. This hints there is possibly something in this representation of the texts that does not pass correctly through the system and needs to be thoroughly investigated. Further research should be conducted with different character representations and refined data preprocessing. Also in-depth investigation of how the strings are passed to the ASR system internally could reveal more information about how, for example, different combining characters are treated. As we have published the training data and trained models openly, this examination can be continued easily.

7 Conclusion

We conducted several experiments on speech recognition of endangered and extinct languages. The most significant result is that we can identify a clear threshold of few hours after which the current models do not show clear decrease in the label error rate. We also show that differences in transcription system do not cause significant differences between models, although there is enough variation that ideal representation should be investigated. We do not see large differences in whether we use IPA or a project's internal transcription convention. As all these transcriptions are phonemic at fairly same level, the lack of differences is not surprising. Best results were achieved without word boundaries, but also the experiments with all word boundaries were encouraging enough that we would suggest testing a model trained with those intact. We aligned transcriptions to detect only word boundaries that correspond to natural pauses in speech, but the results of this were not better than in other experiments. We presume that the question here lies in the possibly weak quality of this automatic segmentation, and the experiment should be repeated with manually corrected version of the data.

As the transcription bottleneck is a major problem in linguistic fieldwork and documentation of endangered languages, our work sheds light to possible emerging working methods. This is also a question of resource allocation: when the language is rapidly disappearing, should we focus into recording more or into refinement and transcription of the existing materials? The situation is especially concerning when there are only a few individual elder speakers.

We hope our work offers new insight into this complicated question. Kamas has been extinct for more than 30 years, yet we are able to build a relatively good speech recognition model from just two hours of transcribed speech. This is a realistic amount, and not particularly much in the context where contemporary language documentation projects usually have budgets that cover several researchers work for multiple years.

Automatic transcription with ASR tools is a fast moving target. The results in few years will certainly be entirely different from what we are seeing now. Thereby making recommendations is also a complicated matter. However, we would argue, based to our results, that after one hour is transcribed for an individual speaker, training an ASR model to speed up the transcription work should already take place. In the same vein, this could be taken into account when working with endangered languages with very small speech communities. Recording and transcribing different speakers widely, with substantial transcription base for each speaker, seems to be the best way to take advantage from currently available ASR systems. We are clearly moving into a situation where manual transcription of everything is not the only option.

Future research should also focus into moving from single speaker systems into ASR that can work with multiple speakers, including unknown speakers. Very encouraging results were reported recently with only 10 minutes of training data, with the use of pretraining on unannotated audio and using a language model (Baevski et al., 2020). Since unannotated audio is available for virtually all language documentation projects, and also text corpora are becoming increasingly available and have proven useful (Hjortnaes et al., 2020a), there are certainly possibilities to experiment with these methods also in what it comes to language documentation context.

There is also evidence that systems other than Persephone could deliver better results, which is to be expected as the field evolves. Gupta and Boulianne (2020) reached a phoneme error rate of 8.7% on 3.1 hours of Cree training data, and their comparison of different systems showed significant improvement to other currently available methods, among those Persephone. This would suggest that there are possibilities to improve also from the Persephone results we have reported now.

The Persephone models that we have trained can be used with Cox's (2019) ELAN extension, or programmatically using Python. We have published both the models and the training datasets[11] in order to encourage further experiments on this important topic, and also to allow Nganasan researchers to benefit from our results. Although the majority of Kamas materials are already transcribed, we believe our results are relevant and valuable for the work being done with endangered and extinct languages.

[11] https://zenodo.org/record/4029494

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Maria Brykina, Valentin Gusev, Sandor Szeverényi, and Beáta Wagner-Nagy. 2018. Nganasan Spoken Language Corpus (NSLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 2018-06-12. Available online at http://hdl.handle.net/11022/0000-0007-C6F2-8.

Maria Brykina, Svetlana Orlova, and Beáta Wagner-Nagy. 2020. INEL Selkup Corpus. Version 1.0. Publication date 2020-06-30. http://hdl.handle.net/11022/0000-0007-E1D5-A. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmberg, editors, *The INEL corpora of indigenous Northern Eurasian languages*.

Josefina Budzisch, Anja Harder, and Beáta Wagner-Nagy. 2019. Selkup Language Corpus (SLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0.0. Publication date 2019-02-08. http://hdl.handle.net/11022/0000-0007-D009-4 .

Christopher Cox, 2019. *Persephone-ELAN: Automatic phoneme recognition for ELAN users*. Version 0.1.2.

Konrad Ehlich and Jochen Rehbein. 1976. Halbinterpretative arbeitstranskriptionen (hiat). *Linguistische Berichte*, 45(1976):21–41.

István Endrédy, László Fejes, Attila Novák, Beatrix Oszkó, Gábor Prószéky, Sándor Szeverényi, Zsuzsa Várnai, and Wágner-Nagy Beáta. 2010. Nganasan–computational resources of a language on the verge of extinction. In *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May 2010*, pages 41–44.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). In *SLTU*, pages 205–209.

Marc Garellek, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, Daniel Recasens, Timo Roettger, Adrian Simpson, et al. 2020. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science*, 9(1).

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Vishwa Gupta and Gilles Boulianne. 2020. Speech transcription challenges for resource constrained indigenous language cree. 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 362–367.

Valentin Gusev, Tiina Klooster, and Beáta Wagner-Nagy. 2019. INEL Kamas Corpus. Version 1.0. Publication date 2019-12-15. http://hdl.handle.net/11022/0000-0007-DA6E-9. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmberg, editors, *The INEL corpora of indigenous Northern Eurasian languages*.

Mika Hämäläinen and Simon Hengchen. 2019. From the paft to the fiiture: a fully automatic NMT and word embeddings method for OCR post-correction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 431–436.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020a. Improving the language model for low-resource ASR with online text corpora. In Dorothee Beermann, Laurent Besacier, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st joint SLTU and CCURL workshop (SLTU-CCURL 2020)*, pages 336–341, Marseille. European Language Resources Association (ELRA).

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020b. Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Juha Janhunen and Ekaterina Gruzdeva. 2020. Nganasan: A fresh focus on a little known Arctic language. *Linguistic Typology*, 24(1):181–186.

Kaisla Kaheinen. 2020. Nganasanin itsekorjaus: Huomioita korjaustoimintojen rakenteesta ja korjauksen merkityksistä vuorovaikutuksessa. MA thesis, University of Helsinki.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: the use case WebMAUS. In *Digital Humanities Conference 2012*.

Tiina Klooster. 2015. Individual language change: a case study of Klavdiya Plotnikova's Kamas. MA thesis, University of Tartu.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Larisa Leisio. 2006. Passive in Nganasan. *Typological Studies in Language*, 68:213.

AK Matveev. 1964. Kamassi keele jälgedel. *Keel ja Kirjandus*, 3:167–169.

Alexis Michaud, Oliver Adams, Christopher Cox, Séverine Guillaume, Guillaume Wisniewski, and Benjamin Galliot. 2020. La transcription du linguiste au miroir de l'intelligence artificielle: réflexions à partir de la transcription phonémique automatique.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing, 3rd edition. Online version: http://www.unesco.org/languages-atlas/.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era".

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. Walter de Gruyter GmbH & Co KG.

Niko Partanen and Michael Riessler. 2019a. Automatic validation and processing of ELAN corpora for spoken language data. Presentation in: Research Data and Humanities – RDHum 2019. University of Oulu, August 14–16, 2019. URL: https://www.oulu.fi/suomenkieli/node/55261.

Niko Partanen and Michael Rießler. 2019b. An OCR system for the Unified Northern Alphabet. In *The fifth International Workshop on Computational Linguistics for Uralic Languages*.

Niko Partanen and Michael Rießler. 2018. langdoc/iwclul2019: An OCR system for the Unified Northern Alphabet – data package, December. https://doi.org/10.5281/zenodo.2506881.

Niko Partanen. 2017. Challenges in OCR today: Report on experiences from INEL. In *Èlektronnaâ pis'mennost' narodov Rossijskoj Federacii: Opyt, problemy i perspektivy. Syktyvkar, 16-17 marta 2017 g.*, pages 263–273.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (doreco). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2657–2666.

Uwe D Reichel and Thomas Kisler. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, pages 42–49.

Uwe D Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *Proc. Interspeech*.

Jack Rueter and Mika Hämäläinen. 2017. Synchronized Mediawiki based analyzer dictionary development. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 1–7, St. Petersburg, Russia, January. Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Florian Siegl. 2013. *Materials on Forest Enets, an indigenous language of Northern Siberia*. Number 267 in Mémoires de la Société Finno-Ougrienne. Société Finno-Ougrienne.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Claudia Matos Veliz, Orphée De Clercq, and Véronique Hoste. 2019. Benefits of data augmentation for nmt-based text normalization of user-generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 275–285.

Beáta Wagner-Nagy. 2014. Possessive constructions in Nganasan. *Tomsk Journal of Linguistics and Anthropology*, (1):76–82.

Beáta Wagner-Nagy. 2018. *A grammar of Nganasan*. Brill.

Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari-a high-performance tensorflow-based deep learning package for optical character recognition. *arXiv preprint arXiv:1807.02004*.

Raphael Winkelmann, Jonathan Harrington, and Klaus Jänsch. 2017. EMU-SDMS: Advanced speech

database management and analysis in R. *Computer Speech & Language*, 45:392–410.

Guillaume Wisniewski, Alexis Michaud, and Séverine Guillaume. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from muyu. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2893–2900.

Neural Machine Translation from Historical Japanese to Contemporary Japanese Using Diachronically Domain-Adapted Word Embeddings

Masashi Takaku Tosho Hirasawa Mamoru Komachi Kanako Komiya

Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
{16t4027n, kanako.komiya.nlp}@vc.ibaraki.ac.jp

Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo, 191-0065, Japan
{hirasawa-tosho@ed., komachi@}tmu.ac.jp

Abstract

This paper describes the first trial of neural machine translation (NMT) from historical Japanese to contemporary Japanese. To compensate for the lack of parallel data, we used pre-trained word embeddings for the input of the system and performed diachronic domain adaptation in the order of time. We investigated and compared an NMT system without pre-trained word embeddings, an NMT system with pre-trained word embeddings trained with contemporary Japanese, an NMT system with word embeddings diachronically domain-adapted at one time, and NMT systems with word embeddings that were gradually domain-adapted in the order of time. Although our system did not outperform statistical machine translation, experiments revealed that diachronic domain adaptation is effective, especially if it is performed in the order of time.

1 Introduction

This paper describes a neural machine translation (NMT) system for translation from historical Japanese to contemporary Japanese. In recent years, machine translation using deep learning, or NMT, has been intensively studied. Although there is a study of statistical machine translation (SMT) from historical Japanese to contemporary Japanese (Hoshino et al., 2014), to the best of our knowledge, there have been no studies on an NMT system that translates historical Japanese to contemporary Japanese. NMT systems generally output fluent translations. Therefore, NMT is expected to improve the fluency of contemporary Japanese translation. However, it is difficult to obtain a model with high performance when only small parallel corpora are available, as NMT systems usually require large parallel corpora for training (Koehn and Knowles, 2017). Because the available parallel corpus of historical and contemporary Japanese is small, NMT is not appropriate for translation from historical Japanese to contemporary Japanese.

To improve translation performance, translation models are sometimes initialized with pre-trained word embeddings trained with a large corpus for language pairs that do not have sufficient parallel corpora (Qi et al., 2018). We believe that this method can also be effective for the translation model from historical Japanese to contemporary Japanese.

To obtain high-quality word embeddings, it is desirable to train them using a large training corpus. Therefore, the use of word embeddings trained with a contemporary Japanese corpus that is much larger than the historical Japanese corpus is expected. However, when the word embeddings trained with contemporary Japanese are directly used for the translation model, the model is expected to have poorer performance because the domains of the word embeddings and inputs differ from each other.

In addition, because the parallel corpus of historical and contemporary Japanese contains literature from different time periods, the meaning of the words or the words themselves may change according to the period. We thus propose a method for initializing the translation model with word embeddings generated from a contemporary Japanese corpus and perform gradually diachronic domain adap-

tation by fine-tuning using the corpus written in each period in the order of time (see Section 3). We compared other initialization methods in which pre-trained word embeddings were directly used or were fine-tuned using the entire historical corpus at one time (see Section 4).

The findings of this paper are listed as follows:

(1) The bilingual evaluation understudy (BLEU) score of the proposed method outperformed that of other methods, as discussed in Section 5,

(2) The quality of translation displayed improvement when diachronically domain-adapted word embeddings in the order of time were used; the model could translate words that appeared only in a corpus of a specific time period (see Section 6),

(3) However, the translation model that was diachronically domain-adapted up to a certain period did not always exhibit the best translation performance on the test set of that period, as discussed in Section 6.

2 Related Work

Hoshino et al. (2014) proposed a method for obtaining a sentence-based parallel corpus using a rule-based score function for aligning sentences from a paragraph-based parallel corpus of historical and contemporary Japanese. They translated historical Japanese to contemporary Japanese using a SMT system trained with the corpus obtained by the proposed method, and demonstrated the effectiveness of their proposed sentence aligning method. We believe that our study was the first to utilize NMT for translation from historical Japanese to contemporary Japanese.

Much research has been conducted on natural language processing using word embeddings, or distributed representations. Classification tasks, or sequence labeling tasks, using deep learning usually utilize pre-trained word embeddings; however, pre-trained word embeddings are rarely used in NMT. This is because the translation model itself learns suitable word embeddings when it is trained with a large parallel corpus. However, the initialization of inputs with pre-defined word embeddings offers the potential for improved translation performance when the translation model is trained with only a small parallel corpus. Qi et al. (2018) demonstrated that the use of pre-trained word embeddings for the training of NMT improved the translation performance for language pairs that had only a small parallel corpus.

For the domain adaptation method of word embeddings, we employed fine-tuning. Faruqui et al. (2015) proposed the retrofitting method, demonstrating that fine-tuning with another corpus improved the quality of the word embeddings. Yaginuma et al. (2018) performed word sense disambiguation in Japanese using fine-tuned word embeddings.

In addition, Kim et al. (2014) performed diachronic fine-tuning. They automatically detected changes in language over time through a chronologically trained neural language model. They obtained word embeddings specific to each year and demonstrated that some words had changed their meanings. Based on their research, we believe that diachronically domain-adapted word embeddings can capture changes in language meanings over time.

3 Diachronic Domain Adaptation of Word Embeddings Using Historical Corpus

In this study, we propose the initialization of inputs to the NMT model with diachronically domain-adapted word embeddings. Following the study by Kim et al. (2014), we chronologically fine-tuned the word embeddings starting from the newest corpora (contemporary Japanese corpus) to the oldest corpora. This is the domatin adaptation of time.

We fine-tuned the word embeddings in the order of time to minimize the drift in meaning over time. We employed the fine-tuning method used by Yaginuma et al. (2018).

We used the parallel corpus of historical and contemporary Japanese from four periods: the modern period (after the Edo period), Muromachi period, Kamakura period, and Heian period [1].

The procedures of diachronic domain adaptation are as follows:

[1] Edo, Muromachi, Kamakura, Heian, and Nara Periods are from 1603 to 1868, from 1336 to 1573, from 1185 to 1333, from 794 to 1185, and from 710 to 794, respectively. These periods are defined according to the political systems by historians.

(1) Fine-tune word embeddings pre-trained with contemporary Japanese using the modern corpus,

(2) Fine-tune the corpus obtained in step (1) using the Muromachi corpus,

(3) Fine-tune the corpus obtained in step (2) using the Kamakura corpus,

(4) Fine-tune the corpus obtained in step (3) using the Heian corpus.

4 Experiments

4.1 Model

In our experiments, we employed an encoder-decoder model[2] based on long short-term memory (LSTM) with attention. OpenNMT[3], an open-source NMT tool, was used for implementation. We utilized two unidirectional LSTM layers for the hidden layers and global attention (Luong et al., 2015) for attention. The vector sizes of the word embeddings and the hidden layers were set to 200 and 512, respectively, for both the encoder and decoder. Adam was used as the optimization algorithm, and the learning rate was set to 0.001. The vocabulary size treated by the model was limited to 20,000 for each of the source and target data, and unknown words were processed as `<unk>` tokens. The hyper parameters were determined according to preliminary experiments.

We initialized the weights of the word embedding layer of the translation model. In this study, the word embeddings pre-trained with the contemporary Japanese corpus were diachronically domain-adapted using the historical Japanese corpus, and used to initialize the weights of the word embedding layer of the encoder of the translation model. These word embeddings were also directly used for the decoder of the translation model.

We used the BLEU score (Papineni et al., 2002) to evaluate the translation model. Each method was given different seeds validated on each 5,000th step and was tested with the translation model with the highest BLEU score. The average BLEU scores over three trials using different seeds were evaluated as the scores of the translation models.

For comparison, we conducted experiments in which the weights of the word embedding layer were directly initialized with word embeddings pre-trained with the contemporary Japanese corpus without fine-tuning. In addition, we performed initialization with word embeddings pre-trained with the contemporary Japanese corpus and fine-tuned with the entire historical corpus at one time. Furthermore, we evaluated ensemble methods of models of diachronic domain adaptation at one time and models of diachronic domain adaptation in order of time. For the ensemble methods, we used the ensemble option of OpenNMT [4].

4.2 Data Set

We utilized a parallel corpus of historical and contemporary Japanese extracted by Hoshino et al. (2014) for translation. The sentences in this corpus were extracted from the corpus we used for fine-tuning: The Complete Collection of Japanese Classical Literature published by Shogakukan [5]. This corpus can be classified into five sub-corpora based on the periods in which each piece of literature was written. The statistics of the corpus are presented in Table 3. The periods in which each piece of literatures was written was determined by referring to the guide to the contents on the official website of the corpus [6].

Following Hoshino et al. (2014), we used three sub-corpora for the training and test data of the translation model. The modern Japanese corpus consisted of texts written after the Edo period, while the Kamakura corpus consisted of texts written in the Kamakura period. The Heian corpus consisted of texts written in the Heian period. The number of sentences in the modern, Kamakura, and Heian corpora were 4,577, 30,075, and 52,032, respectively, for a total of 86,684 sentences. The Muromachi corpus consisting of literature written in the Muromachi period was used only for fine-tuning because previ-

[2] We also tried a transformer model but the performance greatly varied depending on each trial. Also, the averaged performance did not surpass the encoder-decoder model. Therefore we decided to use an encode-decoder model.

[3] https://github.com/OpenNMT/OpenNMT

[4] https://github.com/OpenNMT/OpenNMT-py/pull/732

[5] https://japanknowledge.com/en/contents/koten/

[6] https://japanknowledge.com/en/contents/koten/title.html

ous research did not use it for testing. We did not use literature written in the Nara period at all because previous research did not use it for testing. Literature written in this period is also not suitable for fine-tuning because it is the oldest literature.

We divided the parallel corpus into training, development, and test corpora following previous research. The number of sentences were 82,591, 2,093, and 2,093, respectively (see Table 1). We randomly selected the test sentences. The ratio of the number of sentences of the entire corpus for each period in which the sentences were written is identical to that of the text examples. The number of sentences in the modern, Kamakura, and Heian corpora was 4577, 30,075, and 52,032, respectively, totaling to 86,684 sentences. Therefore, the number of sentences in the modern, Kamakura, and Heian corpora was 123, 739, and 1231, respectively. (123:739:1,231) = (4,577:30,075:52,032)

The number of examples in the test set is presented in Table 2. We used MeCab v0.996[7] as a morphological analyzer and UniDic for Early Middle Japanese v1.3[8] (Ogiso et al., 2012) and UniDic v2.3.0[4] (Maekawa et al., 2010) as dictionaries for historical and contemporary Japanese, respectively. We limited the length of an input or output sentence to 100 words.

Historical Japanese	
Total Number of Sentences	86,684
Vocabulary Size	49,200
Number of Tokens	2,774,745
Contemporary Japanese	
Total Number of Sentences	86,684
Vocabulary Size	45,690
Number of Tokens	3,611,783

Table 1: Parallel corpus of historical and contemporary Japanese.The data for translation are extracted from parallel corpus of Complete Collection of Japanese Classical Literature (see Table 3).

4.3 Word Embeddings

We used NWJC2vec (Shinnou et al., 2017) for the word embeddings for contemporary Japanese.

[7] https://taku910.github.io/mecab/
[8] https://unidic.ninjal.ac.jp/

Period	Number of Example
Modern Test Set	123
Kamakura Test Set	739
Heian Test Set	1,231

Table 2: Number of test example according to period

These word embeddings were generated from the NWJC-2014-4Q dataset (Asahara et al., 2014), which is an enormous Japanese corpus developed using the word2vec tool (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c). Tables 6 and 4 present summary statistics for the NWJC-2014-4Q data and the parameters used to generate the word embeddings, respectively.

CBOW or skip-gram	-cbow	1
Dimensionality	-size	200
Number of surrounding words	-window	8
Number of negative samples	-negative	25
Hierarchical softmax	-hs	0
Minimum sample threshold	-sample	1e-4
Number of iterations	-iter	15

Table 4: Parameters used to generate NWJC2vec

We followed Yaginuma et al. (2018) for the parameters for fine-tuning NWJC2vec (see Table 5). The other parameters were set to the default settings.

CBOW or skip-gram	-cbow	1
Dimensionality	-unit	200
Number of surrounding words	-window	5
Number of negative samples	-negative	5
Batch size	-batchsize	1000
Number of iterations	-iter	10

Table 5: Parameters used to fine-tune NWJC2vec

5 Results

Table 7 presents the BLEU scores of the entire test data according to each model. *SMT* (Hoshino et al., 2014) refers to the results of Hoshino et al. (2014), who used an SMT system to perform translation from historical Japanese to contemporary Japanese. *Baseline* refers to the results of the NMT model when only the parallel corpus of historical and con-

	Total Number of Sentences	Vocabulary Size	Number of Tokens
Modern	22,485	25,584	544,293
Muromachi	12,640	14,931	386,101
Kamakura	35,020	29,062	933,190
Heian	59,744	29,520	1,543,102
Nara	4,832	6,013	112,094
Total	134,721	61,345	3,518,780

Table 3: Parallel corpus of Complete Collection of Japanese Classical Literature published by Shogakukan

Number of URLs collected	83,992,556
Number of sentences (Some are overlapped)	3,885,889,575
Number of sentence (No overlapping)	1,463,142,939
Number of words (tokens)	25,836,947,421

Table 6: Statistics for the NWJC-2014-4Q dataset

temporary Japanese was used without using pre-trained word embeddings. *NWJC2vec* refers to the results of the NMT model when NWJC2vec was directly input to the system without fine-tuning. *Entire historical corpus* represents the results of the translation model when NWJC2vec was fine-tuned with the entire historical corpus at one time. *+Ensemble* refers to the results of an ensemble method. We evaluated two types of ensemble methods. The first, *+Ensemble at one time* is an ensemble method using the top three translation models of the *Entire historical corpus*. *(1) Modern*, *(3) Modern → Muromachi → Kamakura*, and *(4) Modern → Muromachi → Kamakura → Heian* are the results of translation models in which NWJC2vec was diachronically domain-adapted in the order of time. We did not evaluate (2) Modern → Muromachi because previous research did not use texts written in the Muromachi period for the training and test data. The second ensemble method, *+Ensemble in order of time*, is the ensemble method of models (1), (3), and (4).

According to Table 7, the best method among all NMT systems was the ensemble methods using diachronic domain adaptation in the order of time. Particularly, it outperformed the ensemble method of the top three models of diachronic domain adaptation using the entire contemporary Japanese corpus at one time. In addition, all diachronic domain adaptation methods in the order of time surpassed the *Entire historical corpus*. These results imply that diachronic domain adaptation in the order of time is effective.

Moreover, all methods using fine-tuning outperformed the baseline, although NWJC2vec, the model that directly used the word embeddings pre-trained with contemporary Japanese, were unable to outperform the baseline. This result indicates that fine-tuning using the historical corpus is effective. However, even the best NMT model did not surpass the SMT model. We believe that this is because NMT requires more parallel data than SMT. The BLEU score of the best NMT method was more than 2 points higher than that of the baseline method. However, the differences between models of diachronic domain adaptation in the order of time and models of diachronic domain adaptation at one time were not very large.

In addition, Table 7 indicates that the BLEU score decreased for earlier time periods when the word embeddings were diachronically domain-adapted in the order of time. Table 8 lists the BLEU scores of the proposed method according to each period. According to this table, the translation model in which (1) only the modern corpus was used for fine-tuning was the best for the modern test set. In addition, (4) diachronic domain adaptation up to the Heian period was the best for the Kamakura test set and outperformed (3) diachronic domain adaptation up to the Kamakura period. Furthermore, (1) diachronic domain adaptation up to the modern period was the best for the Heian test set. The second best was (3) diachronic domain adaptation up to the Kamakura

Method	BLEU
SMT (Hoshino et al., 2014)	28.02
Baseline	19.22
NWJC2vec	19.16
Entire historical corpus	19.24
+Ensemble at one time	20.94
Diachronic domain adaptation in order of time	
(1) Modern	**19.43**
(3) Modern → Muromachi → Kamakura	19.33
(4) Modern → Muromachi → Kamakura → Heian	19.29
+Ensemble in order of time	21.59

Table 7: BLEU scores of entire test data according to each model

period, whereas (4) diachronic domain adaptation up to the Heian period was the worst.

6 Discussion

Although the BLEU score did not significantly improve for the proposed method, some examples demonstrated the effectiveness of the proposed method.

Table 9 presents translation examples in which (a) diachronic domain adaptation up to the Kamakura period was better and (b) the ensemble method of the proposed methods was better.

For example, (a) "やまとうた (Japanese poems)" could not be translated to "和歌 (Japanese poem)" until only the modern corpus was used for fine-tuning; however, it could be translated correctly after the Kamakura corpus was used. In this example, diachronic domain adaptation improved the translation.

Furthermore, example (b) in Table 9 demonstrates that the word "騒がしき (noisy)" was erroneously translated when the word embeddings were fine-tuned with the entire data set at one time, but were correctly translated when they were gradually domain-adapted in the order of time. This example also demonstrates the effectiveness of the proposed method.

Next, we consider the effects of domain adaptation in the order of time. We hypothesized that the translation model would exhibit the best performance when the model was diachronically domain-adapted up to the time when the test set was written; however, Table 7 indicates that this hypothesis was not always correct. When the proposed method was used, unknown words decreased for earlier time periods because the fine-tuning method we used following Yaginuma et al. (2018) added a new entry when the new corpus included a new word that exceeded the threshold value. However, (1) diachronic domain adaptation up to the modern period was the best not only for the modern test corpus but also for the Heian corpus. However, the difference between the models for the Heian corpus was rather subtle compared to that for the modern corpus or Kamakura corpus; it was only 0.13 for the Heian corpus but 1.65 for the modern corpus and 0.6 for the Kamakura corpus. In other words, for the Heian corpus, there was no large difference based on the translation model of time.

Future work will include the investigation of effects of other word embeddings, such as fastText and Glove. In addition, diachronic domain adaptation using contextual word representations, such as ElMo and BERT, would be interesting.

7 Conclusion

This paper is the first to present an NMT system that translates historical Japanese to contemporary Japanese. We proposed diachronic domain adaptation of word embeddings in the order of time. We gradually fine-tuned the word embeddings for the input of the system in the order of time using a corpus written in each period. The NMT results were unable to surpass the results of the SMT system due to the lack of sufficient parallel data. In addition, the hypothesis that the translation model should have

	Modern	Kamakura	Heian
(1) Modern	5.24	25.16	19.53
(3) Modern → Muromachi → Kamakura	4.09	25.65	19.43
(4) Modern → Muromachi → Kamakura → Heian	3.59	25.76	19.40

Table 8: BLEU scores of the proposed method according to each period

(a) An example where diachronic domain adaptation until Kamakura period is better	
Input Sentence:	やまとうた (Japanese poems) の 道 、 浅き に 似 て 深く 、
Reference translation:	和歌 (Japanese poems) の 道 は 、 浅い よう で い て じつ は 深く 、
English translation:	The soul of Japanese poems seems shallow, but it is in fact profound.
Baseline:	<unk> の 道 は 、 浅い の に 似 て 深く 、
Modern:	<unk> の 道 、 浅い の と 同様 に 、 深く て 、
Modern → Muromachi → Kamakura:	和歌 (Japanese poems) の 道 は 、 浅い 時代 に 似 て 深く 、
(b) An example where ensemble method of proposed methods is better	
Input sentence:	大饗 に 劣ら ず 、 あまり 騒がしき (noisy) まで なん 集ひ たまひ ける 。
Reference translation:	大 饗 の とき に 劣ら ない ほど 、 あまり に 騒がしい (noisy) まで 大勢 お 集まり に なる の だっ た 。
English translation:	There came together as many people as people at a royal party and it was almost too noisy.
Baseline:	大 饗 に も 劣ら ず 、 あまり に も あわただしい (hasty) くらい に お 集まり に なっ た 。
Entire historical corpus:	大 饗 に 負け ず 、 あんまり 暑い (hot) まで 集まっ て お いで に なっ た 。
Diachronic domain adaptation:	大 饗 に 劣ら ず 、 あまり に 騒がしい (noisy) まで 集まっ て お いで に なっ た 。

Table 9: Translation examples

the best performance when the input of the system is diachronically domain-adapted up to the period in whihc the test corpus is written was not always correct. However, the translation performance when the word embeddings were domain-adapted in the order of time was better than that when the embeddings were domain-adapted at one time. In addition, some examples in which diachronic domain adaptation improved the translations were observed.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17H00917 and a project of the Center for Corpus Development, National Institute for Japanese Language and Linguistics. We would like to thank Hikaru Yokono, who gave us helpful advice and the data set used in previous research.

References

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria*, 25(1-2):129–148.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June. Association for Computational Linguistics.

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.

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, Baltimore, MD, USA, June. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *In Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics*, pages 28–39.

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.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop 2013*, pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 1–9.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*, pages 746–751.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 2002 Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? 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 529–535.

Hiroyuki Shinnou, Masayuki Asahara, Kanako Komiya, and Minoru Sasaki. 2017. Nwjc2vec: Word embedding data constructed from NINJAL Web Japanese Corpus. *Journal of Natural Language Processing (In Japanese)*, 24(5):705–720.

Daiki Yaginuma, Kanako Komiya, Hiroyuki Shinnou, et al. 2018. WSD of domain adaptation by distributed representation of fine tuning. *IPSJ SIG Technical Reports (NL) (In Japanese)*, 2018(1):1–5.

Improving Semantic Similarity Calculation of Japanese Text for MT Evaluation

Yuki Tanahashi
WANTS Inc.
Nagoya, Japan
tanahashi236@wantsinc.jp

Kyoko Kanzaki
Toyohashi University of Technology
Toyohashi, Japan
kanzaki@cite.tut.ac.jp

Eiko Yamamoto
Gifu Shotoku Gakuen University
Gifu, Japan
eiko@gifu.shotoku.ac.jp

Hitoshi Isahara
Toyohashi University of Technology
Toyohashi, Japan
isahara@tut.jp

Abstract

In recent years, the quality of machine translation has significantly improved, and it is considered that translation of a novel becomes possible. However, when performing translation for novels by a machine translation, it needs an evaluation method that performs comparison using a vector of distributed representations. BERTScore is one of the methods for performing automatic evaluation using a distributed representation. In this research, we examined the optimal setting for applying the BERTScore to the evaluation of translations of Japanese novels, and improved the method by introducing penalties for named entities based on idf values calculated from large corpora. The introduction of the penalty has made it possible to mitigate the false matching of personal names caused by distributed representation. We verified the method by calculating the Pearson correlation between the modified BERTScore and human-rated scores. Furthermore, we set four BERT models and two kinds of corpora to calculate idf value, and investigated which setting is most suitable for evaluation of novel translation. As a result, the setting with the model based on novel corpus, the idf based novel corpus and the penalty had the highest correlation with human-rated scores.

1 Introduction

In recent years, machine translation quality has dramatically improved due to the development of neural translation models that utilize deep learning, such as the sequence transformation model (Sutskever et al., 2014) and the attention model (Dzmitry et al., 2015. Luong et al., 2015), which is an application of the attention mechanism, and improved computer performance. Due to these improvements, not only documents consisting of formal expressions such as patent sentences and academic papers that have been fixed to some extent but also informal expressions such as novels and colloquial expressions could be machine translated. However, previous research has revealed that problems that have not been considered as important in machine translation research so far have a great influence on the learning and results of novel translation. Among them, the variety of text expressions is a serious problem. The problem is that when the author or translator in a novel is different, or even if the same author/translator has a different story speaker, an English sentence is translated into a Japanese sentence with a distinctly different translation but with a similar meaning.
Specific examples are shown below.

English	Japanese
My name is John.	俺はジョンという。 (I am called John.)
My name is Maria	私の名前はマリアよ。 (My name is Maria.)

When a language resource (corpus) containing such parallel translations is used as learning data in neural translation that receives the whole sentence as input and learns so as to maximize the likelihood that a correct word sequence will be output, learning becomes difficult and the output of the translation system becomes unstable. In addition, different expressions cause problems not only in learning but also in translation performance evaluation. In machine translation, BLEU (Papineni et al, 2002) is the de facto standard as an automatic evaluation method for evaluating the performance of translation systems in many previous studies. BLEU is a n-gram matching that scores the translation quality between 0.0 and 1.0 by counting the number of matching words n-grams between the reference sentence that is the human-translated correct data and the sentence to be evaluated output by the translation system. BLEU is used in many machine translation studies because it is a simple and easy-to-interpret method, but it is necessary that the surface text strings of the words in the reference sentence and the sentence to be evaluated are exactly the same.

Therefore, even if two sentences appear to be semantically identical to each other by humans, BLEU gives a low rating if the words used are different (different expressions). In novel translation, where the description of expressions is likely to be different when the translator and the speaker in the story are different for a certain English sentence, even if there is a system that can translate high quality, BLEU will not perform correctly.

As described above, an automatic evaluation method such as BLEU that considers only the surface of a sentence cannot correctly evaluate two sentences that have different expressions but have similar meanings. Therefore, in order to facilitate future novel translation research in Japanese, it is necessary to consider an automatic evaluation method that can accurately evaluate novel sentences from a certain language to Japanese before developing a translation model.

We apply BERTScore (Zhang et al., 2019) to the semantic similarity evaluation of Japanese novels. BERTScore uses BERT (Devlin et al., 2018) which generates a general-purpose linguistic expression for automatic sentence similarity evaluation. In applying the BERTScore, we proposed a modification that reduces the problem of similarity calculation in Japanese novels by imposing the editing distance of word reading (pronunciation) as a penalty. In addition, in order to adapt the BERTScore to novel evaluation, we constructed a BERT pre-learning model using a monolingual novel corpus consisting of sentences collected from the novel posting site. This model, and other existing BERT pre-learning model were applied to BERTScore to investigate which model is most suitable for novel evaluation.

The contributions of this research are the following three points.

1. Investigation of optimal settings for applying BERTScore to Japanese novels
2. Clarification and correction of problems that occur when calculating the similarity of novel Japanese sentences using BERTScore
3. Construction of BERT pre-learning model using large-scale novel corpus

2. Applying BERTScore to Japanese

Zhang et al. conducted experiments on BERTScore using the test set provided in the Metric Shared Task of WMT2017, and confirmed its usefulness in sentence pair evaluation on the English side in several language pairs. However, its usefulness has not yet been verified in Japanese sentences in English-Japanese language pairs. Therefore, in this research, we search for the optimal setting for applying BERTScore to the sentence pairs on the Japanese side in English-Japanese language pairs.

Since this study has the goal of improving the translation evaluation of English-Japanese novel translations, we consider its application especially to the evaluation of Japanese sentences in novel sentences. In other words, we consider a method to correctly evaluate an example in which the reference sentence and the sentence to be evaluated have different expressions but the same meaning. Such differences in expressions can be absorbed to some extent by using a distributed expression vector optimized for the meaning of words created by BERT. However, the use of distributed expressions by BERT causes another problem for the evaluation of sentences that include proper nouns such as person names and place names. In BERT, when considering the

meaning of a word, a vector is defined by surrounding words and their arrangement. Consider the following two sentences;

"Mr. Tanaka bought a bottle of juice."
"Mr. Sato bought a bottle of juice."

These two sentences represent distinctly different situations for humans. Because the nouns that are the subject are different, i.e. Mr. Tanaka and Mr. Sato, these clearly indicate another person. However, since these sentences have the same sentence structure of "someone", "juice", and "buy", the words around "Tanaka" and "Sato" that correspond to "someone" are the same. Therefore, the vectors of the distributed expressions for the proper nouns "Tanaka" and "Sato" can be relatively close. However, when these sentences appear in a novel, the difference in proper nouns such as a person's name or a place's name greatly affects the reading comprehension of the story. The two sentences above need to be clearly distinguished.

Also, proper names such as person names and place names appear in the corpus less frequently than ordinary nouns and verbs. Therefore, there is a high possibility that it will be out of vocabulary and will be treated as unknown words. The translation system is likely to output incorrect translations for such person names and place names. The possibility of mistranslation is even higher when translating unusual or fictitious names of people or places, or when the translation system itself is learned from a low-resource corpus.

3 Penalty by Edit Distance

In the evaluation of Japanese sentences in novel translation, it is necessary to give a low evaluation value if the reference sentence and the sentence to be evaluated have different corresponding named entity expressions. In particular, we deal with proper nouns in which the flow of the story collapses due to incorrect output of person names and place names among proper expressions.

In this study, the edit distance (Gusfield, 1997) between the tokens that forms the proper nouns in two sentences is calculated, and that value is reflected as a penalty in the final evaluation. The edit distance is a distance indicating how similar two strings are. The editing process that inserts, deletes, and replaces one character in one of the two character strings is repeated until it matches the other character string. Then, the minimum number of processes required to match two character strings is defined as the edit distance.

Since proper nouns appear in the corpus less frequently than general words, in the following steps, low-frequency tokens are assumed and treated as tokens that make up proper nouns. BERTScore calculates idf values to pay attention to the characteristic expressions of sentences. In this study, all tokens whose idf value exceeds the threshold are treated as LFT (Low requent Token). Then, referring to Maximum Similarity, we obtain the paired token of the sentence on the other side that maximizes the cosine similarity for all the low-frequency tokens included in the sentence. For the low-frequency token set thus obtained and the token set of the pair corresponding to the low-frequency token set, the edit distance is calculated after converting them into Japanese pronunciation character, hiragana. As a result, it is possible to give a high editing distance to low frequency words with completely different meanings, and conversely, a low editing distance can be given to the orthographic variants in the same low frequency token, i.e. kanji, katakana, and hiragana in Japanese.

In this study, the edit distance obtained in this way is divided by the longer of the reading lengths of low-frequency tokens, normalized to the range [0,1], and the value is subtracted from 1. Then the edit distance matchs coefficient M expressed between [0,1]. If the concordance coefficient is close to 0, it indicates disagreement, and if it is close to 1, it indicates coincidence. The agreement coefficient M obtained from the edit distance between two sentences is shown in the following Equation.

$$M = 1 - \frac{EditDist(x^{(LFT)}, \hat{x}^{(LFT)})}{max(|x^{(LFT)}|, |\hat{x}^{(LFT)}|)}$$

Here, x(LFT) represents the entire reading string of the low-frequency token in the reference sentence. x^(LFT) represents the entire string of reading of infrequent tokens in the sentence to be evaluated. EditDist(s1,s2) represents the edit distance between sentences s1 and s2.

Then, the matching coefficient is set for all tokens in the evaluation target sentence. A value obtained by dividing the sum of these by the length of the sentence is reflected as a penalty in the precision and recall in BERTScore. However, the concordance coefficient is set to 1 for tokens for which the

idf value does not exceed the threshold and is not a low-frequency token. The respective penalties are shown in the following Equations.

$$Pnl^{(P)} = \frac{\sum_{\hat{x}_i \in \hat{x}} p(\hat{x}_i)}{|\hat{x}|}$$

$$Pnl^{(R)} = \frac{\sum_{x_i \in x} p(x_i)}{|x|}$$

$$p(x) = \begin{cases} M & (idf(x) > threshold) \\ 1 & (otherwise) \end{cases}$$

By these penalties, BERTScore is modified as follows.

$$P_{rev} = P_{BERT} \times Pnl^{(P)}$$

$$R_{rev} = R_{BERT} \times Pnl^{(R)}$$

$$F_{rev} = 2\frac{P_{rev} \cdot R_{rev}}{P_{rev} + R_{rev}}$$

This makes it possible to give an appropriate penalty to the similarity evaluation of two sentences that are mostly similar but differ only in proper nouns.

4 Experiments

In this section, we explain the test set and procedure in the experiment to evaluate the performance of the proposed method, and explain various tools related to the experiment. In order to evaluate the performance of the proposed method, we first performed a preliminary experiment under multiple settings and confirmed the setting with the best performance. Then, we conducted a verification experiment comparing it with the existing method to examine whether the proposed method is effective. In order to distinguish the similarity between tokens based on the cosine similarity, the evaluation on how the two sentences are semantically similar, performed by the human or the automatic evaluation system, is called “similarity score”.

4.1 Data

The experimental data and the experimental procedure are based on the evaluation sharing task (Metric Shared Task) (Ma et al, 2018) in WMT. Verification experiments to be described later and comparison experiments with other existing methods were also performed under the same experimental settings.

For the experiment, we extracted 100 sentences, 10 sentences each from 10 novels randomly selected from Project Gutenberg Canada (http://gutenberg.ca/index.html). Then, we asked expert translator to translate the extracted English sentences into Japanese. The 100 translated Japanese sentences obtained were used as the reference sentences in the test set.

When translating, we requested that the translation be performed based on the fact that it was one sentence included in the novel, but on the other hand, we instructed that the information carried over across sentences would not be added. The purpose of this research is to properly evaluate the variety and difference of expressions in novels. It is not intended to evaluate other characteristics included in the novel (such as abbreviations of words and free translations of pronouns considering the preamble). For this reason, the ten sentences selected from one novel were chosen so that the connections of the stories were located far enough away from each other so that they could not be seen as much as possible. We shuffled 100 sentences and asked for translation. We translated these 100 English sentences by Google translation and by our novel translation system using the novel corpus developed by ourselves, i.e. we obtained two types of machine translated Japanese sentences for each English sentence. In the output set obtained, some of the expressions such as personal names, unknown words, and other apparently broken sentences were modified.

The reason for making modifications to translations is that it is difficult for current machine translation systems to translate novel sentences in high quality, so almost all sentences may be classified as low quality by humans. However, in order to properly measure the performance of the proposed method, a high score is given to the evaluation target sentence judged by the human evaluator to be good, and a low score is given to the evaluation target sentence judged to be bad. Even if the proposed method gives a high correlation to a dataset in which all sentences are evaluated low by humans, it cannot be judged that they have been evaluated correctly. Therefore, we intentionally mix high-quality translations to avoid the judgment that all sentences are of low quality. Here, the definition of a good evaluation

target sentence is a sentence whose meaning is similar to that of the reference sentence (regardless of difference in expression), and the definition of a bad evaluation target sentence is a sentence whose meaning is different from reference sentence, e.g., a sentence whose subject of action is different.
The reason why two types of sentences are prepared for one reference sentence by two translation systems is to confirm the correlation of evaluations between humans and systems for a large number of variations of expressions, and to check how the evaluations change by correcting the person's name etc. in one sentence.
Since one reference sentence and two evaluation sentences were set for one English sentence, a total of 200 reference sentence and evaluation sentence pairs were obtained. For these pairs, we performed a human evaluation task to evaluate the validity of the content of the reference sentence in the sentence to be evaluated. Using three Japanese native speakers as evaluators, we instructed to score how much the “contents/meanings” of the given pairs match, using a value between 0 and 100.
In scoring, we did not consider the fluency of the sentence to be evaluated, and requested to evaluate in the same way as a fluent sentence if the meaning was unbroken. In addition, even if the reference sentence and the sentence to be evaluated have similar meanings, we asked for a low score if the subject and/or object of the action were different. Since the reference sentence and the sentence to be evaluated are sentences in the novel, different proper nouns indicate different situations.
Normalization was performed to eliminate bias in the scoring of the three evaluators. We calculated the Pearson's correlation coefficient r, which examines the linear correlation between the two variables, and confirmed how well the scores of the evaluators agree. The correlation coefficient r was obtained from following Equation. Here, x represents the average value of the scores evaluated by x.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Since the Pearson's correlation coefficient is an index for evaluating the correlation between two variables, we calculated the correlation for every two evaluators. As a result of the calculation, the average value of the correlation was 0.657 to 0.548, and it was confirmed that the evaluations of the three parties were in good agreement. Therefore, we averaged the evaluation scores of each of the three evaluators, and defined them as the correct answer score by human evaluation in the test set.
In the following preliminary and verification experiments, the human correct answer scores obtained by the above procedure are used as references.
For the system scores obtained by the BERTScore and the existing method, we confirmed the agreement between the human evaluation and the evaluation by using Pearson's correlation coefficient.

4.2 Settings

In the BERTScore, which is the basis of the proposed method, the BERT pre-learning model used, the difference in the document set for which the idf value is calculated, and the presence or absence of a penalty affect the score. Therefore, it is necessary to confirm which of these settings has the best score.
In this study, we prepared following four pre-learning models for comparison when generating a token vector using BERT.

- ✓ Multilingual Model
  https://github.com/google-research/bert/
- ✓ Wikipedia Model
  http://nlp.ist.i.kyoto-u.ac.jp/
- ✓ SNS Model
  https://github.com/hottolink/hottoSNS-bert
- ✓ Novel Model
  Novel Model was developed by ourselves.

The BERT pre-learning model with Wikipedia as the learning corpus uses a very large scale of data for distributed representation. Although optimization is possible, almost all sentences in the learning corpus are composed of formal expressions, so they are not compatible with novels. Especially, since the first and second person are rarely included, there is a possibility that learning is not fully optimized for these words that appear frequently in novels.
On the other hand, the BERT pre-learning model, which uses a group of sentences posted on SNS as a learning corpus, has many colloquial sentences, and it is presumed that relatively many first-person and second-person sentences are included compared to Wikipedia sentences. However, due to the characteristics of media such as SNS, it contains

unnecessary information that cannot be seen in novels such as emoticons and URLs.
From such a background, texts which includes colloquial expressions rather than Wikipedia and is less than SNS, i.e. positioning model including intermediate expressions in colloquialism, were needed. Therefore, the novel was used as the learning corpus for BERT pre-learning model. We collected 6,876,198 novel sentences from the novel submission site "小説家になろう, (Become a Novelist)". Morphological analysis and subword conversion were applied the sentences. BERT was trained using this learning corpus.
The list that defines the correspondence between each token and idf value is called the idf dictionary. In BERTScore, which is the basis of the proposed method, each sentence included in the reference side that is the correct answer in the sentence pair set to be evaluated is regarded as one document, and the idf value for each token is calculated. However, with this method, if tokens that should be judged to be a unique expression such as a person's name frequently appear on the referrer side, the idf value may be lower than intended and it may not be regarded as a unique expression.
Therefore, in this study, we prepared a special document set separately from the test set, and modified it so that high idf values could be given to unusual words by calculating idf values using them. As a large-scale corpus for idf value calculation, we utilized a Wiki corpus consisting of all 882892 documents acquired from Japanese Wikipedia and 229 documents obtained from novel corpus, and examined how different idf dictionaries affect the score.
The reason why we prepared two kinds of corpora is that the Wiki Corpus has articles that describe various things and can cover a wider range of vocabulary. Novel corpus has many novel-specific expressions such as first person and second person that are rarely seen in Wikipedia and can give a low idf value to the word. Therefore, such words will be not mistakenly processed as a low frequency token.
Throughout the experiment, the threshold value of idf to be processed was set as follows.

1. Sort all token and idf value pairs obtained from large corpus by idf value.
2. Tokens are divided into 70% and 30% of the whole, and 70% tokens are not processed and 30% takens are processed
3. Set the threshold value at the boundary between two groups.

4.3 Preliminary Experiment

In the preliminary experiments, four types of pre-learning models, two types of idf dictionaries and the presence/absence of an edit distance penalty for low-frequency words are combined and examined, i.e. a total of 16 experimental environments. The effectiveness of the method was verified by comparing the performance of the experimental environment with the best performance with other existing methods. We set SentBLEU and Quick-Thought (Lajanugen and Honglak, 2018) as the methods to be compared.
We calculated the Pearson's correlation coefficient between the correct (reference) scores by the three evaluators and the predicted scores output by the system, and confirmed how well the human evaluation and the system evaluation agree. Following tables show the result when the idf dictionary was constructed based on the novel corpus, and the result when the idf dictionary was constructed based on the Wiki corpus.

Model	With Penalty	Without Penalty
Multilingual	0.169	0.197
Wikipedia	0.395	0.387
SNS	0.153	0.094
Novel	0.456	0.451

With idf dictionary by novel corpus

Model	With Penalty	Without Penalty
Multilingual	0.303	0.292
Wikipedia	0.405	0.397
SNS	0.094	0.083
Novel	0.438	0.451

With idf dictionary by wiki corpus

From the results of the preliminary experiment, when the Japanese novel sentence is evaluated using the modified BERTScore, using the BERT pre-learning model by the novel model, and the penalty with the idf dictionary constructed by the novel corpus has the highest correlation with humans.

5 Consideration

We conducted a comparative evaluation of the proposed method using the setting that achieved the highest performance in a preliminary experiment

and other existing methods. Following table shows the results of comparing the Pearson's correlation coefficient with the human evaluation for each of these methods and the proposed method.

Method	Pearson's correlation
Our method	0.456
Sent BLEU	0.093
Quick-Thought	0.067

The proposed method showed higher correlation with human evaluation than other comparison methods.

Next, we check how the presence or absence of a penalty affects the similarity evaluation. Using the novel model and the idf dictionary constructed from the novel, the similarity score with no penalty and the similarity score with penalty were compared with the similarity score by human evaluation. We investigate which sentence is more similar to human evaluation in which sentence.

Table below shows the comparison results. The first column of the table shows which is more similar to the human evaluation between presence and absence of penalty.

Of the two sentences in the second column, the upper sentence is the reference sentence and its English translation, and the lower sentence is the evaluation target sentence and its English translation.

In the third column, base indicates no penalty and penalized indicates penalty. The parentheses following the similarity in the table show the difference between the similarity by the human evaluation and the similarity by the system evaluation. The lower (the smaller the difference), the higher the correlation with the human evaluation.

In the first example of the table, the subject of the reference sentence is “Saito” and the subject of the sentence to be evaluated is “Ozaki”, indicating a completely different person. Therefore, as instructed when creating the data, the average human evaluation for these sentences is 0.477, which is a relatively low score. On the other hand, comparing the system-based scores, the Penalized score, which has a similar score lower due to the penalty, is closer to the manual evaluation than the original Base score. Therefore, the penalty works as intended.

Setting close to human evaluation	Sentence pair	Similarity Value
With penalty	斉藤さんは今最高の環境にいると思った。 (Mr. Saito thought he was in the best environment right now.) 尾崎は自分がこれまでにないほど良い場所にいると思った。 (Ozaki thought he was in a better place than ever before.)	Human: 0.477 (-) Base: 0.514 (0.037) Penalized: 0.492 (0.015)
Without Penalty	あなたは猟師のように槍を握っていなかったから失敗したのよ。 (You failed because you didn't hold your spear like a huntman.) あなたはハンターがすべきように槍を持っていなかった、そしてあなたが逃したのでそれはあなたのせいです。 (You didn't have a spear as a hunter should, and it's your fault because you missed.)	Human: 0.790 (-) Base: 0.617 (0.173) Penalized: 0.605 (0.185)

Contrary to the first example, in the second example of the table, the Base score is closer to the human evaluation than the Penalized score. A closer examination revealed that there was a penalty for the token “ハンター(hunter)” which corresponds to “猟師 (huntsman”). This is because the idf became high as the token “ハンター(hunter)” was not sufficiently included in the novel corpus, and it was considered a penalty target.

Considering this example, we suppose that there are other examples in which penalties are erroneously given because the idf calculated from the corpus is

high, even though it is a general token that is not a proper noun. In order to solve this problem, it is conceivable to expand the corpus with a wider range of novel documents, or to not use idf as the penalty granting criterion in the first place. In order to recognize proper nouns more accurately, it is possible to add labels to words in the reference sentence and the sentence to be evaluated by using named entity extraction technique.

Examining the base score that does not give a penalty, the correlation of human evaluation has hardly changed in the novel model and the Wiki model. However, when a penalty is added, the correlation is improved in many models when the idf dictionary based on the novel corpus is used. On the other hand, in the case of using the idf dictionary based on the Wiki corpus, there is not much improvement or rather a decrease compared to the case of using the novel corpus.

This is because the model based on the novel corpus has a low idf value for words often used in novels such as personal pronouns, while the Wiki corpus has a high idf value for these expressions, so that no penalty is imposed. Then, it results improperly penalizing the sentence pair.

6 Conclusion

In this study, we applied BERTScore, which is an automatic evaluation method using distributed expressions of words by BERT, to Japanese novels. When two sentences with different expressions but having similar meanings were compared, it was possible to make a more accurate evaluation than BLEU, which only considers the surface form of words. We investigated which BERT pre-learning model and which idf dictionary should be used when applying BERTScore to Japanese sentences. In addition, we improved the tendency in which high similarity is given to different named entity expressions, which is a problem when evaluating novels, by giving a penalty using the edit distance when calculating the score.

We evaluated the similarity between two sentences using the BERTScore with a penalty, and confirmed the correlation with human evaluation under multiple settings. As a result, it was confirmed that a higher correlation was shown by matching the type of the learning corpus and idf dictionary used for the BERT pre-learning model with the sentence to be evaluated. In addition, a improvement in correlation was seen by giving a penalty.

However, due to the characteristics of the novel corpus used for BERT pre-training and idf dictionary construction, some expressions, such as vector representations of real-world place names, are insufficiently optimized, and idf for general nouns is high. In the verification experiment, the corrected BERTScore showed higher correlation than SentBLEU, which considers only the surface form of words, and Quick-Thought, which obtained the cosine similarity from the distributed representation of sentences.

In the future, it is conceivable to create a large-scale and reliable Japanese corpus with a score so that the score can be predicted by regression analysis from the BERT distributed expression of two sentences. In BERT, the distributed expression of the obtained sentence was used as an explanatory variable to meet various downstream tasks of natural language processing. It may be possible that the automatic evaluation method can also be handled by setting a task that directly predicts the similarity score from the distributed expression of two sentences.

Acknowledgement

This research is (partially) supported by "Knowledge Hub Aichi", Priority Research Project from Aichi Prefectural Government.

References

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.

Bahdanau Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR 2015.

Dan Gusfield. 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY, USA.

Logeswaran Lajanugen and Lee Honglak. 2018. An efficient framework for learning sentence representations. In International Conference on Learning Representations.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attentionbased neural machine translation. In Proceedings of the 2015

Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 671–688, Belgium, Brussels. 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 the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014.Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 3104–3112.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav. Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Workshop on Multiword Expressions in Asian languages

Predicative multi-word expressions in Persian

Jens Fleischhauer
Department of General Linguistics
Heinrich-Heine-Universität, Düsseldorf, Germany
fleischhauer@phil.uni-duesseldorf.de

Abstract

Persian, like many other Asian languages, licenses the use of bare nouns in object position. Such sequences are often treated as multi-word expressions (compound verbs/light verb constructions, and pseudo-incorporation constructions). In the paper, I argue against a uniform treatment of all 'bare noun + verb' sequences in contemporary Persian. The paper presents criteria which allow to distinguish light verb constructions from other superficially similarly looking predicational construction types.

1 Introduction: Predicative construction types

Persian, an Iranian language of Western Asia, has a rather small set of lexically full verbs. Mohammad & Karimi (1992, 195) mention a number of around 115 lexically full verbs; others, for example Samvelian & Faghiri (2016, 212) and Samvelian (2018, 256), mention somewhat higher numbers (about 250) but state that only around half of them are still in use. The lack of full verbs is compensated by the use of light verb constructions (sometimes also called 'compound verbs'). Light verb constructions (LVCs) consist of a semantically reduced verb and a non-verbal element (NVE). The NVE is prototypically an NP (cf. the LVC *sedâ dâdan* 'to produce a sound' in (1)).[1]

(1) *Sag sedâ dâd.*
dog sound give.PST
'The dog made a sound.'

Light verb constructions are multi-word expressions since they form a predicational unit consisting of (at least) two words. LVCs are fixed expressions since the set of light verbs is rather restricted (according to Family (2006, 8), around 20 Persian full verbs exhibit light and heavy uses). Furthermore, the combination of a light verb with a specific type of NVE is not fully predictable. Nevertheless, LVCs are semantically compositional as they show properties very similar to those found with idiomatically combining expressions (in the sense of Nunberg et al. 1994); LVCs license internal modification and form families (Family, 2011; Samvelian, 2012; Samvelian & Faghiri, 2014, 2016; Fleischhauer et al., 2019; Fleischhauer & Gamerschlag, 2019; Fleischhauer & Neisani, 2020).

Light verbs are formally identical to the lexically full verbs as the examples in (2) show. The light use of *xordan* 'eat' is illustrated in (2a), there it heads the LVC *qose xordan* 'to worry about' (lit. 'concern eat'). The example in (2b) illustrates the heavy use of *xordan* which denotes an event of eating food.

(2) a. *hâmiše qosey-e fârda-ra*
always concern-EZ tomorrow-ACC
mi-xor-âd.
IMPF-eat-3SG
'She always worries about the future.'
(Family, 2006, 85)

[1] Glossing of the examples follows the Leipzig glossing rules; the following abbreviations are used: ABS: absolutive, ACC: accusative, CL: classifier, DEM: demonstrative, EMPH: emphatic, ERG: ergative, EZ: ezâfe, IMPF: imperfective aspect, INDEF: indefinite, NEG: negation, PST: past tense, PL: plural, SG: singular, SUB: subjunctive.

b. *Bâčče-ha qazâ* râ *xord-and.*
child-PL food ACC eat.PST-3PL
'The children ate the food.'

There is some debate whether all sequences of bare noun plus verb should be conceived as instances of the same type of complex predicate or not. Some authors (e.g., Ghomeshi & Massam 1994; Vahedi-Langrudi 1996; Mahmoodi-Bakhtiari 2018, 295) treat the 'N + V' sequences in (3) equally as 'compound verbs.' Others (e.g., Mohammad & Karimi, 1992; Lazard, 1992; Nemati, 2010; Megerdoomian, 2012; Modaressi, 2014) argue that the two sequences in (3) look superficially similar but exemplify different types of constructions. The example in (3a) is analyzed as an instance of pseudo-incorporation (e.g. Nemati 2010; Modaressi 2014, 2015); the one in (3b) is an LVC.

(3) a. *Bâčče-ha qazâ xord-and.*
child-PL food eat.PST-3PL
'The children ate (food).'
b. *Doxtar dârad jiq mi-zan-ad.*
girl has scream IMPF-hit-3SG
'The girl is screaming.'

Thus, Persian poses the problem of distinguishing between at least three predicate construction types: (i) regular predicate-argument constructions (2b), (ii) light verb constructions (1), (2a) and (3b), and (iii) pseudo-incorporation constructions (3a). I define a predicational construction type as a specific morphosyntactic construction which realizes the sentence predicate. Most importantly, the set of light verbs is partially overlapping with the set of heavy verbs which show pseudo-incorporation of bare nouns in object position. An example is the verb *xordan* 'to eat', which has already been illustrated in the examples in (2a) and (3a) above.

Irrespective of the question whether one analyzes examples like in (3a) as multi-word expression or not (see, e.g., Hüning & Schlücker 2015), one needs criteria to distinguish between LVCs on the one hand and pseudo-incorporation constructions (PICs) on the hand. I will argue that such a set of criteria will also allow to distinguish between LVCs and regular predicate-argument constructions. Although the current analysis focuses on Persian only, the question discussed in the paper is highly relevant for a larger number of Asian (but also non-Asian) languages (e.g. Turkish, Hindi/Urdu, Kurdish).

2 Pseudo-Incorporation

The term 'pseudo-incorporation' goes back to Massam's (2001) analysis of Niuean, which is a Malayo-Polynesian language. Niuean exhibits a grammatical phenomenon which is reminiscent of the nominal incorporation attested, for example, in many American languages (e.g., Mohawk or Tiwa, see, Mithun 1986). In Niuean, a nominal element is usually preceded by a case marker, as indicated in (4a). The absence of a case marker preceding *ika* 'fish' (4b) results in a change from a transitive case frame (ergative case for the subject, absolutive case for the object) to an intransitive one (absolutive case for the subject).

(4) a. *Takafaga tūmau nī e ia e*
hunt always EMPH ERG he ABS
tau ika.
PL fish
'He is always fishing.'
b. *Takafaga ika tūmau nī a ia.*
hunt fish always EMPH ABS he
'He is always fishing.'
(Massam, 2001, 157)

A crucial aspect of the 'V + N' sequence *takafaga ika* 'hunt fish' in (4b) is that the two do not form a morphological word (for details, the reader is referred to the original discussion in Massam 2001). Subsequent work has shown that pseudo-incorporation is widespread among the world's languages. The literature on (pseudo)-incorporation has identified a number of stable semantic properties which are cross-linguistically attested in (pseudo)-incorporation this issue, see Borik & Gehrke 2015). Pseudo-incorporated nouns tend to be non-referential and show the following properties: (i) they have obligatory narrow scope with respect to scope bearing elements, e.g., negation; (ii) they are number neutral; (iii) they are discourse opaque; and (iv) they show restrictions with respect to modifiability. One has to mention that there is some cross-linguistic variance with respect to these properties; especially the property of discourse opacity is somewhat relaxed in some lan-

guages (see Farkas & de Swart 2003 on Hungarian). I illustrate these properties by the use of Persian language data.

Starting with the first property, the bare noun *gorbeh* 'cat' in (5a) has narrow scope with respect to the negation operator. The only interpretation of the sentence is that the subject referent did not see any cat. The non-bare, i.e., case marked, noun in (5b) has wide scope with respect to the negation operator. The sentence means that there is a particular cat which the subject referent did not see.

(5) a. *Gorbeh na-did-âm.*
cat NEG-see.PST-1SG
'I didn't see any cat.' [$\neg > \exists$]
b. *Gorbeh-râ na-did-âm.*
cat-ACC NEG-see.PST-1SG
'I didn't see the cat.' [$\exists > \neg$]

Number neutrality is illustrated by the example in (6a). The noun *gorbeh* 'cat' is used without a plural marker but licenses a singular as well as plural interpretation. With respect to non-bare nouns, number interpretation depends on number marking. If the noun neither bears plural marking nor is preceded by a number word, it only licenses a singular reading (6b).

(6) a. *Gorbeh did-âm.*
cat see.PST-1SG
'I saw (a) cat/cats.'
b. *Gorbeh-râ did-âm. #Xeili ziba*
cat-ACC see.PST-1SG very pretty
bood-ând.
be.PST-3PL
'I saw the cat. #They were very pretty.'

Bare nouns are non-referential and therefore do not introduce discourse referents. The bare noun *šer* 'poem' in (7a) cannot serve as the antecedent of a (null) anaphora.[2] In non-bare use, the noun introduces a discourse referent and cannot be picked up anaphorically (7b).

(7) a. *Ali bâyad šer be-xân-ad. #Ân*
Ali must poem SUB-read-3SG DEM
(*šer*) *tavasote yek šâer-e arab*
POEM by INDEF poet-EZ Arabic
sorude šod.
written become
'Ali must read a poem. That poem was written by an Arabic poet.'
b. *Ali bâyad šer-i bexânad.*
Ali must poem-INDEF SUB-read-3SG
Ân (*šer*) *tavasote yek šâer-e*
DEM poem by INDEF poet-EZ
arab sorude šod.
Arabic written become
'Ali must read a [specific] poem. That poem was written by an Arabic poet.'

Finally, pseudo-incorporated nouns are restricted with respect to modification. Attributively used adjectives require a linking element — called 'ezâfe' -– which is an affix placed between the modified noun and its modifier. As (8a) shows, the bare noun does not license the adjective *ziba* 'beautiful' as an attributive modifier. Modification is, as shown in (8b), restricted to kind-level modifiers (a similar restriction is mentioned by Espinal & McNally 2011 for Spanish and Catalan).

(8) a. **Ketab-e ziba nevešt-âm.*
book-EZ beautiful write.PST-1SG
'I write (a) beautiful book/books.'
b. *Mân ketab-e ghesseh mi-xâr-âm.*
I book-EZ story IMPF-buy-1SG
'I buy story books.'
(Modaressi, 2014, 23)

The properties associated with pseudo-incorporation are only found with bare nouns in object position. Other types of bare nouns, especially those figuring as the subject argument, do not show these properties.

3 Semantic differences between LVCs and pseudo-incorporation constructions

The current section aims at demonstrating that the semantic function of the verb is different in LVCs and PICs. Light verbs are semantically reduced compared to their heavy verb use. They do not have

[2] Modaressi (2014, 2015) as well as Krifka & Modarresi (2016) show that Persian bare nouns are not fully discourse opaque but are discourse translucent, following the terminology of Farkas & de Swart (2003). For a discussion of this issue, the reader is referred to the mentioned literature.

full predicational content and therefore do not denote an event of their own (e.g., Butt & Geuder, 2001, 356). Rather, the denoted eventuality is mainly determined by the NVE. This becomes clear from Fillmore et al.'s (2003) discussion of the differences between the English verb *decide* and the LVC *make a decision*. They write that both sentences in (9) "report on the same event, that of deciding something" (Fillmore et al., 2003, 244).[3] Although the LVC is headed by the light verb *make*, the authors state that sentence (9b) is "not about an event of making." This is tantamount to saying that the LVC denotes a different situation-type -– or event-type -– than the one denoted by the heavy correspondent of its verbal head.

(9) a. *The committee decided to convene again next month.*
b. *The committee made a decision to convene again next month.*

Building on the above-mentioned idea, I propose the working definition of a light verb construction presented in (10).

(10) A light verb construction is a complex predicate consisting of a semantically light verb and a non-verbal element. The situation type denoted by the light verb construction is not a subtype of the situation type denoted by the heavy verb but is dependent on the NVE.

The basic idea is that the light verb construction denotes a different type of situation than the heavy verb. Since this is a crucial part of the argumentation, I like to illustrate the definition by use of the Persian examples in (11).

(11) a. *Ân mard be ân zan yek*
DEM man to DEM woman INDEF
ketâb dâd.
book give.PST
'The man gave a book to the woman.'
b. *Sag sedâ dâd.*
dog sound give.PST
'The dog made a sound.'

[3]Fillmore et al. (2003) do not use the term 'light verb' but speak of 'support verbs.'

In (11a), *dâdan* is used as a heavy verb and denotes a giving-situation, which requires a special relation between an agent (the giver), a theme (the given), and a recipient. In this type of situation, the referent of the theme is transferred from the giver to the recipient. The LVC *sedâ dâdan* 'produce a sound', in (11b), denotes a situation of sound emission, which involves an emitter and an emittee (the emitted sound). A sound emission-situation is not a specific subtype of a giving-situation. That the two constructions denote different situation types is evidenced by the fact that only the example in (11a) allows adding 'and she is still in possession of it.' Thus, the working definition captures the basic idea that the main predicational content of an LVC is contributed by the non-light element, whereas the light verb merely adds information to the event predication (e.g., Butt & Geuder, 2001, 2003).

The definition in (10) allows us to distinguish between LVCs on the one hand and PICs on the other. As mentioned above, in the case of an LVC like *sedâ dâdan* 'produce a sound', the denoted situation type is not determined by the verb but by the NVE. In the case of a PIC like in (12), the verb determines the denoted situation. *Gorbeh didan* 'see a cat(s)' is a specific subtype of a seeing situation, i.e., it is a seeing of cats rather than of some other stimulus. The pseudo-incorporated noun further specifies the situation type denoted by the verb.

(12) *Gorbeh did-âm.*
cat see.PST-1SG
'I saw (a) cat/cats.'

This brief discussion gives rise to a first distinguishing property of light verb constructions and pseudo-incorporation constructions:

(13) LVCs differ from PICs with respect to the lexical element(s) determining the denoted situation type.

The contrast with respect to the element(s) determining the denoted situation type is a direct consequence of a difference regarding the status of the verb in the two types of complex predicates. The verbal head of an LVC is a light verb, whereas it is a heavy verb in the case of a PIC. Evidence for this fact is gained from the interpretation of the examples

discussed above. *Sedâ dâdan*, as already discussed in some detail, does not mean 'to give a sound to someone'. Thus, *dâdan* does not contribute its full lexical content. In the case of the PIC *gorbeh didan* 'cat see', the verb contributes its full lexical content. Only if a verb is used as a heavy verb and contributes its full lexical content is it able to determine the denoted situation type. I summarize this as a second distinguishing feature between the two types of constructions:

(14) The verbal head of an LVC is a light verb; a PIC is headed by a heavy verb.

After having present semantic differences between LVCs and PICs, I turn next to the discussion of the role bare nouns play in the two predicational construction types.

4 Bare nouns as NVEs

The semantic properties of pseudo-incorporated nouns are usually only found with nouns showing "some degree of bareness" (Borik & Gehrke, 2015, 12). In some languages, pseudo-incorporation is restricted to nouns without any functional morphology (e.g., number, case, in/definiteness marking), while other languages show weaker restrictions. Hungarian (Farkas & de Swart, 2003) and Greek (Gehrke & Lekakou, 2013) license accusative case marking on pseudo-incorporated nouns, whereas Hindi allows plural marking (Dayal, 2011). The discussion in Section 2 revealed that Persian restricts pseudo-incorporation to bare nouns.

The current section aims at investigating two different albeit related questions: First, do bare nouns used as NVEs of light verb constructions show the same semantic properties than pseudo-incorporated nouns? Second, is the nominal element within an NVE necessarily bare or does is license functional morphology?

4.1 The interpretation of bare noun NVEs

In section 2, it was shown that pseudo-incorporated nouns show a number of recurrent properties: they have narrow scope with respect to scope bearing elements, they are number neutral, they are discourse translucent, and, finally, they only license kind-level modifiers. The crucial question to be answered in the current section is whether bare nouns used as the NVE of a light verb construction show the same properties.

For the purpose of illustration, I will use the LVC *sedâ dâdan* 'produce a sound'. As the example in (15) shows, the bare noun *sedâ* does not introduce a discourse referent. Furthermore, the noun is interpreted as number neutral; the LVC either refers to situations of emitting a single sound or of emitting a number of (different or non-different) sounds.

(15) *#Âbgarmkon sedâ dâd. Ân (sedâ)*
boiler sound gave DEM sound
boland bud.
loud be.PST
Intended: 'The boiler produced (a) sound(s). It was loud.'
(Fleischhauer & Neisani, 2020, 13)

The bare noun NVE also has narrow scope with respect to negation. It is understood that the boiler did not produce any sound rather than that there is a particular sound which it did not produce.

(16) *Âbgarmkon sedâ na-dâd.*
boiler sound NEG-give.PST
'The boiler did not produce any sound.'

In contrast to pseudo-incorporated nouns, bare noun NVEs show fewer restrictions with respect to modification. As (17) shows, the bare noun *sedâ* licenses modification by the adjective *boland* 'loud' which is not a kind-level modifier.

(17) *Sedâ-ye boland dâdan nešân az*
sound-EZ loud give sign from
qodrat nist.
strength NEG.be.3SG
'Producing a loud sound/loud sounds is not a sign of strength.'
(Fleischhauer & Neisani, 2020, 13)

Bare noun NVEs indeed share a number of properties with pseudo-incorporated nouns. One might take this as evidence that there is no real distinction between LVCs on the one hand and PICs on the other. Contrary to this assumption, LVCs and PICs show a number of differences, for example, regarding restrictions on nominal morphology.

4.2 Morphosyntactic properties of pseudo-incorporated nouns and NVEs

With respect to light verb constructions, the question is whether LVC-formation is – like pseudo-incorporation – similarly restricted to nominal elements showing some degree of bareness. Persian has nominal morphology for the expression of indefiniteness, number as well as case. The three categories are briefly discussed subsequently.

Indefiniteness marking

Persian has different grammatical means for expressing indefiniteness: the indefinite article *yek* – which is identical to the numeral 'one' – and the phrasal suffix *-i*. The two markers have an overlapping but not identical distribution and can also be used in combination (Ghomeshi 2003, 65, Paul 2008, 322, Fleischhauer & Neisani 2020, 11). Within the limits of the current paper, I cannot present a detailed discussion of the similarities and differences of the indefiniteness markers. For the current discussion, it seems sufficient to say that *-i* signals specificity, whereas *yek* does not. Only *-i* but not *yek* can be used in referentially opaque contexts, like in (18a). In the example in (18b), either *yek* or *-i* but also both together can be used.

(18) a. *Ahmad mi-xâst bâ* (**yek*)
Ahmad IMPF-want.PST with INDEF
zan-e puldâr-i ezdevâj
woman-EZ rich-INDEF marry
kon-ad ammâ na-tavânest
do-3SG but NEG-could
kas-i-râ peidâ kon-ad.
one-INDEF-ACC find do-3SG
'Ahmad wanted to marry a rich woman but could not find one.'

b. *Ahmad mi-xâst* bâ (*yek*)
Ahmad IMPF-want.PST with INDEF
zan-e puldâr(*-i*) *ezdevâj*
woman-EZ rich-INDEF marry
kon-ad ammâ u tark-aš kard.
do-3SG but she leave-3SG did
'Ahmad wanted to marry a rich woman but she left him.'
(slightly adapted from Fleischhauer & Neisani 2020, 11f.)

Nouns marked for indefiniteness – either by *yek* or *-i* – do not show the properties of pseudo-incorporated nouns. Rather, despite the fact that *-i* can be used in referentially opaque contexts, indefinite nouns are discourse transparent (18b). Additionally, nouns marked for indefiniteness receive a number specific interpretation, i.e., they are not number neutral. This is evidenced in (19): *yek* as well as *-i* enforce a singular interpretation of the noun *medad* 'pencil'. Thus, a specification on the number of pencils cannot be added to the sentences in (19).

(19) a. *Yek medad avord-âm*,
INDEF pencil bring.PST-1SG
#yek-i bâraye khod-âm va
one-INDEF for self-1SG and
do-ta bâraye Leila.
two-CL for Leila
'I brought a pencil (one for me and two for Leila).'

b. *Medad-i avord-âm*,
pencil-INDEF bring.PST-1SG
#yek-i bâraye khod-âm va
one-INDEF for self-1SG and
do-ta bâraye Leila.
two-CL for Leila
'I brought a [specific] pencil (one for me and two for Leila).'
(based on Modaressi 2014, 24)

The nominal element within an NVE can be marked for indefiniteness, as the example in (20) shows, although the eventive noun *sedâ* 'sound' receives a referentially specific interpretation and introduces a discourse referent. This is evidenced by the fact that the referent introduced by *sedâ* can be anaphorically picked up. Thus, although *sedâ* is used referentially, it still forms a complex predicate with the verb *dâdan*. Irrespective of whether *sedâ* is marked for indefiniteness or not, the combination of *sedâ* with *dâdan* is interpreted as 'produce (a) sound(s)' rather than 'give someone a sound'. Thus, *dâdan* is still used as a light verb in (20) rather than as a heavy verb. This demonstrates that there is no relevant difference between the indefiniteness marking of NVEs and that of 'regular' nouns in argument position.

Nominal morphology	Pseudo-incorporation	Light verb construction
case	no	yes
indefiniteness	no	yes
number	no	yes

Table 1: Nominal morphology in Persian complex predicates.

(20) *Âbgarmkon sedâ-i dâd. Ân*
boiler sound-INDEF gave DEM
(*sedâ*) *boland bud.*
sound loud be.PST
'The boiler produced a [specific] sound. That (sound) was loud.'
(Fleischhauer & Neisani, 2020, 13)

Number marking

Persian has a binary number system distinguishing between an unmarked singular and a morphologically expressed plural. The plural marker *-hâ* is optional in contexts in which number is already expressed by other means, e.g., number words. The example in (21) shows plural marking of the NVE *sedâ*; the interpretation of the example is that the subject referent produces a number of (different) sounds.

(21) *In mâšin šab-hâ sedâ-i-hâ*
DEM car night-PL sound-INDEF-PL
mi-dah-ad.
IMPF-give-3SG
'This car produces some [specific] sounds at night.'
(Fleischhauer & Neisani, 2020, 12)

Case marking

Persian has a binary case system: it possesses a morphologically unmarked nominative case and the phrasal case affix *-râ* which marks accusative case. The language displays definiteness-based differential object marking, restricting accusative case marking to nouns that have a referentially specific interpretation (see, e.g., Bossong 1985; Lazard 1992, and Ghomeshi 1997). Since NVEs take the specificity marker *-i*, it is not surprising that they also license accusative case marking. An example taken from Karimi-Doostan (2011, 89) is shown in (22).[4]

[4]For more data on the case marking of the non-verbal element of Persian LVCs, see, e.g., Samvelian & Faghiri (2014, 51) and Karimi-Doostan (1997/2012, 203ff.).

The LVC under discussion is *râhnamâʔi kardan* 'advice/give advice'; the NVE *râhnamâʔi* 'advice' bears accusative case marking and is separated from the light verb by the indirect object *be Sasan* 'to Sasan'.

(22) *Ali in râhnamâʔi-râ be Sasan kard.*
Ali DEM advice-ACC to Sasan do.PST
'Ali gave Sasan this advice.'

Interim summary

The nominal elements used within the two types of complex predicates have different morphosyntactic properties: pseudo-incorporated nouns have a higher degree of bareness than NVEs. Whereas pseudo-incorporated nouns have to be bare, NVEs do not carry restrictions with respect to number, indefiniteness, or case marking. This does not mean that any NVE licenses all types of functional morphology; accusative case marking, for example, is restricted to NVEs which are realized as the light verb's direct object. So far there has been no systematic investigation of which NVEs are realized as a direct object.

The morphological properties of the two predicative construction types are summarized in table 1. Pseudo-incorporation but not LVC-formation is restricted to non-referential nouns, i.e., nouns which are neither marked for case nor for indefiniteness or number. Thus, nominal morphology allows us to distinguish LVC-formation from pseudo-incorporation but does not provide clear-cut criteria for identifying NVEs. Bare noun NVEs superficially look like pseudo-incorporated verb complements, while non-bare noun NVEs superficially look like non-pseudo-incorporated verb complements.

5 Conclusion

The starting point of the current paper was the question whether all instances of 'bare noun + verb'

Predicational construction type	**Noun**	**Verb**
pseudo-incorporation construction	semantics: non-referential object argument morphosyntax: bare direct object noun	heavy verb determines situation type
regular (transitive) predicate-argument construction	semantics: referential object argument morphosyntax: non-bare direct object noun	heavy verb determines situation type
light verb construction	semantics: *no restrictions on referentiality* morphosyntax: *no restriction* determines situation type	light verb

Table 2: Summary of the semantic and morphosyntactic properties of Persian predicational construction types.

exemplify the same type of complex predicate or not. The current paper argues against this view and presents evidence for a distinction between LVCs and PICs. LVCs can be distinguished from PICs by a number of properties. First, the verbal head of a light verb construction is semantically light, i.e., it does not contribute its full lexical content. The verbal head of a PIC, on the other hand, is a heavy verb.

Second, the denoted situation type of an LVC is determined by the NVE but not by the light verb. This point is closely related to the first one mentioned above since light verbs are defective event predicates. In the case of pseudo-incorporation, the denoted situation type is a subtype of the situation type denoted by the verbal head, i.e., food-eating is a subtype of eating but sound emission (literally 'give a sound') is not a subtype of giving-situations.

Third, nominal morphology does not block LVC-formation. Rather, NVEs are basically compatible with all types of nominal morphology. Most crucially, NVEs license case and number as well as indefiniteness marking, which is compatible with the fact that LVC-formation is not restricted to non-referential NVEs. Case as well as specificity marking block pseudo-incorporation.

The morphosyntactic as well as semantic differences between the three basic predicational construction types discussed in the current paper are summarized in table 2.

Among the questions which need to be addressed in future work is the following: Are there also syntactic differences between LVCs and PICs? Such differences are expected given that pseudo-incorporation seems to be restricted to bare nouns in immediately preverbal position. On the basis of the criteria present in the current paper, a corpus-based study on the syntactic behavior of LVCs and PICs is planned.

The identification of semantic, morphosyntactic and syntactic properties of different types of MWEs will hopefully a better identification of these expression in language corpora.

Acknowledgments

The research was carried out as part of the research project 'Funktionsverbgefüge: Familien & Komposition' ('Light verb constructions: Families & composition'; HE 8721/1-1) funded by the Deutsche Forschungsgemeinschaft (DFG). I like to thank Mozhgan Neisani for help with the language data.

References

Borik, Olga & Berit Gehrke. 2015. An Introduction to the Syntax and Semantics of Pseudo-Incorporation. In Olga Borik & Berit Gehrke (eds.), *The Syntax and Semantics of Pseudo-Incorporation*, 1–43. Leiden/Boston: Brill.

Bossong, Georg. 1985. *Empirische Universalienforschung: Differentielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Gunter Narr.

Butt, Miriam & Wilhelm Geuder. 2001. On the (Semi)Lexical Status of Light Verbs. In Norbert Corver & Henk van Riemsdijk (eds.), *Semilexical Categories: On the content of function words and*

the function of content words, 323–370. Berlin: Mouton.

Butt, Miriam & Wilhelm Geuder. 2003. Light verbs in Urdu and grammaticalization. In Regine Eckardt, Klaus von Heusinger & Christoph Schwarze (eds.), *Words in Time*, 295–350. Berlin/New York: Mouton de Gruyter.

Dayal, Veneeta. 2011. Hindi pseudo-incorporation. *Natural Language & Linguistic Theory* 29(1). 123–167.

Espinal, M. Theresa & Louise McNally. 2011. Bare nominals and incorporating verbs in Spanish and Catalan. *Journal of Linguistics* 47. 87–128.

Family, Neiloufar. 2006. *Explorations of Semantic Space: The Case of Light Verb Constructions in Persian*. Paris: Ecole des Hautes Etudes en Science Sociales dissertation.

Family, Neiloufar. 2011. Verbal islands in Persian. *Folia Linguistica* 45(1). 1–30.

Farkas, D. & H. de Swart. 2003. *The Semantics of Incorporation: From Argument Structure to Discourse Transparency*. Stanford: CSLI Publications.

Fillmore, Charles, Christopher Johnson & Miriam Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3). 235–250.

Fleischhauer, Jens & Thomas Gamerschlag. 2019. Deriving the meaning of light verb constructions – a frame account of german *stehen* 'stand'. In Constanze Juchem-Grundmann, Michael Pleyer & Monika Pleyer (eds.), *Yearbook of the German Cognitive Linguistics Association, Vol. 7*, 137–156. Berlin/Boston: Mouton de Gruyter.

Fleischhauer, Jens, Thomas Gamerschlag, Laura Kallmeyer & Simon Petitjean. 2019. Towards a compositional analysis of German light verb constructions (LVCs) combining Lexicalized Tree Adjoining Grammar (LTAG) with frame semantics. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 79–90. Gothenburg, Sweden: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-0407.

Fleischhauer, Jens & Mozhgan Neisani. 2020. Adverbial and attributive modification of Persian separable light verb constructions. *Journal of Linguistics* 56. 45–85.

Gehrke, Berit & Marika Lekakou. 2013. How to miss your preposition. *Studies in Greek Linguistics* 33. 92–106.

Ghomeshi, Jila. 1997. Topics in Persian VPs. *Lingua* 102(2–3). 133–167.

Ghomeshi, Jila. 2003. Plural marking, indefiniteness, and the noun phrase. *Studia Linguistica* 57(2). 47–74.

Ghomeshi, Jila & Diane Massam. 1994. Lexical/syntactic relations without projections. *Linguistic Analysis* 24 (3–4). 175–217.

Hüning, Matthias & Barbara Schlücker. 2015. Multi-word expressions. In Peter Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word formation: An International Handbook of the Languages of Europe, Vol. 1*, 450–467. Berlin/Boston: De Gruyter Mouton.

Karimi-Doostan, Gholamhossein. 1997/2012. *Light Verb Constructions in Persian*. Saarbrücken: Lambert Academic Publishing.

Karimi-Doostan, Gholamhossein. 2011. Separability of Light Verb Constructions in Persian. *Studia Linguistica* 65(1). 70–95.

Krifka, Manfred & Fereshteh Modarresi. 2016. Number neutrality and anaphoric update of pseudo-incorporated nominals in Persian (and weak definites in English). In Mary Moroney, Carol-Rose Little, Jacob Collard & Dan Burgdorf (eds.), *Proceedings of Semantics and Linguistic Theory (SALT 26)*, 874–891. Washington DC: Linguistic Society of America.

Lazard, Gilbert. 1992. *A grammar of contemporary Persian*. Costa Mesa, CA/New York: Mazda Publishers.

Mahmoodi-Bakhtiari, B. 2018. Morphology. In A. Sedighi & P. Shabani-Jadidi (eds.), *The Oxford Handbook of Persian Linguistics*, 273–299. Oxford: Oxford University Press.

Massam, Diane. 2001. Pseudo Noun Incorporation in Niuean. *Natural Language & Linguistic Theory* 19. 153–197.

Megerdoomian, Karime. 2012. The status of the nominal in Persian complex predicates. *Natural Language & Linguistic Theory* 30. 179–216.

Mithun, Marianne. 1986. On the nature of noun incorporation. *Language* 62(1). 32–37.

Modaressi, F. 2014. *Bare nouns in Persian: Interpretation, Grammar and Prosody*. Ottawa/Berlin: University of Ottawa & Humboldt-Universität zu Berlin dissertation.

Modaressi, F. 2015. Discourse Properties of Bare Noun Objects. In Olga Borik & Berit Gehrke (eds.), *The Syntax and Semantics of Pseudo-Incorporation*, 189–221. Leiden/Boston: Brill.

Mohammad, Jan & Simin Karimi. 1992. Light verbs are taking over: Complex predicates in Persian. In J. A. Nevis & V. Samiian (eds.), *Proceedings of the Western Conference on Linguistics, vol. 5*, 195–212. Fresno: California State University.

Nemati, Fatemeh. 2010. Incorporation and complex predication in Persian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG2010 Conference*, 395–415. Stanford: CSLI Publications.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Paul, Daniel. 2008. The individuating function of the Persian 'indefinite suffix'. In Simin Karimi, Vida Samiian & Donald Stilo (eds.), *Aspects of Iranian linguistics*, 309–328. Cambridge: Cambridge Scholars Press.

Samvelian, Pollet. 2012. *Grammaire des prédicats complexes – les constructions nom-verbe*. Paris: Lavoisier.

Samvelian, Pollet. 2018. Specific features of Persian syntax. In Anousha Sedighi & Pouneh Shabani-Jadidi (eds.), *The Oxford Handbook of Persian Linguistics*, 226–269. Oxford: Oxford University Press.

Samvelian, Pollet & Pegah Faghiri. 2014. Persian Complex Predicates: How Compositional Are They? *Semantics-Syntax Interface* 1(1). 43–74.

Samvelian, Pollet & Pegah Faghiri. 2016. Rethinking compositionality in Persian Complex Predicates. In *Proceedings of the Berkeley Linguistic Society 39th Annual Meeting, February 16-17 2013*, 212–227. Berkeley: Berkeley Linguistics Society.

Vahedi-Langrudi, Mohammad-Mehdi. 1996. *The syntax, semantics and argument structure of complex predicates in modern Farsi*. Ottawa: University of Ottawa dissertation.

Forms and Meanings of Lexical Reduplications in Cantonese: a corpus study

Charles Lam
Department of English
The Hang Seng University of Hong Kong
Shatin, N.T., Hong Kong
charleslam@hsu.edu.hk

Abstract

Lexical reduplications (LR) in Cantonese are multiword expressions (MWEs) that are frozen and unproductive. The meanings of LR are often non-compositional, such as *hung4 bok1 bok1* 'bright red', where *bok1 bok1* does not contribute any meaning to *hung4* 'red'. This can pose a challenge in parsing and natural language understanding. LRs can be misinterpreted to bear literal meaning, if they are mistakenly treated as decomposable chunks. Identifying LRs is therefore an important step to reduce errors in word segmentation and natural language understanding. This study discusses a collection of LRs extracted from two data sets. Some common patterns are also identified in this study, which may guide parsers to automatically identify items that are novel to the system.

1 Introduction

Reduplication in general is a ubiquitous phenomenon in Cantonese. There are many types of reduplication and they are mostly productive. This means that reduplications are templatic, and their compatibility with the lexical items is rule-governed. For example, non-stative verbs (i.e., verbs involving actions) may be reduplicated to show durative events, as in (1a). Reduplication of the classifier (or measure words) denotes the meaning of 'every', as shown in (1b). To express the short duration of an event, one may use the 'V+one+V' reduplication in (1c). Since these reduplication types are rule-governed, their meaning and structure are predictable and cause no problem in NLP.

(1) a. haang6 haang6 haa5
walk walk PRT
'while walking'

b. zek3 zek3 gau2
CL CL dog
'every dog'

c. mong6 jat1 mong6
look one look
'to take a look'

d. coeng4 coeng2 dei2
long long DEI
'long-ish; fairly long'

Unlike the productive reduplications described above, lexical reduplications (LR) in Cantonese are multiword expressions (MWEs) that cannot be formed productively. LRs are considered frozen and are not coined by individual users. Many LRs are idiomatic in that their meanings are non-transparent. In (2a), the single use of *dai2* can only mean 'under / beneath', but not undergarment. Example (2b) does not have a corresponding base form (i.e. non-reduplicated) *tiu3 zaat3* either. When the item is used, *tiu3* must be reduplicated.

(2) a. dai2 dai2
under under
'undies' (casual term for undergarment)

b. tiu3 tiu3 zaat3
jump jump tie
'bouncy and active'
(Attested example from words.hk)

What makes LR tricky is that they may carry literal meaning in some cases. This can cause parsers to an-

alyze them as compositional and miss the intended idiomatic use. Example (3) contains the parallel of *haang4* 'walk' and *kei2* 'stand':

(3) haang4 haang4 kei5 kei5
walk walk stand stand
'being idle and aimless'
(From Cheung, Ngai and Poon (2018))

Taken literally, the phrase could mean walking and pausing intermittently. However, it actually means being idle and not doing anything and is often used to describe laziness. Its meaning is therefore unpredictable and must be learned by heart for speakers.

Some LRs contain parts that do not contribute meanings at all, most notably in some color terms. The reduplicated syllables *bok1* in (4b) and *gam4* in (5a) do not appear to bear any meaning[1], and there is no non-reduplicated 'base form', i.e., **hung4 bok1* alone is an illicit form.

(4) a. hung4 'red'
b. hung4 bok1 bok1
red BOK BOK
'bright red'

(5) a. wong4 gam4 gam4
yellow GAM GAM
'bright yellow'
b. *hung4 gam4 gam4
red GAM GAM
Intended: 'bright red'
c. *wong4 bok1 bok1
yellow BOK BOK
Intended: 'bright yellow'

In addition, examples (5b) and (5c) show that the combinations are fixed and cannot be changed. The same also applies to several other colors or adjectives. As the examples above have demonstrated, LRs are unproductive and frozen forms. The data presented in this study will potentially be useful for error / grammar detection (Jiang et al., 2012), or facilitate other studies on idiomatic expressions (Wang et al., 2019).

[1]Throughout this paper, these meaningless elements, reduplicated or not, are glossed with the romanization in all capital letters.

2 Related Works

Studies on lexical reduplication in Cantonese are limited. In the linguistics literature, most studies lie in the areas of phonology and phonetics, which deals with the relation between the underlying form of the reduplication and its realied pronunciation.

Since the seminal study by Wilbur (1973), there has been great progress in the investigation of sound systems (Botha, 2006; Frampton, 2009; Inkelas and Zoll, 2005). However, sound patterns might not be directly useful for LRs in the present study. For the morphosyntax and semantics of reduplication, previous studies focused on productive and predictable forms (Hurch, 2005; Francis et al., 2011; Štekauer et al., 2012), which cannot cover the LRs in the present study either.

Specific to Sinitic languages, Cheng (2012) and Lee (2020) provide thorough explanations on the mechanism of classifier reduplication in example (1b). Lam (2013) and Basciano and Melloni (2017) both discussed verbal reduplication in Cantonese and Mandarin. It seems that unproductive and frozen forms like LR have not received much attention in linguistics. This is probably because they do not display particular patterns, and therefore are not seen as theoretically important.

For idioms in Sinitic languages, the focus is often placed in learning enhancement, either with digital materials (Chung and Hsieh, 2017), or extracting a list of idioms from corpus data (Wang et al., 2013). The present study falls under the latter type. However, given the paucity of Cantonese teaching materials for children (which typically contain more idioms than materials for adults), this study uses dictionaries, both online and print, as the sources of data. As our data show, Cantonese LRs typically would not be considered a part of Chinese idioms or four-character expressions. It is therefore necessary to investigate LRs as a separate category from idioms.

Corpus resources for Cantonese are scarce, both for LRs specifically and for idiomatic expressions at large. While there are several Cantonese corpora that are widely used (Luke and Wong, 2015; Lee and Wong, 1998; Leung and Law, 2001), no previous works have been conducted on LR and their distribution. The present study is an attempt to in-

vestigate LRs through a comparison across corpora. As Cantonese is known to have a rich inventory of idiomatic expressions, many of the LRs identified in this study are not found in Mandarin. It is therefore not practical to assume Mandarin resources can be borrowed to handle Cantonese data. Therefore, this study aims to provide a constructed data set specific to LRs with the description of the data. The next section provides the statistics of the data sets, and the section after discusses their significance and potential uses.

3 Data sets

3.1 Cheung, Ngai and Poon (2018)

This study extracted two sources for the data set of LR. The first source of data was the Cantonese dictionary (Cheung et al., 2018), henceforth CNP. The CNP data set was extracted and digitized manually from the print dictionary. Out of the total of 12,000 entries in the dictionary, 756 entries contain reduplicated elements, which makes approximately 6.30% of the data set. Table 1 summarizes the data set.

Length of LR	Types	% of LR
2 characters	23	3.04%
3 characters	268	35.45%
4 characters	302	39.95%
5 characters	27	3.57%
6 characters	23	3.04%
7 characters	33	4.37%
8 characters	9	1.19%
$\geq$ 9 characters	71	9.39%
Total	756	100.00%

Table 1: Summary of LR in the CNP data set

Since the source is a dictionary, each entry represents a unique type and there is no number of tokens available.

3.2 Words.hk

The second source is the site `https://words.hk`, which is a crowd-sourced effort to create an online dictionary established in 2014. This study uses its data set of Cantonese articles, which includes lexical items that are attested in written articles but excludes the items that are only found in the constructed dictionary section of the site. This approach provides the number of tokens used in the texts, which can better reflect the use of LR, rather than the constructed list of types in the dictionary format. In the `words.hk` data (henceforth WHK), there are 30,821 unique types and 2,938,248 tokens from this data set. The longest lexical items have 4 characters. More details are listed in table 2:

Length	Types	Tokens
1 char	3,878 (12.59%)	2,253,458 (76.69%)
2 char	20,592 (66.88%)	636,566 (21.66%)
3 char	3,116 (10.12%)	32,408 (1.10%)
4 char	3,205 (10.41%)	15,419 (0.52%)

Table 2: Summary of all types in the Words.hk data set

Among these unique types, 881 are found to contain reduplicated elements. Since reduplicated forms entails more than one character, table 3 excludes one-character types. The percentages in table 3 are based on the types / tokens of LR.

Length	Types	Tokens
2 characters	138 (15.66%)	9,870 (67.92%)
3 characters	242 (27.47%)	2,302 (15.84%)
4 characters	501 (56.87%)	2,359 (16.23%)
Total	881 (100%)	14,351(100%)

Table 3: Summary of LR in the Words.hk data set

4 Distribution and Patterns of LRs

This section describes the patterns of the attested LRs in the two data sources. The reduplicated elements can form several patterns that are logically possible, but they are not equally distributed. While these sources are not meant to be exhaustive, the proportion of the different LR categories are similar, indicating that they are representative of LRs in the Cantonese language.

LRs with two characters cannot display any variation in pattern, due to their lengths. More fine-grained analysis on their parts of speech, meanings or syntactic distribution will require further investigation. The patterns of 3- and 4-character LRs will be discussed below.

4.1 3-character LR

LRs with three characters are attested in both sources and can be found in three templates.

(6) AAB-template:
- a. laap6 laap6 ling3
  LAAP LAAP shiny
  'shiny'
- b. cyun3 cyun3 gung3
  sacarstic sacarstic GUNG
  'sacarstic'

Similar to the examples of *hung4 bok1 bok1* 'bright red' in (4b) and *wong4 gam4 gam4* 'bright yellow' in (5a), some reduplicated elements are meaningless, as in (6). However, it is not always the case that the unreduplicated element denotes the meaning of the whole LR. Example (6b) shows that in some cases, it is the reduplicated element that indicates the meaning of the whole phrase, and the unreduplicated element is meaningless.

In the ABA-template, while examples like (7) and (7b) contain meaningful elements, the meaning of the entire phrase is not always compositional:

(7) ABA-template:
- a. gau2 m4 gau2
  long.time not long.time
  'once in a while'
- b. daap3 soeng6 daap3
  contact over contact
  'to liaise through a third party'

The ABB template includes many adjectives. As shown in the first section, color terms like examples (4b), (5a) and (8) often appear in the ABB template. In example (8b), *jin6* is likely a truncation of *jin6gam1* 'cash'. While the term is somewhat transparent, the whole phrase is used as an adverb that modifies paying events, not as a noun (as 'cash' is). Therefore, the function of the full phrase is not completely predictable by its elements.

(8) ABB-template:
- a. baak6 syut1 syut1
  white snow snow
  'very white'
- b. jin6 dau1 dau1
  now DAU DAU
  'in cash'

Table 4 below shows the identified LRs and their distribution in the two data sets:

	Types	
Category	**Words.hk**	**CNP Dictionary**
AAB	88 (36.36%)	89 (31.79%)
ABA	31 (12.81%)	11 (3.93%)
ABB	122 (50.41%)	180 (64.29%)
AAA	1 (0.41%)	0 (0%)
Total	242 (100.00%)	280 (100.00%)

Table 4: Subcategories of 3-character LRs by types

More than half of the unique types belong to the ABB pattern across the two sources; and the second largest group is the AAB pattern, followed by the alternating ABA pattern. The distribution is the same across the two data sources, suggesting that it reflects the general pattern in the language as well.

On a side note, the AAA pattern is extremely rare in Cantonese. There are more than one attested entries from the WHK data, but a few of them were removed from LR, as they do not reflect idiomatic use of the language. For example, the entries of '999' (in arabic number) and *gau2 gau2 gau2* 'nine nine nine' (in Chinese character) are the emergency number in Hong Kong, so they do not reflect on the word formation in Cantonese. The only AAA item that ends up in the data set is *paak1 paak1 paak1*, which is onomatopoeic for sexual activities. The two data sets contain duplicated entries, such as:

(9) a. zing6 zing6 gai1
  quiet chicken chicken
  'very quiet'
- b. suk6 hau2 suk6 min6
  familiar mouth familiar face
  'very familiar'

These duplicates were not removed, so that the two data sources are accurately represented. Researchers who want to make use of these data should remove the duplicated items. With the numbers indicating the unique types, table 4 provides a direct comparison between the two sources[2].

[2]Since the CNP dictionary data do not include naturalistic use in prose, there is no statistics on tokens. The numbers

4.2 Patterns of 4-character LR

LRs with four characters can be found in the following six templates.

(10) loi4 loi4 heoi3 heoi3
come come go go
'always' AABB-template

(11) sei2 sei2 dei6 hei3
die die ground air
'reluctantly' AAXY-template

(12) sai1zong1 gwat1 gwat1
suits bone bone
'being dressed up' XYBB-template

(13) bei2ci2 bei2ci2
each.other each.other
'same to you' ABAB-template

(14) mou5 jan4 mou5 mat6
no person no thing
'having nothing at all' AXAY-template

(15) sau2 ting4 hau2 ting4
hand stop mouth stop
'living from hand to mouth' XBYB-template

Similar to LRs with three characters, these items are idiomatic and cannot be modified. Some examples may seem to be transparent. However, the word order cannot be changed, even for the parallel ones. For example, (10) cannot be rephrased as **heoi3 heoi3 loi4 loi4*, and (14) cannot be **mou5 mat6 mou5 jan4*. This shows that these items should be considered frozen multiword expressions.

Table 5 shows the distribution of the types of LRs with four characters in the two data sets[3]. In both data sets, the AXAY template appears most frequently, followed by the AABB form. The two data sets diverge in the less frequent categories. For the two data sets, it is not clear why such difference exists, or whether such difference is representative.

	Types	
Category	**Words.hk**	**CNP Dictionary**
AABB	83 (16.57%)	61(20.07%)
AAXY	68 (13.57%)	7 (2.30%)
XYBB	39 (7.78%)	22 (7.24%)
ABAB	8 (1.60%)	5 (1.64%)
AXAY	267 (53.29%)	181 (61.18%)
XBYB	36 (7.19%)	23 (7.57%)
Total	501 (100.00%)	304 (100%)

Table 5: Subcategories of 4-character LRs by types

It is worth pointing out that the ABAB form results in rather low frequency, when compared to what one might expect from Mandarin. In Mandarin, the ABAB form is a productive morphological process to show tentative events. For instance, *chángshì* 'to try' can be reduplicated as *chángshì chángshì* 'to give it a try'. The process cannot be applied to Cantonese: While the cognate *soeng4 si3* 'to try' exists as a verb, the form **soeng4 si3 soeng4 si3* is not acceptable at all and therefore cannot be used to mean 'to give it a try'. This difference between Mandarin and Cantonese may explain the low frequency of ABAB reduplications. In all the attested Cantonese examples, the ABAB reduplications are not formed productively and therefore genuine multiword expressions.

5 Potential use of the data

It is possible that these forms of LRs can be used as part of novel term detection, or as a flag for idioms. From the examples above, it is clear that LRs are idiomatic and do not denote compositional meanings. Productive reduplications in Cantonese often come with specific morphemes, such as *dei2* for adjectival reduplication denoting diminution in example (1d), or *haa5* in V-one-V reduplication in example (1c). Reduplications that do not come with any particular marker can be a sign for LRs. At this point, only color terms in the ABB template form an observable pattern. Further studies on the form-meaning correlation in LRs will facilitate better recognition and comprehension for these idiomatic expressions. It is possible to identify and predict the meanings of based on the individual components. For example, LRs in ABB format with color terms and adjectives

of tokens in the Words.hk data set are as follows: AAB: 618 (26.85%); ABA: 778 (33.80%); ABB: 884 (38.40%); AAA: 22 (0.96%). In total, there are 2,302 tokens of LR with 3 characters.

[3]The tokens of 4-character LRs in the Words.hk data set are as follows: AABB: 472 (20.01%); AAXY: 274 (11.62%); XYBB: 149 (6.32%); ABAB: 16 (0.68%); AXAY: 1,265 (53.62%); XBYB: 183 (7.76%). In total, there are 2,359 tokens of LR with 4 characters.

as the first character are highly likely to be intensification of the denoted color or attribute.

6 Conclusion

This paper has highlighted the need for resources on the lexicial reduplication phenomenon (LR) and shown that the LRs are idiomatic and unproductive, which can be an issue for parsing or natural language understanding. Especially because many of them are unique to Cantonese, but not Mandarin, existing resources from Mandarin cannot be borrowed to recognize LRs, despite the abundance of Mandarin data.

The data show that LRs constitute a significant amount in the vocabulary. The reduplicated element may occur in various places within a Cantonese LR. Such unpredictability of LRs makes the phenomenon a challenge for natural language understanding. While some reduplication templates appear more frequently, it remains unclear how exactly the templates or forms correlate with the meanings.

Given that the two data sets are not exhaustive in listing the LRs in the Cantonese language, there will be novel LRs in NLP tasks in the real world. Therefore, the distribution from the presented data can potentially be used for identification of novel, undiscovered LRs.

References

Bianca Basciano and Chiara Melloni. 2017. Event delimitation in Mandarin: The case of diminishing reduplication. *Italian Journal of Linguistics / Rivista di linguistica*, 29(1):147–170.

Rudolf P. Botha. 2006. *Form and meaning in word formation: A study of Afrikaans reduplication*. Cambridge University Press.

Lisa Lai-Shen Cheng. 2012. Counting and classifiers. In Diane Massam, editor, *Count and mass across languages*, pages 199–219.

Lai Yin Cheung, Lit Wai Ngai, and Lai Mei Poon. 2018. *The Dictionary of Hong Kong Cantonese*. Hong Kong: Cosmo Books.

Liang-Yi Chung and Sheng-Min Hsieh. 2017. Using graphic digital materials in language learning. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 295–298. IEEE.

John Frampton. 2009. *Distributed reduplication*, volume 52. MIT Press.

Elaine J Francis, Stephen Matthews, Reace Wing Yan Wong, and Stella Wing Man Kwan. 2011. Effects of weight and syntactic priming on the production of Cantonese verb-doubling. *Journal of psycholinguistic research*, 40(1):1–28.

Bernhard Hurch. 2005. *Studies on reduplication*. Walter de Gruyter.

Sharon Inkelas and Cheryl Zoll. 2005. *Reduplication: Doubling in morphology*. Cambridge University Press.

Ying Jiang, Tong Wang, Tao Lin, Fangjie Wang, Wenting Cheng, Xiaofei Liu, Chenghui Wang, and Weijian Zhang. 2012. A rule based Chinese spelling and grammar detection system utility. In *2012 International Conference on System Science and Engineering (ICSSE)*, pages 437–440. IEEE.

Charles Lam. 2013. Reduplication across categories in Cantonese. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, pages 277–286.

Thomas Lee and Colleen Wong. 1998. CANCORP: The Hong Kong Cantonese child language corpus. *Cahiers de Linguistique Asie Orientale*, 27(2):211–228.

Peppina Po-Lun Lee. 2020. On the semantics of classifier reduplication in Cantonese. *Journa of Linguistics*, (online first).

Man-Tak Leung and Sam-Po Law. 2001. HKCAC: the Hong Kong Cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325.

Kang Kwong Luke and May Lai-Yin Wong. 2015. The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics*, 25:309–330.

Pavol Štekauer, Salvador Valera, and Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: a typological survey*. Cambridge University Press.

Zhimin Wang, Li He, and Yanqiu Shao. 2013. The idiom investigation of Chinese undergraduate textbook and the extraction of common used idioms. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 208–212. IEEE.

Chengyu Wang, Yan Fan, Xiaofeng He, Hongyuan Zha, and Aoying Zhou. 2019. Idiomaticity prediction of Chinese noun compounds and its applications. *IEEE Access*, 7:142866–142878.

Ronnie Wilbur. 1973. *The phonology of reduplication*. Indiana University Linguistics Club Bloomington.

Abstract Meaning Representation for MWE: A study of the mapping of aspectuality based on Mandarin light verb *jiayi*

Lu Lu
Hong Kong Polytechnic University
lu-cbs.lu@polyu.edu.hk

Nianwen Xue
Brandeis University
xuen@brandeis.edu

Chu-Ren Huang
Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Abstract

Multiword expression (MWE) refers to various types of linguistic units that are made up of more than one word. Light verb constructions (LVCs), as one of the least explored areas in MWEs, have idiosyncratic features that are difficult to capture in computational linguistics. In this paper, we addressed the aspectual differences between LVCs and their corresponding regular verb constructions with corpus data. Specifically, the *jiayi*-LVC in Mandarin Chinese was investigated as a case study, where idiosyncratic aspectual information in the LVC was proposed. This feature was not yet previously represented in abstract meaning representation (AMR), in which LVCs and its regular verb counterparts shared the same AMR. Given the semantic difference in *jiayi*-LVCs, we expand AMR by introducing aspect as the root node, while maintaining the core predicative component in meaning representation.

1 Introduction

Light verb constructions (LVCs) are a crucial type of multiword expression (MWE) that can be found across different languages (Butt 2010). One of the prominent semantic features in LVCs (e.g. *have a bath* in English and *jinxing yanjiu 'carry out a study'* in Mandarin Chinese) is that the light verb (LV; i.e. *have* and *jinxing*) bears little semantic content and the actions are largely described by the eventive nominals (i.e. *bath* and *yanjiu 'research'*), encoding the same semantic information as their regular verb counterparts. Since the meaning of LVCs differ from usual predicative structures or the direct aggregation of its semantic components, LVCs, as one of the least explored areas of MWEs in computational linguistics, pose a number of challenges in computational grammar, such as automatic word alignment, annotation and semantic representation. In the model of Abstract Meaning Representation (AMR), it is often assumed that the LVCs and its corresponding regular verb construction (RVC) share the same representation (Banarescu et al. 2012; Flanigan et al. 2014, Bu et al. 2016). AMR is a semantic framework addressing the predicate-argument relation of the whole sentence (to be more fully described in Section 4).

However, corpus data suggest that LVCs and RVCs have slightly different semantic meaning. In Urdu, for example, LVs play a central role in the meaning and morphosyntactic choices of the whole construction in (1). Although the LVs *par* 'fall' and *ḍaal* 'put' both occur with *ciik*^{h} 'scream' in Urdu, the LV in (1a), which involves an involuntary action, is preceded by an unmarked nominative subject, whereas the LV *ḍaal* 'put' in (1b), denoting a conscious control over the action, requires the marked ergative case on the subject

argument. Based on this observation, Butt (1995) argues that it is the particular LV *ḍaal* 'put' that contributes to the meaning of conscious choices of a given action, which further assigns a marked ergative case to the subject in (3b). In this analysis, the so-called 'LV' is not completely 'bleached', and has important role to play in AMR.

(1) a. vo ciik^{h} paṛ-aa
pron=NOM scream fall-PRF.M.SG
'He began screaming suddenly (despite himself).'

b. us=ne ciik^{h} ḍaal-aa
pron=ERG scream put-PRF.M.SG
'He screamed violently (on purpose).'
(Butt 1995: 110)

Xu et al. (2020) focussed on another set of near synonyms like properties of LVCs in Mandarin Chinese. They first show that a subset of LVs is interchangeable in some contexts, which support the hypothesis of their lack of contribution to the meaning of the sentences. However, it is also shown that in other contexts, only a certain LVs can be used. For example, the LV *jiayi* tends to take accomplishment events in Mainland Mandarin Chinese, whereas other LVs do not show the same inclination. This observation underlines the fact that even though LVCs share largely similar grammatical patterns with their corresponding RVCs or other LVCs, each of them still contains some lexical meaning that can interact with the context to generate new meaning. Xu et al. (2020) used these distributional grammatical features to classify and predict different LVs. Our current study tackles the challenge of how to systematically represent such meaning differences, especially within the Universal Meaning Representation (UMR) initiative (Xue et al. 2019). Given the variations in meaning that an LV contributes to, a simple AMR graph may not be adequate. In this regard, a deeper understanding of the semantic content in LVCs contributes to our understanding of the meaning representation in natural language processing, revealing parameters of variations and commonalities between different structures and modelling systems. We also expect that the results can shed light on how to represent MWEs in AMR/UMR.

This study will use the corpus data of *jiayi*, with the abstract literal meaning of *give* in Mandarin Chinese, as a case study. It investigates the semantic features that are idiosyncratic to *jiayi* with data from Chinese Gigaword and presents the preliminary thoughts on how AMR can be refined to represent the semantic differences between LVCs and RVCs.

2 Corpus observation

In this study, the most recent edition of the newswire corpus—Chinese Gigaword (fifth edition)[1]—was consulted. To limit the influence of the varieties of Chinese language on the research results, only the corpora of Mainland Mandarin Chinese were employed at this initial stage of research. That is to say, texts from the corpora of Guangming Daily, People's Daily, People's Liberation Army Daily and Xinhua News Agency were selected for this paper. Queries of the LV *jiayi* were entered into AntConc and 55 instances of qualified *jiayi* were returned.

A closer look at the concordances of *jiayi*-LVCs show its distinctive aspectual feature. In the Chinese Gigaword corpus, *jiayi*-LVCs cannot take any aspect markers, such as *-le* and *-guo*, immediately after the LV, see (2). However, other semantically similar LVs, such as *geiyu*, can go with an aspect marker, see (3). Such properties have also been observed in Hu and Fan (1995), Diao (2004) and Kuo (2011).

(2) You ren suiyi jiang
have people randomly with

zhexie lishi geming gequ
these history revolution song

cuangai jiayi xiedu.
falsify LV blaspheme
'Someone randomly falsified these revolutionary songs and blasphemed against them.'
(Chinese Gigaword)

(3) Yingguo Bominghan Daxue

[1] The corpus can be downloaded from the Linguistic Data Consortium catalogue.

UK Birmingham University

zhuanmen wei zhekuan che pingjia.
specifically for this car evaluation
'The University of Birmingham carried out an examination specifically for the car and spoke highly of it.' (Chinese Gigaword)

jinxing-le jiance, jiyu-le jigaode
LV-ASP examine LV-ASP excellent

Table 1. Tokens of different syntactic structures in the Chinese Gigaword corpus

	a. RV+OBJ1+OBJ2	b.OBL+RV+OBJ2	c. RV+OBJ2	d. LV+OBJ1+AN	e. OBL+LV+AN	f. LV+AN	Total
jiayi	0	0	0	0	9	46	55
geiyu	29	19	148	307	89	371	963

[Examples:

a. Ta **geiyu**-le [wo]$_{OBJ1}$ [rensheng zhong gengwei baogui-de jingyan]$_{OBJ2}$
he give-PRF me life in more precious experience
'He gave me more precious life experience.'

b. Guojia [dui huojiang qiye]$_{OBL}$ **geiyu** [300wan jiangli]$_{OBJ2}$.
nation to award-winning company give 3.million reward
'The country rewarded 3 million yuan to award-winning companies.'

c. Guojia ji yao xiang minzu diqu touru gengduode zijin,
nation not.only need to ethnic area invest more money

you yao **geiyu** [geng youhuide zhengce]$_{OBJ2}$
but.also need give more preferential policy
'The country not only needs to make more investment in ethnic areas, but also needs to give more preferential policies (to them).'

d. Ta zai shenghuoshang **geiyu** [ta]$_{OBJ1}$ [wuweibuzhide zhaogu]$_{AN}$.
he on life LV her meticulous care
'He cared for her meticulously in life.'

e. Ta [dui zhe yi zuofa]$_{OBL}$ **geiyu** [chongfen kending]$_{AN}$.
he to this one practice LV adequate confirmation
'He highly confirmed this practice.'

f. Ri mei ye **geiyu**-le [zugou guanzhu]$_{AN}$.
Japanese media also LV-PRF enough attention
'Japanese media also paid enough attention (to this).']

In what follows, we will explain and represent the aspectual properties of the *jiayi*-LVC observed in and generalised from corpus data: none of the aspect markers can be found in *jiayi*-LVCs.

3 Aspectuality in *jiayi*-LVCs

As noted earlier, an LV is assumed to be form identical to its corresponding regular verb (RV) in a language. In Table 1, we summarised the syntactic structures of *jiayi*-LVCs evidenced in the Chinese Gigaword corpus, and further, for a better understanding, contrasted it with another LV *geiyu* 'give' (bearing the same abstract meaning as *jiayi*), which can be used as an RV in a sentence (see the examples from a-c right after Table 1).

In Table 1, structures (a), (b) and (c), occurring in RVCs, can only be found in the contemporary *geiyu*, whereas the three structures were readily available in Classical Chinese for

both verbs. We can also see that a mixture of RVCs and LVCs is widely manifested in the Contemporary *geiyu*-LVCs, but such mixture is by no means occurs in *jiayi*-LVCs. As LVCs arguably enter the grammaticalisation cline developing from RVs to grammatical morphemes, it implies that *jiayi* is much closer to the grammatical end of the cline, compared to *geiyu*.

Additionally, from the perspective of semantics, the transition from (a) to (f) implies the generalisation in meaning, whereby new context entails more general meaning. This is, as observed in Heine and Kuteva (2007), one of the important factors responsible for grammaticalisation. However, as for *jiayi*, it has lost lexical content to a great extent that their contemporary syntactic context mostly favours the last two structures listed in Table 1. This supports the above claims that *jiayi* is at a later stage of grammaticalisation, compared to *geiyu*. In other words, *jiayi* is more grammaticalised than *geiyu*.

The percentage in Figure 1, based on the frequencies of RVCs and LVCs in Table 1 lends further weight to the claim that *geiyu* is at the earlier stage of grammaticalisation compared with *jiayi*.

Figure 1. The percentage concerning the distribution of LVCs and RVCs

	RVC	**LVC**
	Lexeme →	Grammatical morpheme
geiyu	20.4%	79.6%
jiayi	-	100%

The earlier/latter stages of grammaticalisation are, interestingly, compatible with the realisation/non-realisation of the perfective aspect markers. It is assumed that the perfective aspectual information is co-provided by the verb and the construction (Michaelis 2004 and Goldberg and Jackendoff 2004) in the *jiayi*-LVCs. Given the usual grammaticalisation path is 'independent lexical verb>grammatical morpheme (i.e. aspect marker in this case)', the more grammaticalised an LV is, the more enriched the aspectual meaning will be (grammaticalised from an RV), and less likely it will resort to the verb to provide aspect information. As for *jiayi*, the verb is bounded and telic in its lexical meaning. This is especially evident in its lexical use, where *jia*, the obsolete form of *jiayi* in Classical Chinese, was used. For example, in *jia bing wo* add military-force I 'add more military force to me', the event encoded by the verb *jia* has a specific end point, thus, when interacting with the context, implying the event can be viewed as a completed whole unit.

Given the above illustration, since *jiayi* is close to the end of grammaticalisation (in contrast to *geiyu*, as shown in Figure 1), we believe that the construction of *jiayi* is sufficient to embody the perfective aspect on its own right. Therefore, the fixed aspectual value internally conveyed in the *jiayi*-LVC makes it incompatible with any perfective aspect markers. This results in the non-realisation of the perfective aspect marker in the *geiyu*-LVCs.

4 Mapping aspectuality into AMR: A preliminary thought

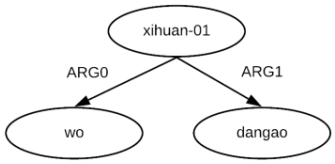
As generalised from the above linguistic observations, while the core predicative-argument relation remains the same in *jiayi*-LVCs and their corresponding RVCs, aspectual information is additionally encoded in LVCs. Therefore, other than the predicate relation and its arguments, which are the core semantic relation in the current AMR, we propose to refine the AMR modelling to represent aspectual information that set a particular LVC apart from its possible syntactic alternatives.

In AMR annotation, canonical meaning of a sentence is represented as a single-rooted, directed, acyclic graph with nodes labelled with concepts and edges labelled with relations. In AMR, the predicate argument structure is the core component. The predicate and its arguments are represented as nodes and the edges represent the relation between the predicate and each of its arguments in the AMR graph. As an illustration, the AMR notation and graph representation of sentence (4) can be found in (5).

(4) wo xihuan dangao.
I like cake
'I like cakes.'

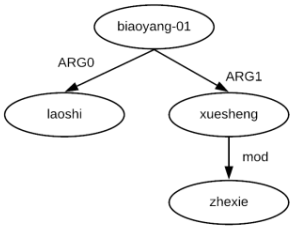
(5) a. (x0/xihuan-01
:ARG0 (x1/wo
:ARG1 (x2/dangao))

b.



As can be seen, the advantage of representing predicative core elements lies in the central positioning of the predicate argument structure. However, AMR does not represent aspectual information (Banarescu et al. 2012). While Bu et al. (2016) add aspect as a non-core semantic relation particularly designed for Chinese AMR, it does not specify how aspect is embedded into Chinese LVCs. As regards to LVCs, Bonial and Palmer (2016) further argue the approach that LVCs and its corresponding RVCs share the same AMR may be adequate for English LVCs, but it needs to be evaluated for other languages, as cross-linguistically there is some semantic space that cannot be covered by their corresponding RVCs. Given the corpus observation and aspectual justification in Sections 2 and 3, we argue that LVCs and its corresponding RVCs encode different aspectual information in Chinese, and argue that the current AMR does not yet properly handle aspectual encodings with different syntactic realisations: for example, the AMR-graph in (6) can represent both LVC and the corresponding RVC in (7).

(6)



(7) a. LVC

Laoshi	dui	zhexie	xuesheng
teacher	to	these	student

jiayi	biaoyang
LV	praise

'The teacher made a praise to these students.'

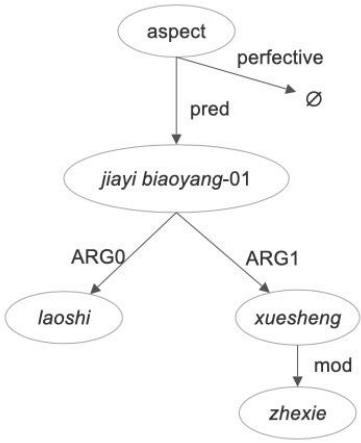
b. RVC counterpart

Laoshi	biaoyang-le	zhexie	xuesheng.
teacher	praise-ASP	these	student

'The teacher praised these students.'

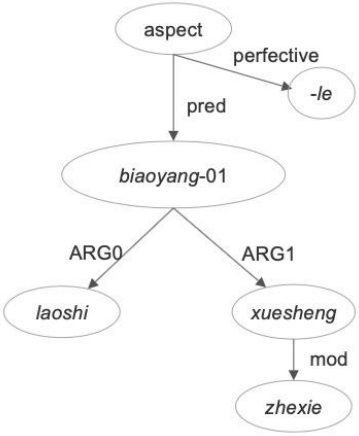
In our research, instead of treating aspect as a modifier, we propose to consider it as a root assigning aspectual value and acting over the predicative node. Consider the AMR graph as shown in (8). The aspect acts as the root node taking aspectual value over the predicate. In example (8), aspect is represented as a property over the predicate (i.e. *jiayi* and the nominal complement), demonstrating the aspectuality in *jiayi*-LVC (i.e. the predicate). In this case, the perfective aspectual value is left empty in AMR, as it is inherently contained in the *jiayi*-LVC. This parallels with the justification of the aspectual features in *jiayi*-LVCs in Section 3. Additionally, from the perspective of lexical semantics, the literal sense of *jiayi*, which is 'add', encodes the telic property on its own, and thus can be viewed as a simple whole. Therefore, the perfective aspect of the *jiayi*-LVC in (7a) is not realised in the node of aspect.

(8) AMR graph of example (7a)



Compared to its corresponding RVC (where the nominal complement in LVC is used as the main predicate), the aspect value, represented in the root node, is shared between the verbal predicate and the aspect marker in RVC. As represented in (9), the aspectual value pertaining to the RV *biaoyang* is specified and realised in the aspect node.

(9) AMR graph of example (7b)



We believe the advantages of the approach can be seen in three folds. It differentiates two syntactically and semantically similar structures, while maintaining both the predicative core elements of the sentence. Further, since the aspect node has the same representation in *jiayi*-LVCs and RVCs (except that the grammatical aspect is realised or not), it lowers the cognitive load for annotation and processing, especially for those who are not familiar with the grammatical system of the language. Lastly, this approach has the potential to generalise into other LVs, leading to a universal representation of aspects and its interaction with the predicate.

5 Conclusion

In this study, generalising from corpus observation, we argue that perfective aspectual meaning is internally encoded in the *jiayi*-LVCs in Mandarin Chinese, thus highlighting the semantic differences between LVCs and corresponding RVCs. Given this, we refined the AMR to capture the aspectual encoding and its interaction with the predicate and the aspect marker. We proposed a root node aspectual feature in Chinese AMR, while maintaining the predicative core element in the original AMR graph. This preliminary work, drawing on the two roughly equivalent constructions of LVCs and their RVC counterparts, enriches the representation of AMRs with the feature of aspect. In the next, we will expand the representation with more corpus data and experiment with small-scale annotation and testing to work on its feasibility regarding universal representation.

References

Banarescu, L. Bonial, C., Cai, S. Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K. Koehn, P., Palmer, M. and Schneider, N. 2013. 'Abstract meaning representation for sembanking'. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Bonial, C., and Palmer, M. (2016). 'Comprehensive and consistent PropBank light verb annotation'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3980–3985.

Bu, L., Y. Wen, B. Li, N. Xue. 2016, 'Zhongwen chouxiang yuyi biaoshi biaozhu guifan (V1.2)' [The annotation guidelines of Chinese Abstract Meaning Representation]. https://www.cs.brandeis.edu/~clp/camr/res/CAMR_GL_v1.2.pdf (accessed on July 10, 2020)

Butt, M. 1995. *The structure of complex predicates in Urdu*. Stanford: CSLI.

Butt, M. (2010). 'The light verb jungle: Still hacking away'. *Complex predicates: Cross-linguistic perpectives on event structure*. 48–78. Cambridge: Cambridge University Press.

Diao, Y. 2004. *Xiandai hanyu yuyi dongci yanjiu* [The research on weak verbs in contemporary Chinese]. Dalian: Liaoning Normal University Press.

Flanigan, J., Thomson, S., Carbonell, J. G., Dyer, C., and Smith, N. A. 2014. 'A discriminative graph-based parser for the abstract meaning representation'. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 1:*1426–1436).

Goldberg, A.E. and Jackendoff, R., 2004. 'The English resultative as a family of constructions'. *Language* 80 (3): 532–568.

Heine, B., and Kuteva, T. 2007. *The genesis of grammar: A reconstruction*. Oxford: Oxford University Press.

Hu, Y. and Fan, X. 1995. *Dongci yanjiu* [The research on verbs]. Kaifeng: Henan University Press.

Kuo, P. J. A. 2011. 'Case study of obligatory object fronting in Mandarin Chinese'. In *Online proceedings of glow in Asia workshop for young scholars.* Accessed http://faculty.human.mie-u.ac.jp/~glow_mie/Workshop_Proceedings/11Kuo.pdf (Accessed on 15 Aug, 2016)

Michaelis, L. A. 2004. 'Type shifting in construction grammar: An integrated approach to aspectual coercion'. *Cognitive Linguistics* 15 (1): 1–68.

Xu, H., Meng, J., Lin, J. and Huang, C.-R. 2020. 'Light verb variations and varieties of Mandarin Chinese: Comparable corpus driven approaches to grammatical variations'. *Corpus Linguistics and Linguistic Theories.* [ahead of print]

Xue, N., Croft, W., Hajic, J., Huang, C.-R., Oepen, S., Palmer, M., and Pustejovksy, J. 2019. Proceedings of the First International Workshop on Designing Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations.*

Formulaic language of Vietnamese children with autism spectrum disorders: A corpus linguistic analysis

Phạm Hiển
Institute of Linguistics
Vietnam Academy of Social Sciences
phamhieniol@gmail.com

Nguyễn Thị Giang
Institute of Linguistics
Vietnam Academy of Social Sciences
blueriver063@gmail.com

Abstract

This study examines formulaic language (based on multiword expressions) in the interactive speech of eleven children with autism spectrum disorders (ASDs). Play sessions were recorded to collect speech samples. Speech-language pathologists (SLP) acted as informants during the recording sessions. Qualitative and quantitative analyses were carried out: a qualitative analysis of situational factors that potentially impacted the prevalence of formulaic language, a quantitative analysis of the prevalence of formulaic language in speech samples using a classification system developed for the study. Various situational factors increased or decreased formulaic language use, though all eleven participants used formulas. Formulas corresponded to several categories and varied in conventionality, whether in form or function. Nonetheless, the qualitative analysis indicated that formulas had several functional uses in the interactions of participants. These findings have implications for future research and language assessment and intervention in ASD.

1 Introduction

The purpose of this study is to survey the predominance and the essence of the use of formulaic language in the interactions of eleven Vietnamese children with ASD (Nguyễn, 2015). It is situated within the view that language is a complex adaptive system in which language emerges from interactions over time and is formed and entrenched by situational factors encompassing each occurrence of usage. Formulaic sequences are ubiquitous in language use and they make up a large proportion of any discourse. Erman and Warren (2000) calculated that formulaic sequences of various types constituted 58.6% of the spoken English discourse they analyzed and 52.3% of the written discourse. Sinclair's (1991) view that language as a whole is organised according to two main structuring principles: an open choice principle and an idiom principle, with the latter involving the widespread use of formulaic stretches of words. Furthermore, this store of formulaic sequences is dynamic and is constantly changing to meet the needs of the speaker (Wray, 2002: 101).

We have decided to use the term formulaic sequence based on a definition by Wray (2002: 9): "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar."

This term covers a wide range of formulaic language, and touches on two key criteria of the emphasis in this book: a) we are concerned with sequences of lexis and b) the mind handles, or appears to handle, these sequences at some level of representation as wholes. However, using this definition, Wray argues that even single words and morphemes can be seen as formulaic sequences.

2 Functions of Formulaic Language

As previously described, formulaic language, as opposed to propositional language, is especially well suited to serving certain functions for speakers. Ellis (1996) contended that language acquisition is essentially sequence learning. In fact, children who were late to combine words were more at risk for future problems with language than children who were late with their first words (Rudolph & Leonard, 2016). Wray and Perkins (2000) group the functions of formulaic language into two categories: (a) devices of social interaction and (b) compensatory devices for memory limitations. The first category includes the use of formulas for the purpose of manipulating others, asserting separate identity, and asserting group identity (Wray & Perkins, 2000). Wray (1999: 216) proposes the following scenario: In a crowded and noisy bar, asking a stranger to move so that you can get past requires attracting their attention and interrupting their conversation. A formulaic expression such as *nói chung là* or *thượng lộ bình an* is easily recognized… In contrast, a less formulaic utterance, such as *Tôi đạp xe dắt theo một con chó*, must be heard more accurately because it is unpredictable, and requires more decoding, so it is more intrusive.

3 Functions of formulaic language in ASD.

To date, there appear to be no survey studies of the functions of formulaic language in ASD of Vietnamese children. However, there is reason to believe that formulaic language may serve different functions in persons with ASD as a result of their impairments in social communication (Wray & Perkins, 2000) and difficulties accomplishing full integral perception of all dimensions in communication. Together, these impairments may lead to a situation in which "formulaicity is not socio-interactionally motivated but rather is a 'Hobson's choice' [i.e., having no real alternative] solution to processing constraints" (Wray & Perkins, 2000: 23). In other words, formulaicity in ASD might not socio-interactionally motivated because of the impairment in social communication (Wray & Perkins, 2000) nor may there be a genuine decision between analytic and holistic processing because difficulties in integrated perception hinder segmentation and thus analytic processing (Noens & Van Berckelaer-Onnes, 2004). With respect to Wray and Perkins' (2000) division in functions, we would thus expect formulaic language to be used for cognitive purposes rather than as social interaction devices. However, impairment in pragmatic abilities in ASD is best described as a *deficiency* than a *complete inability* (Vogindroukas & Zikopoulou, 2011). Indeed, research pertaining to the categories of formulaic language, including immediate and delayed echolalia, politeness sequences (e.g., Volden & Sorenson, 2009), and discourse markers suggest that while the social functions of formulaic language may be impaired or less prevalent, they are not necessarily non-existent. These findings are discussed in greater detail below.

4 Methodology

4.1 Research questions

Q1: Do children on the verbal ASD spectrum with varying language abilities use formulaic language in interactions?

Q2: How are the form and function of formulaic expressions related in the interactions of children on the verbal ASD spectrum with varying language abilities?

4.2 Method

This study approaches the investigation of formulaic language in the conversational speech of children with ASD from a multiple case study design. In a multiple case study, "a number of cases are studied jointly in order to investigate a phenomenon or general condition" (Dörnyei, 2007, p. 152). The cases in this study are children with ASD while the phenomenon of interest is formulaic language use. "Although case studies are typically discussed under the label of qualitative research (because a single case cannot be representative of a population), actual case studies often include quantitative data collection instruments as well" (Dörnyei, 2007, p. 152). Duff (2008) points out that mixed methods data analysis is also appropriate in case studies.

4.3 Participants.

A total of eleven participants (9 males, 2 females) took part in the recording sessions. All participants were ages 36 to 60 months, and had been diagnosed with an autism spectrum disorders by a professional.

Participants were selected using criterion sampling. The key characteristics of interest were age and ASD diagnosis. Additionally, all participants were currently or had previously been clients of our Center for Language Teaching and Rehabilitation, Institute of Linguistics. Not only did this create a point of contact between potential participants and the researcher, but also it ensured that participants would be comfortable interacting with the Center during the play session. No restrictions were made based on sex, but given that ASD affects three to four times more males than females (CDC as cited in Kim & Lord, 2013), it was expected that there would be more male participants.

4.4 Sources of data

For each participant, a one-hour play session with his/her current SLP was audio recorded. The SLP, rather than a parent, was selected as an interlocutor to minimize variations in interpersonal factors across participants. The researchers did not participate as an interlocutor during the play session because previous research has shown that the rate of echolalia tends to increase in interactions with unfamiliar interlocutors. The audio recording device was a Zoom H1 Handy Portable Digital Recorder. The audio recording settings were: WAV at 96Hz at 16Bit.

4.5 Participant Profiles

This section provides background information for each of the participants as well as a brief description of their recording session.

Table 1: *Summary of Demographic Information for Participants*

Pseudonym	Sex	Age at time of recording (months)	Age at ASD diagnosis (months)
T01	Male	37	36
T02	Female	39	30
T03	Male	41	38
T04	Male	44	42
T06	Male	47	40
T07	Male	50	48
T08	Male	53	36
T09	Male	54	48
T10	Male	55	54
T11	Female	59	54

4.6 Quantitative Analysis of Formulaic Language

This section provides the results of a quantitative analysis of formulas from each participant's transcript. This section presents and discusses: (a) the distribution of formulaic and non-formulaic speech by word count, (b) the distribution of formulaic expressions across categories, (c) the distribution of formulaic expressions by function, and (d) the variability of formulaic expressions. Excerpts selected for quantitative analysis were chosen randomly to prevent biases that would confirm or disconfirm the researcher's expectations. They ranged from 4000 to 11000 words each. Word counts were based exclusively on participants' utterances. Coded transcripts for the selected segments are as following Figure 1, along with a description of the ongoing activity. Each coded transcript is followed by a list of formulaic expressions organized by category.

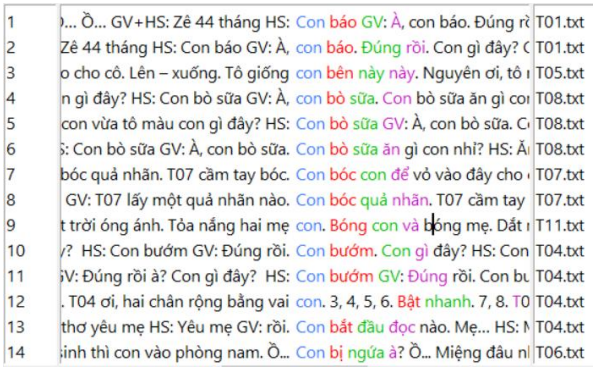


Figure 1: Excerpt concordance of the transcripts

4.7 Distribution of formulaic and novel language

As the actual length of excerpts varied at length, the distribution by word count of formulaic and novel language between participants was compared using percentage scores (WC%) as opposed to raw scores. Figure 2 illustrates the overall distribution

of novel and formulaic language, which has been subdivided into unconventional verbal behavior (UVB) formulas, and conventional formulas. Unconventional verbal behavior formulas include immediate and delayed echolalia, as well as perseveration. Conventional formulas include all other types of formulaic expressions.

Based on the data in Figure 2, the total WC% of formulaic language, including both UVB and conventional formulas, varied between 30% and 80% in the excerpts selected. Thus, while all speakers used formulaic language, they did not use it to the same extent. Furthermore, the figure above indicates that T02 did not use any formulas that were classified as UVB in the excerpt selected for analysis. According to Wray's (2002) model of the balance of holistic and analytic processing, it would be expected that the older participants, T02 and T04, would use more formulaic language than the younger participants, T01 and T04. However, Figure 2 illustrates that this was not the case in the excerpts selected for analysis. T04 used the most formulaic language as measured by WC%, but T02 used the least of all eleven participants. This variability across speakers is not surprising as the excerpts for analysis were randomly selected. Thus, the contextual factors for each segment varied considerably.

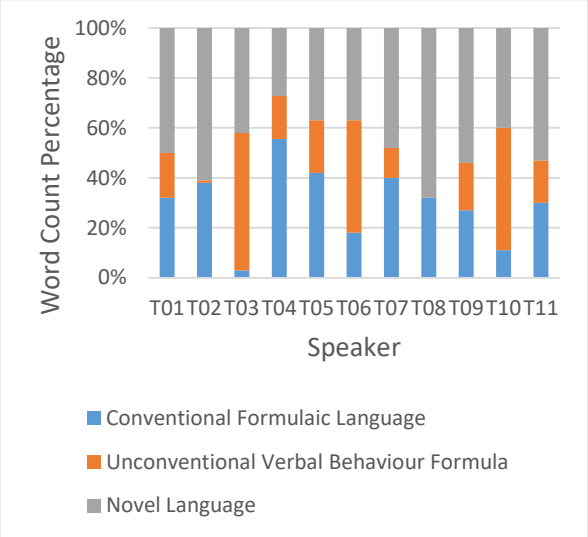


Figure 2. Distribution by word count percentage of novel and formulaic language, by speaker.

As discussed in the previous section, contextual factors influence the prevalence of formulaic language in conversational speech. In terms of ongoing activity, the participants engaged in a number of activities in the segments, including playing with trains, animal figures, and puppets, and participating in conversation without any play. The cognitive and conversational demands placed on the participants in each segment varied according to the type of play, and so the distribution WC% of formulaic language likely reflects these differences in situational factors. If more segments had been analyzed for each participant, age trends may have been clearer. Although there is a relationship between overall number of formulaic expressions and WC% of formulaic and novel language, this measure of formulaicity does not distinguish between the length of sequences and their frequency. By measuring formulaic language use according to WC%, the length of formulas may be a confounding factor in the relationship between formulaic language use and age. Discrepancies such as the difference between T03 and T04 in their respective WC% of conventional formulas are highly related to the length of sequences. While T04's conventional sequences included a number of phrase-long exemplars, T03's longest conventional sequence in the excerpt was 5 words long, while the rest were 3 words or less. The frequency of sequences and its relationship to WC% of formulaic and novel language is addressed later. Another aspect of the distribution of WC% of formulaic and novel language that was of particular interest was the WC% of unconventional verbal behavior compared with all other types of speech. In the Introduction section, an alternate explanation was proposed for Van Lancker Sidtis's (2012) observation-based account of formulaic language in ASD. Van Lancker (2012) put forward the notion that the speech of high-functioning persons with ASD is lacking in formulaic language while the speech of low-functioning persons with ASD is high in formulaic language. However, in school-aged children, the distinction between high-functioning and low functioning autism in school-aged children is based on expressive language abilities, such that strong expressive language is associated with high-functioning autism. Therefore, expressive language abilities can be used to approximate the distinction between high- and low-functioning autism made by Van Lancker Sidtis (2012). Van Lancker

Sidtis's (2012) observations regarding the prevalence of formulaic language do not align with previous research findings, which indicate that formulaic language appears to be characteristic of the entire verbal ASD spectrum. Therefore, it was suggested that Van Lancker Sidtis's (2012) impressions were potentially the result of equating formulaic language with UVB and novel language with all other speech, including conventional formulas. These operational definitions have been observed elsewhere in ASD research. Thus, according to a revised set of definitions, it was proposed that Van Lancker Sidtis (2012) in reality was observing that UVB, and not formulaic language as a whole, decreases with an increase in expressive language abilities. As shown in Figure 2, all participants in this study, regardless of language abilities, used formulaic language. Thus, it is worth investigating the hypothesis discussed above and in the Introduction to attempt to reconcile conflicting accounts of formulaicity in ASD. As the distinction between high-functioning and low-functioning autism in school-aged children is based on expressive language (Tager-Flusberg et al., 2005), participants were ranked in terms of their expressive language abilities to approximate the distinction made between persons with ASD using the previously-mentioned labels. Using this ranking, it was possible to determine whether the participants' WC% of novel and formulaic language as defined in formulaic language research coincided with Van Lancker Sidtis's (2012) observations. If they did not, then the proposal that the observations were based on different operational definitions of formulaic and novel language could be tested. Participants were initially ranked according to their expressive language abilities based on the Child participant profile questionnaire. Based on these measures, T06 had the strongest expressive language skills, as his mother indicated he never has poor expressive language and uses complex sentences. Conversely, T07 had the weakest expressive language skills based on these measures, as his mother indicated that he uses one-word utterances and frequently has poor expressive language. T09 and T10 were more closely matched; while T09 uses compound sentences and T04 uses complex sentences, T10 always has poor expressive language while T09 occasionally does. Thus, based on the questionnaire items, the participants were ranked from strongest to least strong expressive language skills as follows: T03, T06, T09, T10, T11, T01, T02, T04, T05, T08, and T7. The SLP was also asked to rank the participants according to their expressive language abilities based on her professional experience working with the participants and her observations during the recording sessions. Her assessment agreed with the ranking established using the Child participant profile questionnaire. She also indicated that T01 and T04 were difficult to rank as they have different strengths. Therefore, it is with a certain degree of confidence that we can rank the participants based on their expressive language abilities as follows: the strongest is T03, followed by, T01, T04, and T03. With this ranking established, the first step in testing the application of Van Lancker Sidtis's (2012) observations in this study was to compare the ranking by expressive language abilities with the ranking by WC% of formulaic language presented in Figure 2. Based on WC% of formulaic language, T03 used the highest proportion of words in formulaic expressions, followed by. T06, T09, T10, T11. Conversely, according to Van Lancker Sidtis's (2012) observations and the participants' expressive language skills, they should have been ranked according to their WC% of formulaic language as follows: T03, T06, T09, T10, T11, T01, T02, T04, T05, T08, and T7. As the expected ranking and the actual ranking of participants based on WC% of formulaic language did not coincide, it was concluded that the use of formulaic language was moderated by factors other than or in addition to expressive language abilities. As participants' use of formulaic language did not coincide with Van Lancker Idtis's (2012) account of formulaic language in ASD, the proposal that the observations were based on different operational definitions of formulaic and novel language was tested.

4.8 Summary of Findings

This study used several approaches to examine the prevalence and nature of formulaic language use in the interactions of eleven Vietnamese children with ASD. This section summarizes the findings and discussions of the context factors analysis, the quantitative analysis of formulaic language, and the qualitative analysis of formulaic sequences.

The summary is organized in relation to the research questions that guided this study.

The research questions presented in the previous section are repeated below for the purpose of reviewing the issues that were of interest in this study.

Question 1. Do children on the verbal ASD spectrum with varying language abilities use formulaic language in interactions?

This study found that children on the verbal ASD spectrum with varying language abilities did use formulaic language in conversation. The overall balance of formulaic to novel language varied from participant to participant. The percentage of words in formulaic expressions ranged from 30% to 80% of the total word count of each participant's excerpt. This variation is not surprising given the various situational factors at play. While some factors impacted the language system more globally, others had a more direct impact on language use at the moment of speech.

Question 2. How are the form and function of formulaic expressions related in the interactions of children on the verbal ASD spectrum with varying language abilities?

Various form-function combinations of formulas were observed in the speech of the participants. Functional uses of formulas were expected based on results of the quantitative analysis that indicated that all eleven participants used more formulas associated with pragmatic functions than formulas without a pragmatic function or with no function whatsoever. Nonetheless, a qualitative analysis was required to confirm the relationship between form and functional uses of formulaic sequences because of the quantitative analysis's focus on identification and classification by form. In the quantitative analysis only immediate and delayed echolalia were analyzed in terms of function. The functions of all other formulas were implicit based on their respective categories. For example, all collocations belonged to the group of formulas that had no pragmatic functions while all conventional expressions belonged to a subset of formulas that did have a pragmatic function.

5 Conclusions

The purpose of this multiple case study was to examine both the prevalence and the nature of formulaic sequence use in the interactions of eleven children with ASD within the framework of language as a complex adaptive system. To this effect, the participants took part in a one-hour audio-recorded play session with their speech-language pathologist to gather language samples. The participants' parents acted as informants while observing the play session. The audio recordings and the information provided by the parents were used to create participant profiles and analyze the situational factors surrounding the participants' use of language. The researchers transcribed the recordings in order to carry out quantitative analyses of the prevalence and qualitative analyses of the natures of use of formulaic language. The main findings of the study pertain to three domains: situational factors, the prevalence of formulaic expressions, and the nature of use of formulaic expressions. This study found evidence that several situational factors impacted the participants' language use and that these factors did not act in isolation but rather interacted together. The results of the quantitative analyses indicated that all participants, regardless of expressive ability, used formulaic language. All eleven participants used conventional sequences. Conversely only nine of the eleven participants used unconventional sequences characteristic of disordered language, including immediate and delayed echolalia, and perseveration. All eleven participants used sequences from a range of categories of formulaic language and additionally used varied sequences, such that they were more likely to use a new formula than to repeat an old one. Finally, a qualitative analysis of the nature of use of formulaic language was carried out on 36 exemplars selected throughout the transcripts. The sequences were analyzed in relation to the surrounding speech and ongoing activity to determine the markers of formulaicity associated with each and the sequences' functions in context. The sequences were categorized as belonging to one of three form-function pairings: idiosyncratic formulas with functions, conventional formulas with idiosyncratic functions, and conventional formulas with conventional functions. Participants

used the formulas for a range of functional purposes, both interactive and non-interactive.

References

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford, UK: Oxford University Press.

Duff, P. A. (2008). *Case study research in applied linguistics*. New York, NY: Lawrence Erlbaum Associates.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91–126.

Erman, B. and Warren, B. 2000. The idiom principle and the open-choice principle. *Text* 20: 29–62.

Kim, S. H., & Lord, C. (2013). The behavioral manifestations of autism spectrum disorders. In J. D. Buxbaum & P. R. Hof (Eds.). *Neuroscience of autism spectrum disorders* (pp. 25-37). Oxford, UK: Academic Press.

Nattinger, J. R. and DeCarrico, J. S. 1992. *Lexical Phrases and Language Teaching*. Oxford: OUP.

Nguyễn Thị Hoàng Yến, (2015). *Tự kỉ: Những vấn đề lí luận và thực tiễn*. Education Publishing House.

Noens, I., & Van Bercklaer-Onnes, I. (2004). Making sense in a fragmentary world: Communication in people with autism and learning disability. *Autism*, *8*(2), 197- 218.

Rudolph, J. M. & Leonard, L. B. (2016). Early language milestones and specific language impairment. *Journal of Early Intervention*, 38(1) 41 –58.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tager-Flusberg, H., Paul, R., Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (3rd Ed., Vol. 1, pp. 335-364). Hoboken, NJ: John Wiley & Sons.

Van Lancker Sidtis, D. (2012). Formulaic language and language disorders. *Annual Review of Applied Linguistics*, *32*, 62-80.

Vogindroukas, I., & Zikopoulou, O. (2011). Idiom understanding in people with Asperger syndrome/high functioning autism. *Revista da Sociedade Brasileira de Fonoaudiologia*, 16*(4), 390-395.*

Volden, J., & Sorenson, A. (2009). Bossy and nice requests: Varying language register in speakers with autism spectrum disorder (ASD). *Journal of Communication Disorders*, *42*(1), 58-73.

Vũ Thị Bích Hạnh, (2019) *Đối mặt với tự kỉ: Cùng nhau vượt qua*. Tri Thuc Tre books, Women Publishing House.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28.

The Framework of Multiword Expression in Indonesian Language

Totok Suhardijanto
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
totok.suhardijanto@ui.ac.id

Rahmad Mahendra
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
rahmad.mahendra@cs.ui.ac.id

Zahroh Nuriah
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
zahroh.nuriah@ui.ac.id

Adi Budiwiyanto
Department of Linguistics
Faculty of Humanities
Universitas Indonesia
adi.budiwiyanto71@ui.ac.id

Abstract

This paper presents our attempt to develop an Indonesian multi-word expression (MWE) identification framework. The framework consists of three different steps. In the first step, we surveyed any definitions and categorizations of MWEs in the previous studies. In the second step, sentences from our language corpora are segmented and high-frequency n-grams are extracted using statistical methods. The extraction results which consist of word sequences are evaluated and reorganized by using the phraseological analysis procedure. This procedure consists of polylexicality, fixedness, and idiomaticity. The final step is to recategorize and reevaluate data collected in the second step. In this research, data is collected from newspapers, Wikipedia, and scientific papers. The result shows that in terms of Indonesian MWEs, despite polylexicality and fixedness, idiomaticity should be correlated with compositionality to get a better classification of MWEs.

1 Introduction

Since Firth (1957) stated that in language communication, meaning is usually conveyed by word group, many scholars had explored this idea and had ended up with a concept of regularly significant frequent sequences of words called multiword expressions. Different scholars used different terms for these sequential linguistic phenomena (see Burger, 2015; Fleischer, 1997; Sprenger, 2003; Biber & Conrad, 2019). Burger (2015) and Fleischer (1997) mentioned it as phraseology, while Sprenger (2003) and Bider & Conrad (2019), called these word sequences as fixed expressions. Burger (2015) distinguishes phraseological units or MWEs by three features: polylexicality, fixedness, and idiomaticity, where idiomaticity need not be present in all phraseological units.

Up to now, in the field of linguistics and language studies, the classification of MWEs is still considered as a challenging task. Even more so, as there is no general consensus about what counts as an MWE. Masini (2005: 145) viewed MWEs as "lexical units larger than a word that can bear both idiomatic and compositional meanings. (…) the term multi-word expression is used as a pre-theoretical label to include the range of phenomena that goes from collocations to fixed expressions". In a more detailed way, Sailer & Markantonatou (2018: v) defines MWEs as 'any expression that contains more than one basic lexical element and that is lexicalised, fixed, idiomatic, or irregular in one way or the other.'

For both the natural language application and the linguistic theory, Multiword expressions (MWEs)

are challenging because they are often difficult to be classified by an application of the machinery developed for free combinations where the default is that the meaning of an utterance can be predicted from its structure. Compared to MWEs in European languages, there are only a few number of primarily descriptive works on MWEs for Asian languages, except for Chinese, Japanese and Korean. This paper contributes to look at MWEs in Indonesian. They discuss prominent issues in MWE research such as classification of MWEs, their formal grammatical modeling, and the description of individual MWE types from the point of view of different theoretical frameworks, such as Dependency Grammar, Generative Grammar, Head-driven Phrase Structure Grammar, Lexical Functional Grammar, Lexicon Grammar.

Why do we need to identify and deal with MWEs in natural language processing? There are several reasons that can be explained further. First, it is important to recognize MWEs first before processing and implementing POS tagging. Secondly, in current NLP researches, when a cross-lingual or multilingual approach is adopted to a machine learning project, it needs to identify whether the counter translation in a language is a simple or a complex word? Furthermore, identifying opaque or more idiomatic MWEs such as kick the bucket is also a challenging task, but the task can be very helpful when it comes to improving automatic sentiment analysis of data collected from social networks.

In this paper, our research objectives consist of twofold. First, this study attempted to identify MWEs in Indonesian. Second, this paper also classifies Indonesian MWEs into categories with the same distribution and behaviors. [introduction stub]

2 Classification of MWEs

Although some of the classifications made by computational linguists appear to be different, there are actually similarities in the categories of the classifications. The classifications they developed rely on the previous works. Sag et al. (2002) classify MWEs into (1) **institutionalised phrases**, i.e., sets of words which co-occur often but have no syntactic idiosyncrasy, and whose semantics are fairly compositional, and (2) **lexicalised phrases**, presenting some idiosyncratic syntactic or semantic characteristics. The latter can be further divided into three subclasses according to their degree of flexibility: fixed (e.g., *ad hoc, vice versa*), semi-fixed and syntactically flexible expressions (e.g., *cable car*), and proper names (*San Francisco*), as well as non-decomposable idioms (e.g., *kick the bucket, shoot the breeze*).

Meanwhile, Baldwin et al. (2010) categorised MWEs into three types: (1) **nominal MWEs**, (2) **verbal MWEs** and (3) **prepositional MWEs**. Nominal MWEs are one of the most common types (e.g., *golf club*, *connecting flight*, or *open secret*). The verbal MWE consists of (a) verb-particle construction, also termed *particle verbs* or *phrasal verb*, (e.g., *play around*, *take off*, *cut short* and *let go*); (b) prepositional verbs, (e.g., *refer to*, *look for*, *come across*, and *grow on*); (c) Light verb constructions (e.g., *do a demo*, *give a kiss*, *have a drink*, and *take a walk*; and (d) Verb-noun idiomatic combinations (VNICs, also known VP idioms), e.g. *kick the bucket*, *shoot the breeze*, and *spill the beans*. The prepositional MWEs comprises (a) determinerless-prepositional phrases (e.g. *on top*, *by car*, and *high expense*) and (b) complex prepositions (e.g. *on top of*, *in addition to*, and *with regard to*).

Based particularly on the syntactic and semantic properties, MWEs can be classified into (1) **lexicalised phrases** and (2) **institutionalised phrases**. Lexicalised phrases are MWEs with lexical, syntactic, semantic or pragmatic idiomaticity. Lexicalised phrases can be further split into a) fixed expressions (e.g., *ad hoc*, *at first*), *semi-fixed expressions* (e.g., *spill the beans*, *car dealer*, *Chicago White Socks*) and syntactically-flexible expressions (e.g. *add up*, *give a demo*). On the other hand, the class of institutionalised phrases corresponds to MWEs which are exclusively statistically idiomatic (e.g., *salt and pepper*, *many thanks*).

Ramisch (2015: 42-44) makes a simplified typology based on (1) the morphosyntactic role of the whole expression in a sentence and (2) its difficulty to be dealt with using computational methods. They divide MWEs into three categories: (1**) nominal expressions** consisting of nominal compounds (e.g., *traffic light, Russian roulette, degree of freedom, wine glass*), proper names (e.g., *United Nation, Porto Alegre, Alan Turing*), and multiword terms; (2) **verbal expressions** comprising (a) phrasal verbs which are divided into transitive prepositional verbs (e.g., *rely on, agree with*) and more opaque verb-particle constructions where the particle is actually attached to the verb,

forming a cohesive lexical-semantic unit (e.g., *give up, take off*), and (b) light verb constructions (e.g., *take a shower, make a presentation*); and (3) **adverbial and adjectival expressions** (e.g., *upside down; second hand, on fire, at stake,* and *in the buff*).

In addition to those three main types, they also define three orthogonal types that are more related to the computational methods used to treat MWEs. The first class is **fixed expressions** that can be dealt with using relatively simple techniques. They correspond to the fixed expressions of Sag et al. (2002). The second class is **idioms**, that are very hard to recognise and require the use of external semantic resources. The last class is called **true collocations** that correspond to the notion of words that co-occur more often than expected by chance.

Müller et al. (2015: 669) give a list of MWEs though it is not intended as a classification of MWE: (1) proverbs (A bird in the hand is worth two in the bush), quotations (Shaken, not stirred) and commonplaces (One never knows); (2) metaphorical expressions (as sure as eggs is eggs); (3) verbal idioms (to kick the bucket); (4) particle/phrasal verbs (to make up); (5) light verb constructions/composite predicates (to have a look); (6) syntactic/quasi noun incorporation (to wash car, to play piano, to buy (a) house, to have/own a car); (7) stereotyped comparisons/similes (as nice as pie, swear like a trooper); (8) binomial expressions (shoulder to shoulder, by and by, nourish and cherish); (9) complex nominals (man about town, weapons of mass destruction, sheep's clothing) - Collocations (strong tea, hard frost); (10) fossilized/frozen forms (all of a sudden, instead of, depending on); (11) routine formulas (Good morning, How are you doing?, Happy Birthday). Particularly in French language, Müller et al. (2015) classify MWEs into seven categories: (1) nominal sequence, (2) verbal sequence, (3) adjectival sequence, (4) adverbial sequence, (5) prepositional and conjunctive sequence, (6) determinative sequence, and (7) multi-word interjection.

Constant et al. (2017: 6-8) make a list of MWE categories commonly seen in the literature. The categories are non-exhaustive and can overlap. They cover **idioms** (e.g., *to kick the bucket*), **light-verb constructions** (e.g., *to take a shower*), **verb-particle constructions** or phrasal verbs (e.g., *to give up*), **compounds** (e.g., *dry run, bank robbery, stir fry*), **complex function words** (e.g., *as soon as, up until, by and large*), multiword named entities (e.g., *International Business Machines*), and multiword terms (e.g., *short-term scientific mission*).

Laporte (2018) adapts Baldwin and Kim (2010) classification. He proposes two types of classification which are different in terms of the existence of copula in the support-verb constructions. His classification is based on comparison of MWEs in some languages, such as English, French, Romance, Greek, Arabic, Chinese or Korean.

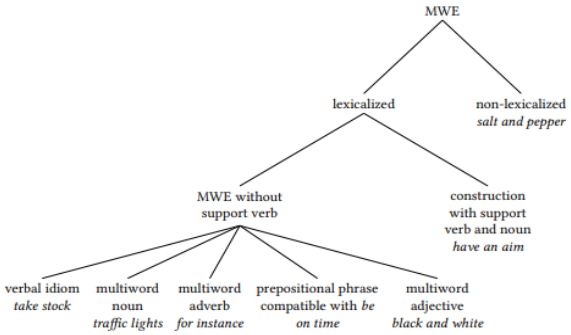


Figure 1: Classification by Laporte (2018) (Escartín et al., 2018)

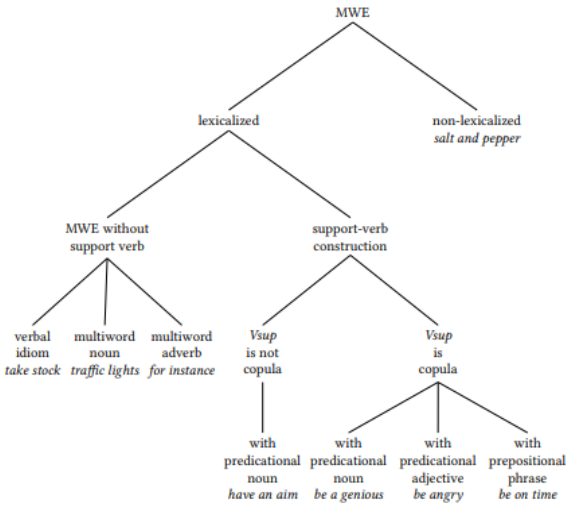


Figure 2: Classification by Laporte (2018) where copula is considered a support-verb (Escartín et al., 2018)

3 MWE Framework in Indonesian

Fleischer (1997) and Burger (2015) mentioned the same three prototypical properties of MWEs: polylexicality, fixedness, and idiomaticity. However, Fleischer proposed one more prototypical property that is not implemented in this study: lexicalization. In our framework, since lexicalization is quite rare in Indonesian MWEs, we did not take it into account to identify and classify MWEs in this paper. According to Fleischer, as MWE is a fuzzy concept, polylexicality is the most obligatory property, while Burger wrote that idiomaticity need not be present in all types of MWEs.

In our framework, MWEs are evaluated from three properties: polylexicality, fixedness, and idiomaticity. First, all MWEs should be made up of more than one word such as *Susilo Bambang Yudhoyono*, *singa laut* 'walrus', and *rumah sakit* 'hospital'. Second, most MWEs cannot be modified such as (i) be inserted by one or more words: *rumah sakit* 'hospital' cannot be inserted with yang 'that' (relative pronoun) become *rumah yang sakit* which literally means 'a sick house'; (ii) be reversed in terms of word order: *singa laut* 'walrus' cannot be reversed into *laut singa* which literally means 'the Lion Sea'. With regard to idiomaticity, we agree with Fleischer and Burger that it is not obligatory to determine whether a word sequence is a MWE or

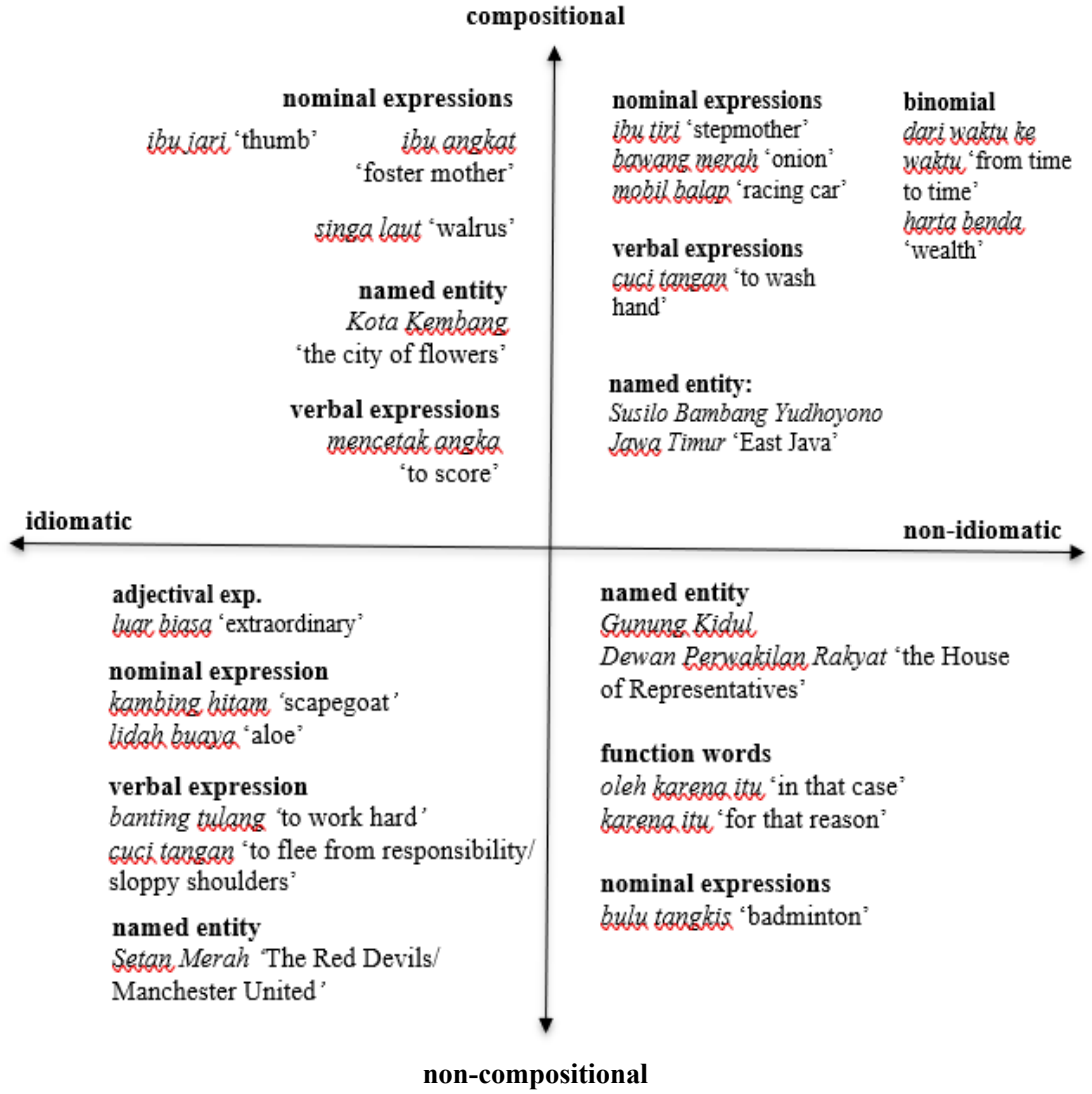


Figure 3: MWE classification based on idiomaticity and compositionality

not. However, in dealing with Indonesian MWEs, we need to consider the idiomaticity as a complex property, rather than a single property. We need to consider the structure of components among MWEs to deal with idiomaticity in Indonesian MWEs.

According to Cruys and Moiron (2007), the linguistic behaviour of MWEs can not be predicted based on the linguistic behaviour of their component words. In addition, Baldwin (2006) characterizes the idiosyncratic behavior of MWEs as "a lack of compositionality manifested at different levels of analysis, namely, lexical, morphological, syntactic, semantic, pragmatic and statistica." MWEs in Indonesian also show idiosyncratic behavior, but some of those still manifest compositionality. Therefore, we classify MWEs in Indonesia based on two axes, that is, idiomaticity and compositionality.

If we look closer at the components, the characteristics of MWEs in Indonesian is not really different from those of MWEs in other languages discussed in the former studies.

- Verbal expressions are verbs that consist of more than one word but refer to one concept. These expressions can consist of a noun, adjective, preposition or other verb next to another verb. Verbal expressions can be an incorporation with a noun as the patient such as *minta izin* 'to ask for permission', *mencetak angka* 'to score', *cuci tangan* 'to wash hand' or 'to flee from responsibility/sloppy shoulder', *banting tulang* 'to work hard' (VN), or a serial verb like *siap saji* 'fast food' (VV), or a verbal combination with an adjective like *bekerja keras* 'to work hard' (VA), or a preposition like *tinggal di* 'to live in' (VP).
- Nominal expressions are nouns that consist of more than one word but refer to one concept. These expressions can consist of an adjective, verb or other noun next to another noun. For example *bawang merah* 'onion', *kambing hitam* 'scapegoat', *ibu tiri* 'stepmother' (NA), *meja makan* 'dining table', *ibu angkat* 'foster mother', *mobil balap* 'race car', *bulu tangkis* 'badminton' (NV), *kata kunci* 'keyword'. The second nouns of nominal MWEs that consist of nouns (N1 N2) semantically can have agentive (*belanja negara* 'state expenditure budget'), patientive (*uji korelasi* 'corellation test'), genitive (*lidah buaya* 'aloe', *ibu jari* 'thumb') or benefactive (*ruang publik* 'public area') or other relation with the first noun (*singa laut* 'walrus'). Some nominal expressions are named entities that consist of a noun and other words (N X(X)) such as *Kota Kembang* 'the City of Flower', *Jawa Timur* 'East Java', *Setan Merah* 'Red Devils/Manchester United', *Dewan Perwakilan Rakyat* 'The House of Representatives', *Gunung Kidul* 'South Mountain', *Susilo Bambang Yudhoyono* (N(X)). There are also binomial with two words or more like *harta benda* 'wealth', *waktu ke waktu* 'from time to time' ((P)N(P)N) and nominal expressions with three words or more *tahun kerja efektif* 'work year' (NVA), *defisit transaksi berjalan* 'current account deficit' (NNV), *berat badan lahir rendah* 'low birth weight' (NNVA).
- Adjectival expressions are adjectives that consist of more than one word but refer to one concept. These expressions can consist of an adjective with another adjective, for example *tegak lurus* 'perpendicular' (AA) or a noun with an adjective for example *luar biasa* 'outstanding' (NA). The construction NA is problematic because this form has no head since the head of a phrase in Indonesian is the word at the beginning (the left one).
- Adverbial expressions are adverbs that consist of more than one word but refer to one concept. These expressions can consist of an adverb with another adverb, for example *lebih kurang* 'approximately' (Adv Adv) or an adverb with a verb for example *lebih lanjut* 'furthermore' (Adv V)
- Function words that consist of more than one word are preposition *di luar* 'outside', *di dalam* 'inside' (PN) or conjunction *akan tetapi* 'but', oleh karena (itu) 'therefore' (PConj).

4 Why Indonesian NLP should care about MWEs

In dealing with identification and classification of MWEs in a particular language, specific problems that are related to the language can arise. With regard to Indonesian MWEs, problems are not only caused by the possible idiomatic meaning of the MWE, but also by the ambiguity of the POS of the MWE's components. For instance, the MWE *minta izin* 'ask permission' consists of the word *izin* 'permission' that is a noun in that case, but in non-formal context, such as in spoken Indonesian, *izin*

can also appear as a verb (*Saya izin datang terlambat* 'I was allowed to come late'*)*.

Another example is shown in the following case. The rule: NP → N V is actually not a valid context-free grammar rule because a NP sequence usually consists of two or more nouns: N N (*batu pasir* 'sandstone') or N N N (*batu pasir Navajo* 'Navajo sandstone'). Meanwhile, a sequence with N followed by V is usually treated as a clause rather than a phrase, such as in *Jakarta banjir* (NV) which means 'Jakarta floods'. However, in another NV sequence, such as *rumah makan* 'restaurant', it is considered as a phrase and categorized as a MWE. If MWEs are not processed first, *rumah* 'house' and *makan* 'eat' are treated as two separated words. The first word would be tagged with the noun category, while the latter would be tagged with the verb category.

Suppose that MWE is not processed, the meaning of *rumah makan* may be derived by applying some inference rules when concatenating the words *rumah* and *makan*". The word *rumah makan* 'restaurant' can be defined as 'the house where to eat'. However, this heuristic should fail in another case like "rumah sakit" (*hospital* in English). If we apply the same rule as previously, combining the word "rumah" and "sakit" (*sick* in English) may come up with a meaning of the house where to be sick.

On the other hand, in spite of following the same pattern with phrase constructions in Indonesian language grammar, several MWEs are idiomatics that have implications of totally different meaning. For example: *kambing hitam* '*scapegoat*' vs *anjing hitam* 'black dog'. So, for those reasons discussed above, the processing of MWE is an ideal step to lighten the burden of lexical semantics processing.

5 Conclusion

In this paper we proposed the framework for identification and classification of Indonesian MWEs. We evaluated MWEs in Indonesian by implementing three characteristics: polylexicality, fixedness, and idiomaticity, to categorize Indonesian MWEs. However, in order to determine a sequence as a MWE in Indonesian, we need to correlate idiomaticity and compositionality to make the classification result more clear. MWE processing should benefit the natural language processing tasks in Indonesian language. We have illustrated a few examples in which ignoring the MWE phenomenon can lead to such problems in POS Tagging, Syntactic Parsing, and Lexical Semantics Processing.

In a future work, we will study how to identify other MWE properties, such as discontinuity and variability, for Indonesian that have not been discussed in this paper. We also expect to extend our work to extract MWE lexicon from large corpora in Indonesian language and incorporate Indonesian MWE lexicon to other language resources and use them in Natural Language Processing tasks. Another interesting direction to further investigate is to include Indonesian MWEs from spoken registers because we found that MWEs in spoken Indonesian are more challenging to deal with.

References

Carla Parra Escartín, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Look*ing* for a taxonomy. In Manfred Sailer & Stella Markantonatou (eds.)*, Multiword expressions*: *Insights from a Multilingual Perspective*, 271–323. Berlin: Language Science Press. DOI:10.5281/zenodo.1182597.

Carlos Ramisch. 2015. *Multiword expressions acquisition*: *A generic and open framewor*k. Cham/Heidelberg/New York/Dordrecht London: Springer.

Éric Laporte. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer and Stella Markantonatou (eds.), *Multiword expression*s: *Insights from a multi-lingual perspective*, 143–186. Berlin: Language Science Press. DOI:10.5281/zenodo.1182597

Harald Burger. 2015. *Phraseologie: Eine Einführung am Beispiel des Deutschen*. 5th edn. Berlin: Erich Schmidt Verlag.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002), 1–15

Manfred Sailer and Stella Markantonatou. 2018. *Multiword Expressions: Insights from a Multilingual Perspective*. London: Language Science Press.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43, no. 4 (2017): 837-892.

Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer. 2015. *Word Formation*: *An International Handbook of the languages of Europe.* Vol. 1. Berlin/Boston: De Gruyter Mouton.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. Boca Raton: CRC Press.

Van de Cruys, T. and Moirón, B.V., 2007, June. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions (pp. 25-32).*

Wolfgang Fleischer. 1997. *Phraseologie der deutschen Gegenwartssprache*. 2nd edn. Tübingen: Niemeyer.

Bilingual Multi-word Expressions, Multiple-correspondence, and their Cultivation from Parallel Patents: The Chinese-English Case

Benjamin K. Tsou
City University of Hong Kong
The Hong Kong University of Science and Technology / Hong Kong SAR
btsou99@gmail.com

Ka Po Chow
Chilin (HK) Ltd.
Hong Kong SAR
Kapo.rclis@gmail.com

John Lee
City University of Hong Kong
Hong Kong SAR
jsylee@cityu.edu.hk

Ka-Fai Yip
Yale University / New Haven
Connecticut, United States
kafai.yip@yale.edu

Yaxuan Ji
The Hong Kong University of Science and Technology / Hong Kong SAR
yjiaf@connect.ust.hk

Kevin Wu
Chilin (HK) Ltd.
Hong Kong SAR
kjwuhk@gmail.com

Abstract

Multi-Word Expressions (MWEs) typically offer challenges in both linguistics and Natural Language Processing (NLP) and their cross-lingual correspondences also introduce new issues. This paper draws on a specially cultivated corpus of more than 300,000 comparable Chinese-English patents over 10 years [Patentlex: http://patentlex.chilin.hk], and focuses on issues related to bilingual correspondence between Chinese and English technical vocabularies extracted from it in terms of: (1) Non-unique correspondence between cross-lingual terms, which so far has not attracted sufficient interests, (2) Means to cultivate good sources for up-to-date technical terms, (3) A network approach to the weighted multilingual alternate renditions and their presentation through knowledge graphs, and (4) Typological differences in the cross-lingual MWEs, including the internal structure of constituent words and their sociolinguistic-discoursal registers.

1. Introduction

Compound words usually contain more than one constituent words (e.g. *watering hole, space station*) and multi-word expressions (MWEs) contain two or more constituent words (e.g. *sodium bicarbonate, subdural hematoma, ASAP, kicking the bucket, still water runs deep* [1]). MWEs are sometimes referred to as phrasal words, and they can be quasi-autonomous constructions within a sentence. Some have *locus classicus* (e.g. "probing for his Achilles' Heel" [2]) and are within the repertoire of only the well-educated or of those in technical and specialized fields. MWEs typically provide learning challenges for non-native speakers of the language as well (Foster et al. 2014, Wray 2002). The use of MWEs in language is quite pervasive (Jackendoff 1997), approaching half of the adult lexicon (Sag et al. 2002) with reference to WordNet (Fellbaum 1998).

The search for more sophisticated and more extensive resources involving MWEs has surged forward following the accelerated developments in science and technology in the run up to the new Millennium, and with the dramatic improvements in computer power, machine learning and AI to handle big data. As shown in

[1] Idiomatic expressions are also examples of MWEs, and they are found in abundance in many Asian languages (Tsou 2012).

[2] This refers to the weakest part of Achilles's body and is a metaphorical reference to an individual's weakness.

Figure 1, patent registration has seen phenomenal growth in recent decades with China taking the lead since 2016.

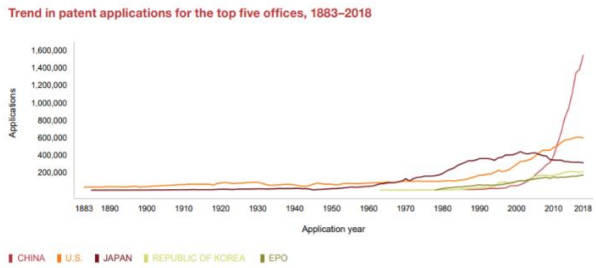


Figure 1. Trend in patent applications

Following rapid changes such as the above and in view of evolving global trade and international relations, there is increasing recognition of the need to overcome cross-lingual communication gaps. Effective efforts to do so entail the processing of MWEs, including handling them in Chinese/English cross-lingual NLP which involves complex texts such as technical manuals, legal documents, contracts and patents in NMT, cross-lingual retrieval and related data analytics.

This paper is organized as follows. Sect. 2 addresses the problems of multiple renditions for MWE translation. Sect. 3 proposes a possible solution, Multiple-Rendition Index, which requires a rigorously cultivated corpus. Sect. 4 introduces our corpus Patentlex and its data cultivation and curation. Sect. 5 compares MWEs in Patentlex with other corpora. Sect. 6 discusses typological differences in the cross-lingual MWEs in terms of correspondence, network representations, stratification and word structure. Sect. 7 concludes.

2. Problems of multiple renditions

There is no simple and regular one-to-one cross-lingual correspondence between any lexical pair of languages, except for very close relatives of the same dialect. For example, the corresponding terms for English mono-morphemic words such as *beef*, *mutton*, and *pork* are respectively bi-morphemic words in Chinese: *niurou* 牛肉, *yangrou* 羊肉, *zhurou* 豬肉 as well as *tunrou* 豚肉. A bilingual dictionary should have the corresponding terms or otherwise inappropriate translations such as the following may result: "I will eat *cow-meat* (for beef), *sheep-meat* (mutton), and *pig-meat* (pork)". We could also have *small cow meat* for veal and either *small lamb* (*xiaoyang* 小羊) or *small lamb meat* (*xiaoyangrou* 小羊肉) for lamb, where the differences between eating meat of the small lamb and eating the small animal could be significant. A single-word or multi-word term may have different renditions in translation according to different contexts. The non-one-to-one correspondence between the simpler kinship terms in Western languages, for example, and the more complex ones in Asian languages are also good illustrations of cultural and cognitive differences being foregrounded. This is especially so for technical terms found in scientific and technical texts and typified by patents. For example, the word "*multiplication*" has more than one word sense.

"Multiplication①: *cheng* 乘"

English: In particular, the invention relates to performing dual complex **multiplication** and complex division using a common circuit.

Chinese Translation: 尤其是, 本发明与使用一共同电路执行双复数乘法及复数除法有关。[3]

"Multiplication②: *fanzhi* 繁殖"

English: Growth and **multiplication** of microbes is substantial when it changes the viscosity, stability, or other important property of the composition.

Chinese Translation: 当微生物的生长与繁殖改变了该组合物的粘性、稳定性或其他重要特性时，微生物的生长与繁殖是实质性的(substantial)。[4]

These examples show that multiplication has a precise meaning in terms of the number of replications [*Cheng* 乘, a simple word in Chinese] in mathematics or physics, as in the first example. It can have an imprecise meaning [*Yansheng* 衍生 "derive-generate", a compound word in Chinese] when referring to biological reproduction in the second example. There are likewise different senses of "base" in botany, chemistry and electronics.[5]

2.1 Value of authentic usage statistics

The following is a good illustration of a common situation when an MWE with multiple renditions is encountered in the translation of technical texts:

"The invention relates to a steam jet enthalpy heat pump air-conditioning hot water unit which at least comprises a compressor, a

[3] According to International Patent Classification (IPC), this word sense is found mostly in domain G (Physics).
[4] Mostly found in domain A (Human Necessities) in patents.
[5] The development of technical terms in Chinese has a relatively short history, see Shen (2001) and Amelung (2004).

four-way reversing valve, the four-way reversing valve, an outdoor heat exchanger……"

There could be several MWEs within the single technical term found in the above passage which may not be familiar to a translator: (1) *steam jet*, (2) *enthalpy*, (3) *heat pump*, (4) *air-conditioning hot water unit*. If s/he is able to look up an appropriate dictionary, s/he may be bewildered by the multiple alternate renditions available and may have difficulty in obtaining useful information to facilitate any translation task at hand. For each constituent term there are multiple renditions available, as can be seen in the top three alternate renditions given in each case in the following examples. It should be clear that among the examples there may not be easy means to decide on an appropriate choice.

(1) Steam jet:
a. 蒸汽喷射(71.42%)
b. 蒸汽射流(21.42%)
c. 蒸汽喷射器(7.14%)

(2) Heat pump
a. 热泵(98.19%)*
b. 加热泵(0.96%)
c. 供热泵(0.69%)

(3) Enthalpy
a. 焓(88.43%)
b. 焓变(6.02%)
c. 热焓(5.04%)

(4) Compressor
a. 压缩机(85.44%)
b. 压缩器(13.22%)
c. 压气机(1.15%)

It may be difficult for a seasoned human translator to make a simple decision just on the basis of several alternate renditions. Yet, given the additional relevant usage frequency of each alternate rendition, an initial choice may be made more easily, as most automated processing system would do, by selecting the one with the highest usage frequency.

However, simply relying on statistical distribution may be inadequate for a human translator or an MT system, as can be seen in the case of *aocao* 凹槽 "groove" in Chinese and its alternate renditions in English.

(5) *aocao* 凹槽
a. groove (36.36%)
b. recess (30.82%)
c. grooves (25.78%)
d. notch (3.32%)
e. indentation (1.42%)
f. recessed portion (0.45%)
g. recesses formed (0.37%)
h. trenches (0.34%)
i. concave groove (0.2%)

These examples show that in the actual translation workflow, the usage difference between the non-unique correspondences "groove" and "recess" may be small, and that more than just the identification of alternate translations may be required.

2.2 Value of authentic usage examples

The problem of making an informed decision merely based on the statistics and top choice among alternate renditions cannot be underestimated. Information on actual usage may be needed, as illustrated by two examples below involving authentic alternate renditions.

E.g.	Rend.	%	Context
(6) eutectic point	a. ***gong jingdian* 共晶点**	0.55	如图 1 的相图所示，饱和溶液当冷却时，随着溶液浓度向**共晶点**变化，先沉淀出一种溶质组分。
	b. ***gong rongdian* 共熔点**	0.26	一般来说，许多特性是本发明的去冰组合物所需要的，如低**共熔点**，值接近 7.0 的 pH 和低腐蚀百分数。
(7) *culi* 粗粒	a. **coarse particles**	0.56	However, the method for controlling the **coarse particles** contained in the hard coat layer is not established.
	b. **coarse grained**	0.16	An upper surface of the wafer has sintered thereon a dispersion of **coarse grained** capacitor grade tantalum powder 12.

Table 1. Alternate renditions of *eutectic point* & 粗粒

While E.g. (6) shows that there may not be critical difference in content of the top choice and the second choice, this is not always the case. In E.g. (7), it can be seen that the choice between the two alternate established renditions of *culi* 粗粒 would not be correct if it is determined only by statistics. The correct choice between **"coarse particles"** and "**coarse grained**" has to be determined by the local grammatical context, which is more readily appreciated by the human translator who may not be familiar with the lexical item and the subject, but whose knowledge of grammar would enable him/her to make the proper selection much more easily than an automated system which is statistically bound. Thus, the need is great for the appropriate curation of data to include authentic examples of actual use. This is also true for the case of E.g. (5f) "recessed portion" and E.g. (5g) "recesses formed" for *aocao* 凹槽 in section 2.1.

3. Quantifying and encapsulating multiple renditions

MWEs may pose challenges to human translators in terms of (1) multiple renditions and (2) technicality. We postulate that the extent of cognitive and other efforts required to process sentences or texts with multiple renditions should have a bearing on the difficulty level in translation. The effort and time needed to translate unfamiliar technical terms should also pose challenges in terms of the lexical lookups and decisions to be made. MWEs may provide even more challenges to L2 speakers because they would have less exposure than L1 speakers to very relevant actual language use contexts (Foster et al. 2014, Wray 2002). Similarly, L2 translators may be at a disadvantage when it

comes to phrasal words and MWEs in L2. There are at least two key variables which require elaboration: (1) the extent of the multiple renditions of the term, and (2) the relative importance of the associated terms within the technical area in terms of usage frequency. If the hypothesis stands, then we could have a preliminary measure of translation difficulty by which amount of efforts and relevant linguistic knowledge may be compared.

The measure we postulate, called Multiple-Rendition Index (MRI), quantifies the relative ease of translation between any two given texts on the assumption that there is correlation between the extent of multiple renditions and the difficulty in translation. The MRI could considerate two features: (1) The lexical gravity of the item in terms of its frequency of occurrences and (2) The type/token ratio of the related items within specific domains, and generally. Additional weighting may be provided to reflect the status complexity in lexical registers and other factors. But foremost in the efforts should include a good database.

4. Data cultivation and curation

The alternate renditions and relevant statistics given in this paper are drawn from the Patentlex corpus (http://patentlex.chilin.hk). It is a very large collection of Chinese and English patents which have been found to be comparable, if not parallel in content. It provides the rare golden standards in translation because its cross-lingual terms have been produced by top language professionals and could have legalistic consequence.

Based on a special collection of 10 years of Chinese and English patents, Patentlex has been cultivated specially for NLP applications. It took several years to identify the patents registered under different jurisdictions: in China SIPO (State Intellectual Property Office of China), in Europe WIPO (World Intellectual Property Organization), and in America USPTO (United States Patent and Trademark Office); and it took longer to build up the pairs or sets of comparable patents written in Chinese and English, whose contents are identical or very comparable as determined by NLP means. This has involved culling very large and separated collections of English and Chinese patents (9 billion characters in total) to identify the bilingually comparable patents (Lu et al. 2010). By means of a combination of search efforts, [6] more than 300,000 such Chinese-English patents were identified. We then applied a series of alignment algorithms and found initially 45 million bilingually aligned sentences or sentence fragments, statistically determined to be good candidates of parallel pairs (Lu et al. 2010). These initial sentences were further refined, and provided more than 30 million top quality bilingual sentence pairs. An initial subset of these bilingual sentences was fruitfully used in two pioneering NTCIR Patent MT competition in Tokyo in 2009 and 2010 as a training corpus and then assessment norms (Goto et al. 2011).

It should be noted that the statistically alignment results were basically strings of characters in Chinese and strings of words in English, which may not be all well-formed terms. To obtain linguistically well-formed words or MWEs, further efforts have produced nearly 3 million candidates of bilingual terms so far (Lu et al. 2011a, 2011b, 2010; Tsou et al. 2017, 2019). Currently on-going semi-supervised efforts have yielded nearly one million top quality terms and their multiple renditions used in the analysis reported here.

The production flow for the current corpus of bilingual terms is shown below.[7]

Corpus Cultivation		Corpus Curation		
Stage 1	Stage 2	Stage 3a	Stage 3b	Stage 4
Search 9 billion chars of C&E patents	*Identify* 300,000 comparable C-E patents	*Align*/get 45M C-E parallel sent. pair candidates	*Refine* 30M good C-E parallel sent. pairs	*Filter/get* 3M bilingual MWE candidates

Table 2. Data cultivation and curation of Patentlex

The technical language found in patents is quite representative up-to-date within a specified period. A major difference between the genre of patents and of general texts, is in the vocabulary. Table 3 below provides useful comparison between Patentlex and a Pan-Chinese media report database LIVAC, [https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus].

	Doc.-lv.: avg. sent./doc	Sent.-lv.: avg. chars/sent.	Word-lv.: avg. chars/word
CN patent	302.8	54.3	2.12
CN media reports ==> LIVAC	11.5	46.6	1.72

Table 3. Comparisons between pantents & media texts

[6] The primary approach is collocational information, as suggested in Church and Hanks (1990), see also Church (2020).

[7] This collection is bigger than the 7000 preliminary parallel Chinese-English patents reported in Lu and Tsou (2009), as it is much more extensive in size.

Excluding diagrams, an average-size Chinese patent document contains about 300 sentences, which is much longer than the average 11.5 sentences of newspaper texts. More specifically, the average number of Chinese characters per sentence at 54.3 is higher than that of media texts (at 46.6). Also, the average number of characters per word at 2.12 in patents is nearly 25% higher than the 1.72 in Chinese media texts (Tsou & Kwong 2015). It can be readily seen that the compound words and MWEs in patents would outnumber media texts.

By providing both usage frequency and authentic examples, associated with MWEs, Patentlex could assist translators, especially when dealing with multiple renditions. It could also form a basis for MRI which characterizes the translation difficulty for a certain task at hand.

5. Patentlex vs. other corpora

The differences between technical vocabulary and ordinary vocabulary may be also explored. To do this, we compare the technical vocabulary from Patentlex and ordinary vocabulary from LIVAC again in Figure 2.

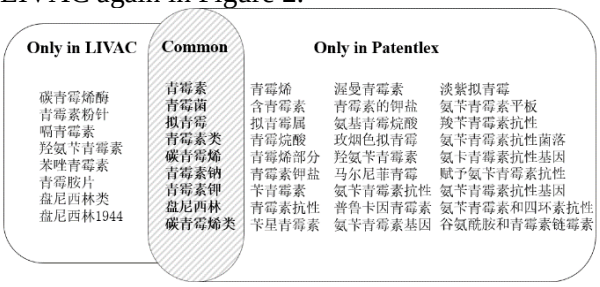


Figure 2. Comparison of the *penicillium* entries from Patentlex and LIVAC

Not surprisingly there are some overlaps. That there LIVAC has some items not found in Patentlex is also not surprising because Patentlex is restricted to a window of 10-year as a retrieval period, in which only new technical developments worthy of protection by law would be reported in the patents. On the other hand, LIVAC reflects topical issues related to the bacteria within their daily life. Notably, a large number of items are found only in Patentlex, which have uncovered just a part of the knowledge base important to our well-being.

Apart from LIVAC, we also compare the Chinese entries containing "青霉" in Sketch Engine and in Patentlex.[8] There are remarkable differences which may result from the data sources of the two corpora. While Patentlex contains 57 terms related to *qingmei* 青霉, Sketch Engine (zhTenTen17) has less (49 items) with 17 overlapping items (see Appendix 1, Table 2 for the examples). However, there is a major difference between the two databases: Patentlex offers also 64 senses in translation related to the 57 terms as well as the distributional statistics of the alternate renditions. Moreover, Patentlex is unique in offering authoritative English translation.

To begin to explore the diachronic development of technical terms, we also compare the Patentlex database with *Modern Chinese vocabulary*, a 1984 Dictionary containing 100,000 entries of Chinese technical terms and ordinary vocabulary. Only 5 terms are found in the 1984 publication, four of which are in common with Patentlex: (a) *qingmei* 青霉 "penicillium", (b) *qingmei-jun* 青霉菌 "penicillium oxalicum", (c) *qingmei-su* 青霉素 "penicillin", (d) *panixilin* 盘尼西林"penicillin", and the fifth is (e) *qingmei-gua* 青霉瓜 "penicillium melon"[9]. This paucity in overlap is not a reflection of extreme developments in the field but of the undeveloped efforts in data cultivation 36 years ago.

6. Cross-lingual MWEs in Patentlex

6.1 Cross-lingual correspondence of MWEs in Chinese and English

The preliminary collection of bilingual technical term candidates exceeds 3 million. While on-going rigorous efforts are made to select the best sets, we can meanwhile report on a preliminary analysis of the lexicons in Chinese and English technical texts on the basis of patents.

Some overall cross-lingual characteristic differences from 100,000 representative items are given below:

	E-to-C	%	C-to-E	%
1 to 1	46843	71.35	74323	85.42
1 to 2	10617	16.17	9302	10.69
1 to 3	3787	5.77	2070	2.38

[8] The source of the Sketch Engine "Chinese Web 2017 (zhTenTen17) Simplified" corpus is mainly the Chinese media and web pages from three Chinese communities, while Patentlex focuses on the patents registered under different jurisdictions.

[9] It has not been possible to trace the source of this term.

1 to 4	1731	2.64	752	0.86
1 to ≥5	2675	4.07	565	0.65
Total	65653	100	87012	100

Table 4: E-to-C & C-to-E multiple renditions

It can be seen from Table 4 that there are noticeable differences among the cross-lingual multiple renditions of the terms. We note that one-to-one translated terms dominate both going from English to Chinese (E-to-C), and from Chinese to English (C-to-E) at 71.35% and 85.42% respectively. Furthermore, 1 to 2 and 1 to 3 alternate translations contribute to the next big group in both directions of translation. It is notable that the percentage of E-to-C multiple renditions (28.65%) almost doubled than that C-to-E (14.58%). Moreover, the correspondence could be as many as 1 to 42 for E-to-C, *and* 1 to 19 for C-to-E cases.

This asymmetry of multiple renditions between the English base and the Chinese base is striking, and invites explanations which may be due to inherent linguistic and lexical differences or due to the direction of translation in the creation of the bilingual documents. If it is the latter, it may be suggested that translation is into a new domain. When there are comparatively inadequate reference materials just as the field of knowledge is developing, there could be many alternate renditions as attempts to create new terms are being made before statistical priorities are established. This is, however, basically a conjuncture which should be evaluated against other more deeply related causes of linguistic difficulty.

As an example, we can compare terms in the field of antibiotics, such as renditions of *penicillium* in Table 5 (see Appendix 1, Table 1 for the examples):

Renditions	E-C	%	C-E	%
1 to 1	17	73.91	26	89.65
1 to 2	3	13.04	3	10.34
1 to 3	3	13.04	0	0

Table 5: Distributions of *penicillium* renditions

For *penicillium*, the gap in asymmetry is larger than the general average in Table 4.

6.2 Network representations of MWEs

The semantically related terms are interrelated and may be represented in a network structure relevant to the mental lexicon. An attempt is made in Figure 3 to show three levels of associated renditions for Chinese *chukou* 出口:

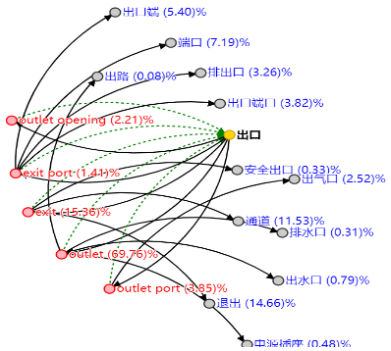


Figure 3: Network representation of *chukou* 出口

In level 2, we have 5 different English renditions, with their usage distributions in the 300,000 patent corpus are given: (a) "outlet" (69.76%), (b) exit (15.36%), (c) outlet port (3.85%), (d) outlet opening (2.21%), and (e) exit port (1.41%).

In level 3, for example, English "outlet" has 4 Chinese renditions (a) *paishuikou* 排水口 (0.31%), (b) *chushuikou* 出水口 (0.79%), (c) *dianyuanchazuo* 電源插座 (0.48%), and the original rendition *chukou* 出口. In addition to "outlet", the other words also have similar multiple renditions. There could be also level 4 and 5. This expanded network of alternate renditions offers a broader view of an important aspect of the lexicological structure of the target terms for the non-causal translators, as well as for the lexicologists and lexicographers.

6.3 Stratification and structure of MWEs

The constructions of MWEs can be quite different in Chinese and in English, as in the *penicillium* case (see Appendix 1, Table 1 for the examples). For Chinese, both semantic adaptations (e.g. No. 3 *qingmeisu* 青霉素 "green-mildew element") and phonetic adaptations (e.g. No. 4 *panixilin* 盘尼西林 "penicillin") are found (and see No. 31 *marnifei-qingmei* 马尔尼菲青霉 "penicillium marneffei" for a hybrid case of phonetic & semantic adaptations). [10] Also the semantic adaptation *qingmei* 青霉 forms the primary base for the majority of terms. The use of Latin and latinate words derived from *penicillium* in English is by far the dominant mode[11]. It can be seen then that the translator from Chinese to English will be at a disadvantage if s/he knows no Latin words

[10] The phonetic mode of adaptation is more common in some Chinese speech communities than others.

[11] See Tsou (2001) and Shen (2001).

associated with bacteria. This is not only found in medicine, but also in law.

The use of Latin or latinate words in English represents an important differentiation between vocabulary in the popular language and that in the High register or learned language. This stratified situation reflects diglossia in general (Ferguson 1959, Tsou 1983). The latinate terms also provide for a common knowledge base for the European languages.

In the case of Chinese, the differentiation in lexical layers to correlate with registers is more subtly manifested through the use of elements from the Classical Chinese language as Table 6 shows:

Class.	***Md.***	***Eng***	***Class.***	***Md.***	***Eng***
fu 腹	*du* 肚	stomach	*lu* 顱	*tou* 头	head
fu 婦	*nv* 女	female	*chi* 齒	*ya* 牙	tooth
kou 口	*zui* 嘴	mouth	*zhi* 脂	*you* 油	fat
zu 足	*jiao* 脚	foot	*yi* 疫	*bing* 病	decease

Table 6: Classical & Modern Chinese morphemes

For the average Chinese speakers, the words from the Classical language are used mainly in professional and learned vocabulary, (e.g. *fuxie* 腹泻 "diarrhoea", *yushi* 浴室 "bath room", *qinshi* 寝室 "bed chamber") and official documents (e.g. *shou* 售 "sell", *gou* 购 "buy") where at least one morpheme belongs to the Classical Chinese language.

However, the use of Classical Chinese elements in a realistically virtual diglossic environment is much more pronounced in the languages of Sinosphere countries, including Japanese, Korean, and Vietnamese, because of their prior incorporation of the Classical Chinese language and the logographic writing system.

The use of Classical Chinese words in these languages serves a function equivalent to Latin and Latinate words in English and European languages and has similarly facilitated intra-regional communication within Sinosphere countries. This deeply rooted tradition also has served to guide the development of new terms.

The diversity in the origin of some Japanese words provides a broader perspective which includes its more recent contact with English. Thus, for describing emotional relationship in Japanese, there can be three basic lexical items: (a) 好き"*suki*", "like" in native Japanese), for a range of casual to deep feelings of positive emotion, (b**)** 愛 ("*ai-suru*", "love" from Sino-Japanese) for a much more serious and intensive feeling, and c) ラブ ("*rabu*", "love" from English for more recent words such as ラブホテル *"love hotel"*, ラブボート *"love boat"*).

It is noteworthy there is two-way flow so that some High register terms in Chinese have come from Japanese through the logographic circle of Sinosphere languages. For example, *hotei (fating* in Chinese*)* 法庭 "court", *minshiu* (*minzhu*) 民主 "democracy", and *sheji (zhengzhi)* 政治 "politics" were first coined in Japan during the Meiji Era before they were introduced into China.[12,13]

The case of "appendectomy" in Japanese and Chinese may provide useful comparison.

(8) *Kyusei-kaifuku-chusui-setsujo*
急性 開腹 蟲垂 切除 (Japanese)
Jixing-mangchangyan-kaidao-shoushu
急性 盲肠炎 开刀 手术 (Chinese)

The Chinese term *mangchang* 盲肠 "blind intestine" refers to the appendage at the end of the digestive track. It first appeared in Japanese as 蟲垂 "hanging worm", a semantic adaptation of the Latin term "vermiform" (worm shape), through the Dutch language, whose speakers along with the Portuguese were the earliest Western visitors to Japan. Lexical stratification is common in languages and serve certain useful social-cognitive functions of differentiation within the society.

Another example is **"subdural hematoma"**, the term referring to "blood clot under the skull",

(9) *Komaku-ka-kesshu* 硬膜下血腫 (Japanese)
Yinnaomo-xia-xuezhong 硬腦膜下血腫 (Chi.)

The English term draws from Latin "sub" and "dura", and from Greek "hematoma", while the Sino-Japanese term refers to "blood swelling under the (hard) skull membrane" [14]. This term would not be easily understood by the man on the street in Japan if it was spoken, but reading

[12] While many words related to modern governance in Chinese have come from Japanese, many words related to cuisine in English have come from French. such as "boeuf" (beef), "mouton" (mutton), and "porc" (pork) as discussed in Section 2.

[13] Similarly, many terms related to Western medicine were first translated in Japan before being adapted in China. This is especially true for matters relating to surgery, such as "appendectomy", which essentially did not exist in China in any significant way. According to Confucian doctrine, the sanctity of body inherited from one's parents should not be violated by unnatural incision. This explains in part why surgery has been a late development in China.

[14] There is simplification in Japanese with the removal of *no* 脳 "brain".

the Kanji characters would improve his comprehension to realize that there is involvement of "brain" and "blood swelling", almost on par with a Chinese man on the street. [15]

6.4 Intra-strata comparison of MWEs

Cross-lingual MWEs may also be compared in terms of headwords. It is noteworthy that while "penicillin" and its derivative terms often function as attributes to other headwords in English, "*qingmei* 青霉" is used more frequently as the headword in Chinese. A good example is "penicillium" in Appendix 1, No. 28 "penicillium citrinum" vs. *juqingmei* 桔青霉.

Another typological difference lies in the number of headwords. A list of 50 common headwords in Chinese MWEs is given in Appendix 2. We note the average number of entries for these top 50 heads is 659, and that the average frequency of occurrence for the 110 top frequency items within the entries at 8100 is quite significant (examples may be seen in Appendix 3). One head may form a number of MWEs in Chinese, but the English headwords involve the use of a large vocabulary of Latin words and display great diversity. The number of headwords in English shall be larger than that in Chinese.[16]

7. Concluding remarks

Among various kinds of MWEs, this paper has singled out bilingual technical terms, purposefully curated from over 300,000 bilingual comparable patents, and has focused on issues related to their non-isomorphic cross-lingual correspondences. We have proposed that the complexity issue may be exasperated by differences in inherent linguistic structure, and that possibly at the inception of terminological development, greater variety and selectivity in the target language may be common when human translation efforts are involved. At the same time, we have pointed out that MWEs in both Chinese and English exemplify passive diglossia with two different lexical layers: the Low language of everyday speech of the population, and the High language known to a much smaller subset of the population. In the case of English, the High register includes Latin or latinate words which are shared by most of the European languages. They also make two important and different contributions: (a) the provision of an easily shared knowledge base in the Western civilization, and (b) differentiation between the status of those who could manage both the Low and High registers in each speech community and those who could not. The same is true of Sinosphere in Asia where some languages have experienced the logographic writing system and the adaptation of some older forms of the Chinese language into their High register vocabulary. This High register layer has contributed to a situation similar to the two-layer system of Europe, with the two different functions. This is also true of speakers of the Chinese language, for whom the contrast between the two registers is less pronounced but no less important. We have also suggested that a Multiple Rendition Index (MRI) measure may be beneficial, and attempted to provide preliminary network representations of MWEs by making use of their multiple renditions and their relative significance within the system, which could have a bearing on the mental lexicon.

Given that technological advancements will always outpace lexicology and lexicography, the integrated study of MWEs through linguistics and natural language processing could go hand in hand to facilitate their management in different applications and our understanding of this important area of language.

Acknowledgements

We wish to thank many individuals who have contributed to the work leading to this publication: Janice Chong, Kenny Mok, Nguyen Thi Hong Quy, Ulrica Nie, Biwei Pan, Belle

[15] There are also instances of portmanteau words in Vietnamese:

(a) Acute appendicitis surgery:
phẫu thuật - viêm - ruột thừa - cấp tính (Vietnamese)
(SinoV.) phẫu thuật 剖術 "operation"; (SinoV.) viêm 炎 "inflammation"; (Viet.) ruột thừa "extra-intestine"; (SinoV.) cấp tính 急性 "acute"

(b) Subdural hematoma:
Tụ - máu - dưới - màng cứng (Vietnamese)
(SinoV.?) Tụ 聚 ? "accumulation"; (Viet.) máu "blood"; (Viet.) dưới "under"; (Viet.) màng cứng "membrane-hard"

Example (a) is more commonly known than (b) and we gather that the latter term is predated by (a). In (a), up to three out of the four constituent words may then root to Sino-Vietnamese words. In the latter case (b), there is only 1 word which might have Sino-Vietnamese origin. These two cases seem to reflect the dwindling use of Sino-Vietnam words in lexical development in Vietnamese, which is different from Japan.

[16] Situations such as this invite comparison between Chinese [Attribute+Head] and English [Head+Attribute] constructions which is beyond the scope of this paper.

Yuan, Skaya Wang, Yuki Wong, and Elaine Zhao.

References

Amelung, Iwo. 2004. Naming Physics. The Strife to Delineate a Field of Modern Science in Late Imperial China. *Mapping Meanings: Translating Western Knowledge into Late Imperial China.* 381-422. Leiden: Brill.

Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. 2009. Statistically driven alignment-based multiword expression identification for technical domains. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 1–8.

Church, K. W. & Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–9.

Church, Kenneth. 2020. Emerging trends: Subwords, seriously? (Keynote Presentation) *The 21st Chinese Lexical Semantics Workshop* (CLSW2020), City University of Hong Kong.

Fellbaum, C. 1998. WordNet: An electronic lexical database. *Language, Speech and Communication*. Cambridge: MIT Press.

Ferguson, Charles A. 1959. Diglossia. *Word* 15(2). 325-340.

Foster, P., Bolibaugh, C. & Kotula, A. 2014. Knowledge of nativelike selections in a L2. *Studies in Second Language Acquisition* 36. 101–132.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin Tsou. 2011. Overview of the patent translation task at the NTCIR-9 workshop. *Proceedings of the NTCIR-9 Workshop*, 559–578.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita, Benjamin Tsou, Masao Utiyama, and Keiji Yasuda. 2013. Database of human evaluation of machine translation systems for patent translation. *Journal of Natural Language Processing 20(1)*: 260-286.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge: MIT Press.

Kwong, Olivia. Y., Benjamin Tsou, and Tom Lai. 2004. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1). 81-99.

Liu, Yuan (ed.). 1984. *Xiandai Hanyu Cibiao* [Modern ChineseVocabulary]. Beijing: China Standard Press.

Lu, Bin and Benjamin K. Tsou. 2009. Towards bilingual term extraction in comparable patents. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation* (PACLIC 23): 755-762.

Lu, Bin, Benjamin K. Tsou, Tao Jiang, Jingbo Zhu and Olivia Y. Kwong. 2011a. Mining parallel knowledge from comparable patents. *Ontology learning and knowledge discovery using the web: Challenges and recent advances*: 247-271.

Lu, Bin, Benjamin Tsou, Jingbo Zhu, Tao Jiang and O.Y. Kwong, 2009. The construction of a Chinese-English patent parallel corpus. *MT Summit XII, 3rd Workshop on Patent Translation*, 17-24.

Lu, Bin, Ka Po Chow, and Benjamin Tsou. 2011b. The cultivation of a trilingual Chinese-English-Japanese parallel corpus from comparable patents. *Proceedings of Machine Translation Summit XIII:* 472–479.

Lu, Bin, Ka Po Chow, and Benjamin Tsou. 2013. Comparable multilingual patents as large-scale parallel corpora. *Building and Using Comparable Corpora (BUCC)* XI: 167-187.

Lu, Bin, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. *Proceedings of The Workshop on Building and Using Comparable Corpora at LREC-2010*, 42–49.

Luk, Robert, Benjamin Tsou, Tom Lai, O. Y. Kwong, Francis Chik, and Lawrence Cheung. 2003. Bilingual legal document retrieval and management using XML. *Software practice and experience* 33, 41-59.

Ren, Z., Y. Lü, J. Cao, Q. Liu, and Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 47–54.

Sag, I., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: a pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, 1–15.

Shen, Guowei. 2001. The creation of technical terms in English Chinese dictionaries from the nineteenth century. *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China.* 287-304. Leiden: Brill.

Tsou, Benjamin K. & Olivia Kwong. 2015. LIVAC as a monitoring corpus for tracking trends beyond linguistics. *Linguistic Corpus and Corpus Linguistics in the Chinese Context (Journal of Chinese Linguistics Monograph Series 25)*, 447-471.

Tsou, Benjamin K. 2001. Language Contact and Lexical Innovation. *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*, 35-56.

Tsou, Benjamin K. 2012. Idiomaticity and classical traditions in some East Asian languages. *26th Pacific Asia Conference on Language, Information and Computation*, 39–55.

Tsou, Benjamin K. 2019. From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration. *Proceedings of the Conference on Building and Using Comparable Corpora (BUCC 2019)*, 29-36.

Tsou, Benjamin K., Derek F. Wong, and Ka Po Chow. 2017. Towards the generation of bilingual Chinese-English multi-word expressions from large scale parallel corpora: An experimental approach. *EUROPHRAS*, 162-168.

Tsou, Benjamin K., Ka Po Chow, Junru Nie, and Yuan Yuan. 2019. Towards a proactive MWE terminological platform for cross-Lingual mediation in the age of big data. *Proceedings of The Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, 21-27.

Tsou, Benjamin K. 1983. Triglossie et realignment sociolinguistique. *Contrastes*. 10-15.

World Intellectual Property Office. 2019. *World Intellectual Property Indicators 2019 -Patents*. Retrieved from https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2019-chapter1.pdf

Wray, A. 2002. *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Appendix 1: MWEs with “penicillium” from different corpora

Table 1: The penicillium words from Patentlex

1	penicillamine	青霉胺
2	penicillic acid	青霉酸
3	penicillin	青霉素
4	penicillin	盘尼西林
5	penicillin acylase	青霉素酰化酶
6	penicillin allergy	青霉素过敏
7	penicillin antibiotic	青霉素抗生素
8	penicillin antibiotic	青霉素类抗生素
9	penicillin antibiotic	青霉素抗生素类
10	penicillin antibiotics	青霉素抗生素
11	penicillin binding protein	青霉素结合蛋白
12	penicillin binding proteins	青霉素结合蛋白
13	penicillin derivative	青霉素衍生物
14	penicillin G	青霉素 G
15	penicillin G	苄青霉素
16	penicillin G sodium	青霉素钠
17	penicillin potassium	青霉素钾
18	penicillin streptomycin solution	青霉素链霉素溶液
19	penicillinase	青霉素酶
20	penicillins	青霉素类
21	Penicillins	青霉素类抗生素
22	penicillins	青霉素类药物
23	penicillium	青霉
24	penicillium	青霉属
25	penicillium	青霉菌属
26	penicillium chrysogenum	产黄青霉
27	penicillium chrysogenum	产黄青霉菌
28	penicillium citrinum	桔青霉
29	penicillium expansum	扩展青霉
30	penicillium italicum	青霉病
31	penicillium marneffei	马尔尼菲青霉
32	penicillium oxalicum	青霉菌

Table 2: MWEs with *qingmei*“青霉” in Sketch Engine (zhTENTEN 2017) and Patentlex

Common	氨苄青霉素 (1818/2643)*, 产黄青霉 (46/64), 桔青霉 (39/13), 拟青霉 (59/102), 拟青霉属(23/67), 羟氨苄青霉素(5/309), 青霉 (1924/995), 青霉胺 (870/408), 青霉病 (205/4), 青霉菌 (1069/11), 青霉属 (115/413), 青霉素 (30886/3753), 青霉素 G (610/352), 青霉素酶 (426/146), 青霉酸 (33/2), 青霉烯 (45/29), 碳青霉烯 (135/103)
Sketch Engine	碳青霉 (939), 碳青霉烯酶 (104), 青霉烯酶 (54), 耐碳青霉 (46), 黄青霉 (40), 青霉索 (34), 青霉菌病 (30), 青霉烷 (21), 蛾拟青霉 (21), 产碳青霉烯酶 (20), 耐碳青霉烯 (16), 青霉烷砜 (15), 绿青霉 (15), 微紫青霉 (12), 蝉拟青霉 (11), 点青霉 (10), 氨苄青霉 (9), 紫青霉 (8), 展青霉 (8), 氨苄青霉霉霉 (SIC) (7), 氧青霉 (6), 青霉类 (6), 苄星青霉霉 (6), 青霉等 (6), 青霉烯酸 (6), 青霉组 (5), 克拉维酸钾/羟氨苄青霉素 (5), 疣孢青霉 (5), 橘青霉 (5), 青霉时 (5), 菌碳青霉 (5)
Patentlex	氨苄青霉素抗性 (393), 青霉素类 (215), 氨苄青霉素抗性基因 (190), 渥曼青霉素 (164), 苄青霉素 (80), 碳青霉烯类 (69), 青霉菌属 (45), 普鲁卡因青霉素 (27),青霉素衍生物 (23), 氨苄青霉素抗性菌落 (20), 扩展青霉 (19), 苄星青霉素 (18),青霉烯部分 (17), 氨苄青霉素平板 (16), 氨苄青霉素和四环素抗性 (15), 青霉素结合蛋白 (15), 产黄青霉菌 (14), 青霉素抗生素 (12), 青霉素链霉素溶液 (12),羧苄青霉素抗性 (12), 青霉素酰化酶 (11), 氨苄青霉素基因 (9), 氨卡青霉素抗性基因 (9), 青霉素钠 (9), 青霉素类抗生素 (7), 氨基青霉烷酸 (6), 青霉烷酸 (6), 赋予氨苄青霉素抗性 (5), 谷氨酰胺和青霉素链霉素 (3), 马尔尼菲青霉 (3), 青霉素过敏 (3), 青霉素抗生素类 (3), 青霉素抗性 (3), 含青霉素 (2),玫烟色拟青霉 (2), 玫烟色拟青霉 (2), 青霉素的钾盐 (2), 青霉素钾 (2),淡紫拟青霉 (1), 青霉素钾盐 (1), 青霉素类药物 (1)

*The numbers refer to the frequencies of occurrence, for common items: (*freq* in Sketch Engine/ *freq* in Patentlex

Appendix 2: 50 Common S/T Headwords from the Patentlex

No.	Headword	Entries
1	器	2022
2	物	1860
3	体	1522
4	(部)件	1275
5	(角)度	1253
6	性	1008
7	量	993
8	剂	980
9	(装)置	962
10	(部)分	931
11	基	772
12	层	738
13	(信)号	730
14	酸	724
15	率	718
16	面	704
17	(系)统	701
18	(分)子	668
19	酯	634
20	(材)料	621
21	(信)息	600
22	(目)的	579
23	(格)式	559
24	线	555
25	化	495
26	(单)元	492
27	点	487
28	(通)道	478
29	数	476
30	构	457
31	力	439
32	(包)括	439
33	素	431
34	(方)法	429
35	(电)流	420
36	(物)质	419
37	(序)列	415
38	(数)据	407
39	(电)路	407
40	胺	396
41	(作)用	395
42	(机)制	391
43	(设)备	388
44	(细)胞	375
45	(类)型	372
46	(区)域	370
47	(反)应	370
48	(组)合	369
49	部	359
50	机	358

No.	Headword	Entries

Appendix3: Examples of MWEs of 10 top frequency Headwords from PatentLex

A. *qi* 器		
English renditions	**Sample entries**	*Freq.*
fully redundant linearly expandable broadcast router	全冗余线性可扩展广播路由器	42
location information domain management server	位置信息域管理服务器	33
fiber bragg grating sensor	光纤布拉格光栅传感器	10
optical recording medium and its corresponding drive	光记录介质及其相应的驱动器	2
complementary metal oxide semiconductor imager	互补金属氧化半导体成像器	1
B. *wu* 物		
acrylate or methacrylate copolymer	丙烯酸酯或甲基丙烯酸酯共聚物	20
aconitrates and citraconates as well as succinate derivatives	乌头酸盐和柠康酸盐以及琥珀酸盐衍生物	20
ethylene alkyl acrylate copolymer	乙烯丙烯酸烷基酯共聚物	2
acrylic emulsions or urethane acrylic copolymer	丙烯酸乳液或氨基甲酸乙酯丙烯酸共聚物	2
ethylene vinyl acetate carbon monoxide terpolymer	乙烯乙酸乙烯酯一氧化碳三元共聚物	2
C. *ti* 体		
diphenylmethane diisocyanate isomers	二苯基甲烷二异氰酸酯异构体	12
cross-linked organopolysiloxane elastomers	交联的有机聚硅氧烷弹性体	10
acrylamide or methacrylamide monomers	丙烯酰胺或甲基丙烯酰胺单体	1
corticotropin-releasing hormone receptor	促肾上腺皮质激素释放激素受体	1
ethylbenzene and all of the xylene isomers	乙基苯和所有的二甲苯异构体	1
D. *jian* 件		
polarization direction rotating elements	偏振方向旋转元件	18
beam shaping optics	光束成形光学器件	18
flip chip semiconductor device	倒装芯片半导体器件	6
optical tool insert	光学加工工具插件	2
conjugated organic semiconductor devices	共轭有机半导体器件	1
E. *du* 度		
low glass transition	低玻璃化转变温度	177
acetylated histone concentration	乙酰化组蛋白浓度	9
bioavailability of metformin	二甲双胍的生物利用度	7
buprenorphine plasma concentrations	丁丙诺啡血浆浓度	2
low crystallinity polymer has a crystallinity	低结晶度聚合物的结晶度	1

F. *xing* 性		
acrylic polymer and a hydrophilic	丙烯酸聚合物和亲水性	3
nitric oxide synthase inhibiting activity	一氧化氮合酶抑制活性	3
dinucleotide repeat polymorphism	二核苷酸重复多态性	2
central dopaminergic neuronal activity	中枢多巴胺能神经元活性	2
acetic acid solution or other acidic	乙酸溶液或其它酸性	1
G. *liang* 量		
comonomer content and molecular weight	共聚单体含量和分子量	3
low density lipoprotein cholesterol levels	低密度脂蛋白胆固醇含量	2
coronary artery blood flow	冠状动脉的血流量	1
average molecular weight of the polyethylene glycols	乙二醇的平均分子量	1
content of vinyl acetate in the copolymer	共聚物中的乙酸乙烯酯的含量	1
H. *ji* 剂		
lipophilic skin moisturizing agent	亲油性皮肤增湿剂	148
phosphite antioxidants	亚磷酸酯抗氧化剂	24
diacylglycerol acyltransferase inhibitors	二酰基甘油酰基转移酶抑制剂	10
ethoxylated alkyl alcohol surfactant	乙氧基化的烷基醇表面活性剂	4
dianionic or alkoxylated dianionic cleaning agent	二阴离子或烷氧基化二阴离子清洗剂	3
I. （*zhuang* 装）*zhi* 置		
portable data storage device	便携式数据存储装置	59
portable inspection data recording device	便携式检验数据记录装置	35
portable radio communication apparatus	便携式无线电通信装置	33
information server memory means	信息服务器存储器装置	13
a portable insulin injection device	便携式胰岛素注射装置	2
J. （*bu* 部）*fen* 分		
erythropoietin portion	促红细胞生成素部分	64
component feed unit control section	元件供送单元控制部分	36
hydrophilic moiety and a hydrophobic moiety	亲水部分和疏水部分	11
donor and corresponding acceptor fluorescent moieties	供体和对应受体荧光部分	1
human monoclonal antibody or a portion thereof	人单克隆抗体或其部分	1