LREC-COLING 2024

The 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia @LREC-COLING-2024 (EURALI)

Workshop Proceedings

Editors
Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu and John P. McCrae

25 May, 2024
Torino, Italia

Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia @LREC-COLING-2024 (EURALI)

ISBN 978-2-493814-33-3
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Introduction

Being the largest continental area on Earth, Eurasia is a hub of more than 2,018 languages from very diverse language families and sub-families, including Afro-Asiatic (Semitic), Austroasiatic, Caucasian, Chukchi-Kamchatkan, Dravidian, Eskimo–Aleut, Indo-European, Japonic, Koreanic, Mongolic, Nivkh, Sino-Tibetan, Tai-Kradai, Turkic, Tungusic, Uralic, and Yeniseian. At the same time, various language communities in Eurasia are under-represented, minoritized, endangered and systematically politically oppressed. Many languages, including Kurdish, Gilaki, Santali, Kashmiri, Laz, and Abkhaz, suffer from a lack of linguistic resources and thus are immediately at risk of digital extinction. Others, such as Shabaki, Talysh, Domari, Korbet, and Bawm, are under-researched in general and run the risk of vanishing completely in the absence of increased support.

Aligned with the pressing need to cultivate language technology for indigenous, endangered, and under-resourced languages across Eurasia, the EURALI workshop is dedicated to catalyzing the development of resources and tools. Our objective is to enhance visibility and foster research for these languages on a global scale. We view the current rapid advancements in language and speech technology, particularly the remarkable progress in large language models, as a unique opportunity for these languages. Moreover, by fostering collaboration among researchers, language experts, and linguists engaged with endangered languages within these communities, our aim is to forge language technology solutions that contribute to the preservation of these languages and elevate their prominence within the realm of language processing.

This year, the EURALI workshop returns for its second edition, set against the vibrant backdrop of LREC-COLING 2024. It offers a thrilling opportunity for our community to reconnect and synergize efforts. However, the presence of numerous concurrent workshops has had a modest impact on our submission numbers compared to EURALI's debut at LREC 2022. The eight selected submissions nonetheless encompass a wide array of aspects and challenges within language technology for Eurasian languages as a whole, with a particular focus on Mambai, Standard Tibetan, Persian, Cantonese, and Khroskyabs.

We extend our gratitude to colleagues who submitted their work to the workshop, the organizers of LREC-COLING 2024, and our dedicated and diligent reviewers; your contributions and support have been vital in making the second EURALI workshop a resounding success.

Workshop Chairs

Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu and John P. McCrae

Workshop Chairs

Atul Kr. Ojha, University of Galway, Galway (Ireland)
Sina Ahmadi, University of Zurich, Zurich (Switzerland)
Silvie Cinková, Charles University, Prague (Czech Republic)
Theodorus Fransen, Università Cattolica del Sacro Cuore, Milan (Italy)
Chao-Hong Liu, Potamu Research Ltd, Dublin (Ireland)
John P. McCrae, University of Galway, Galway (Ireland)

Program Committee:

Abigail Walsh, Dublin City University, Dublin (Ireland)
Aiala Rosá, Universidad de la República - Uruguay, Montevideo (Uruguay)
A. Seza Doğruöz, Ghent University, Ghent (Belgium)
Alina Karakanta, University of Leiden, Leiden (Netherlands)
Alina Wróblewska, Institute of Computer Science, Jana Kazimierza, Warszawa (Poland)
Bogdan Babych, Heidelberg University, Heidelberg (Germany)
Çağrı Çöltekin, University of Tübingen, Tübingen (Germany)
Chao-Hong Liu, Potamu Research Ltd, Dublin (Ireland)
Chihiro Taguchi, the University of Notre Dame, Notre Dame (USA)
Daan van Esch, Google, Amsterdam (Netherlands)
Daniel Zeman, Charles University, Prague (Czech Republic)
Deepak Alok, IIT-Delhi, Delhi (India)
Ekaterina Vylomova, University of Melbourne, Melbourne (Australia)
Elizabeth Sherly, Kerala University of Digital Sciences, Innovation and Technology (India)
George Rehm, DFKI GmbH, Berlin (Germany)
Hiwa Asadpour, Goethe University, Frankfurt (Germany)
Joakim Nivre, Uppsala University, Uppsala (Sweden)
John E. Ortega, New York University (USA)
John P. McCrae, University of Galway, Galway (Ireland)
Jonathan Washington, Swarthmore College, Swarthmore (USA)
Joseph Mariani, LIMSI-CNRS, Paris (France)
Kaja Dobrovoljc, University of Ljubljana, Ljubljana (Slovenia)
Khalid Choukri, ELDA/ELRA, Paris (France)
Luke D. Gessler, University of Colorado at Boulder (USA)
Maitrey Mehta, University of Utah, Utah (USA)
Marie-Catherine de Marneffe, Université catholique de Louvain, Louvain (Belgium)
Mayank Jobanputra, University of Tübingen, Tübingen (Germany)
Olesea Caftanatov, Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chişinău (Moldova)
Ranka Stanković, University of Belgrade, Belgrade (Serbia)
Rico Sennrich, University of Zurich, Zurich (Switzerland)
Ritesh Kumar, Agra University, Agra (India)
Rute Costa, the Universidade NOVA de Lisboa, Lisbon (Portugal)
Saliha Muradoglu, Australian National University, Canberra (Australia)
Sarah Moeller, University of Florida, Gainesville, FL (USA)
Silvie Cinkovà, Charles University, Prague (Czech Republic)
Sina Ahmadi, University of Zurich, Zurich (Switzerland)
Stella Markantonatou, Athena RC, Athens (Greece)
Sourabrata Mukherjee, Charles University, Prague (Czech Republic)
Sylvain Kahane, University Paris Nanterre (France)

Valentin Malykh, MTS AI / ITMO University
Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, Bucharest (Romania)
Victoria Bobicev, University of Moldova, Chișinău (Moldova)
Voula Giouli, Institute for Language and Speech Processing, Athens (Greece)

Table of Contents

Conference Program

Saturday, May 25, 2024

09:00–10:05 **Inaugural Session**

09:00–09:10 ***Welcome***

09:10–10:05 *Keynote talk*
TBD

10:05–10:30 **Oral Session-I**

10:05–10:30 *Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language*
Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo and Ekaterina Vylomova

10:30–11:00 **Coffee break and Poster Session**

10:30–11:00 *Improved Neural Word Segmentation for Standard Tibetan*
Collin J. Brown

10:30–11:00 *Open Text Collections as a Resource for Doing NLP with Eurasian Languages*
Sebastian Nordhoff, Christian Döhler and Mandana Seyfeddinipur

10:30–11:00 *The Extraction and Fine-grained Classification of Written Cantonese Materials through Linguistic Feature Detection*
Chaak-ming Lau, Mingfei Lau and Ann Wai Huen To

10:30–11:00 *Neural Mining of Persian Short Argumentative Texts*
Mohammad Yeghaneh Abkenar and Manfred Stede

Saturday, May 25, 2024 (continued)

11:00–12:15 Oral Session-II

11:00–11:25 *Endangered Language Preservation: A Model for Automatic Speech Recognition Based on Khroskyabs Data*
Ruiyao Li and Yunfan Lai

11:25–11:50 *This Word Mean What: Constructing a Singlish Dictionary with ChatGPT*
Siew Yeng Chow, Chang-Uk Shin and Francis Bond

11:50–12:15 *An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter*
Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout and Reihaneh Rabbany

12:15–13:00 Panel Discussion

12:55–13:00 *Valedictory Session*
Workshop Chairs

Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language

Raphaël Merx[c] **Aso Mahmudi**[m] **Katrina Langford**[t]
Leo Alberto de Araujo[s] **Ekaterina Vylomova**[m]
[c]Catalpa International [m]The University of Melbourne [t]Timorlink [s]Seminario Menor Balide Dili
raphael.merx@gmail.com timorlink@hotmail.com amahmudi@student.unimelb.edu.au
leonberto372@gmail.com vylomovae@unimelb.edu.au

Abstract

This study explores the use of large language models (LLMs) for translating English into Mambai, a low-resource Austronesian language spoken in Timor-Leste, with approximately 200,000 native speakers. Leveraging a novel corpus derived from a Mambai language manual and additional sentences translated by a native speaker, we examine the efficacy of few-shot LLM prompting for machine translation (MT) in this low-resource context. Our methodology involves the strategic selection of parallel sentences and dictionary entries for prompting, aiming to enhance translation accuracy, using open-source and proprietary LLMs (LlaMa 2 70b, Mixtral 8x7B, GPT-4). We find that including dictionary entries in prompts and a mix of sentences retrieved through TF-IDF and semantic embeddings significantly improves translation quality. However, our findings reveal stark disparities in translation performance across test sets, with BLEU scores reaching as high as 21.2 on materials from the language manual, in contrast to a maximum of 4.4 on a test set provided by a native speaker. These results underscore the importance of diverse and representative corpora in assessing MT for low-resource languages. Our research provides insights into few-shot LLM prompting for low-resource MT, and makes available an initial corpus for the Mambai language.

Keywords: low-resource languages, austronesian language, large language models, prompting, dictionary, parallel data

1. Introduction

Large language models (LLM) have shown remarkable abilities to perform natural language processing (NLP) tasks they were not explicitly trained for, including named entity recognition (Mehta and Varma, 2023), text classification (Sun et al., 2023), text summarisation (Zhang et al., 2023b), and machine translation (Hendy et al., 2023; Kocmi et al., 2023; Chowdhery et al., 2022, MT). LLMs can be competitive with traditional encoder-decoder MT models for high-resource languages, but lag behind traditional MT models when translating to and from low-resource languages (Robinson et al., 2023; Hendy et al., 2023; Garcia et al., 2023).

While LLMs can achieve moderately high translation accuracy through zero-shot prompting (Wang et al., 2021), few-shot prompting can improve translation accuracy (Zhang et al., 2023a). Research on the selection of example sentences for use in LLM prompts found that examples close to the source text do not always result in better translation than random examples (Vilar et al., 2023), but that in-domain examples can improve accuracy for technical domains (Agrawal et al., 2023). In particular, for English to Kinyarwanda MT, Moslem et al. (2023) finds an improvement of 11 ChrF points when using in-domain examples instead of random ones.

Using domain adaptation as an analogy, in this paper we explore whether LLMs can be prompted to translate *into* a very low-resource language, through careful selection of sentences and words close to the source text for use in prompting. We work with the Mambai language, a primarily oral language from Timor-Leste with around 200,000 native speakers (Timor-Leste General Directorate of Statistics, 2015). We source prompt examples exclusively from Hull (2001), a language manual which includes parallel English-Mambai sentences and a bilingual word dictionary. We evaluate machine translation quality on both a random subset of sentences from the manual, and on a small corpus of translations collected from a native Mambai speaker.

We find that translation accuracy varies a lot depending on (1) the test set used for evaluation, (2) LLM used for translation, and (3) examples included in the prompt. While 10-shot translation yields BLEU score as high as 23.5 for the test sentences sampled from the language manual used in prompting (with GPT-4 and a mix of sentences retrieved through semantic embeddings and TF-IDF in the prompt), BLEU drops below 5 across all experimental setups for test sentences outside of this domain (novel sentences collected from a native speaker).[1]

Our findings highlight the risks of relying on a

[1]We release the code for extracting the language manual data and for using this data to construct a few-shot prompt given a sentence to translate, as well as the corpus of sentences translated by the paper's author, in https://github.com/raphaelmerx/mambai. The language manual data is available upon request.

2nd EURALI Workshop, pages 1–11

single source when evaluating MT for low-resource languages, especially for languages like Mambai that do not have a standardised vocabulary, orthography, or syntax, where a single corpus can have substantial influence on NLP experiments, despite not always being representative of the language's variations.

2. The Mambai Language

Timor-Leste (also known as East Timor) is a half-island nation in South-East Asia, with a population of 1.3 million as of 2022 (Timor-Leste General Directorate of Statistics, 2022). While its official languages are Portuguese and Tetun Dili (Government of Timor-Leste, 2002, also spelled Tetum), the country has over 30 indigenous languages, from both the Austronesian and Papuan language families (Kingsbury, 2010).

Mambai (also spelled Mambae) is the country's second most common mother tongue after Tetun, with around 200,000 native speakers (Timor-Leste General Directorate of Statistics, 2015). An Austronesian language, it is mostly spoken in the Ermera, Aileu, Manufahi, and Ainaro municipalities (Berlie, 2008), and does not have a standardised orthography (Hull, 2001). It has three distinct varieties, and this article will focus on the southern variety, spoken primarily in the Ainaro, Same, and Hatu-Builico administrative posts (Fogaça, 2013).

Translating to Mambai can bring valuable material closer to Mambai-speaking communities. For example, the Government of Timor-Leste has a mother tongue education program named EMULI, which found that students who were taught in their mother tongue have a higher level in reading comprehension and mathematics than students taught in Portuguese. This program leverages translated material for the curriculum (Gusmão, 2023; Walter, 2016).

Unfortunately, in the taxonomy of Joshi et al. (2020), Mambai would be assigned class 0, "The Left-Behinds", i.e. “languages that have been and are still ignored in the aspect of language technologies”. A search for Mambai sentences on OPUS (Tiedemann, 2009) returns only 36 sentences, all from Tatoeba.[2] To our knowledge, the only NLP tools that claim to support Mambai are language identification models GlotLID (Kargaran et al., 2023) and MMS (Pratap et al., 2023). Mambai does not appear on popular datasets for low-resource languages such as MT560 (Gowda et al., 2021) or FLORES-200 evaluation benchmark (Team et al., 2022).

[2]https://tatoeba.org/

3. Methodology for Data Extraction

As the language does not have any resources in a machine-readable format, we start by digitising the available materials. The general process of data extraction is illustrated in Figure 1.

3.1. Materials

Our primary data source is a Mambai Language Manual (Hull, 2001) that aims to teach the basics of Mambai to foreign speakers, following the Ainaro variety. This 109-page long document includes a pronunciation guide, a grammar, a phrase book, and bilingual dictionaries (English-Mambai and Mambai-English).[3]

To test generalisation of our results, we collaborated with a native Mambai speaker who translated a small corpus of 50 English sentences to Mambai. Since Mambai has no formalised orthography, we tried to keep orthography close to that used in the manual, however we did not aim to produce the same syntactic structures as the manual.

3.2. OCR Process

For the Mambai Language Manual, which we received in paper format, we followed the following OCR process:

1. The book was scanned using an optical zoom camera, which reduces the radial distortion effect and improves the OCR quality;

2. The open-source ScanTailor software[4] was employed to semi-automatically deskew images and make them flat black and white;

3. In the proprietary software ABBYY FineReader 15,[5] we set up a language alphabet, taking into account the characters utilised in each book, with Indonesian (also an Austronesian language) serving as the fallback language, as illustrated on Figure 2. The result of the OCR process was saved in a Word document, preserving font formatting;

4. We then manually separated the extracted data into three collections:

    (a) the section of the manual that contains parallel sentences (14,347 words),

    (b) the section that contains the English to Mambai dictionary (4,023 words),

[3]The author of this book gave his consent to us using it as material, and we acknowledge him as the holder of copyright protecting this intellectual property.

[4]https://scantailor.org/

[5]https://pdf.abbyy.com/

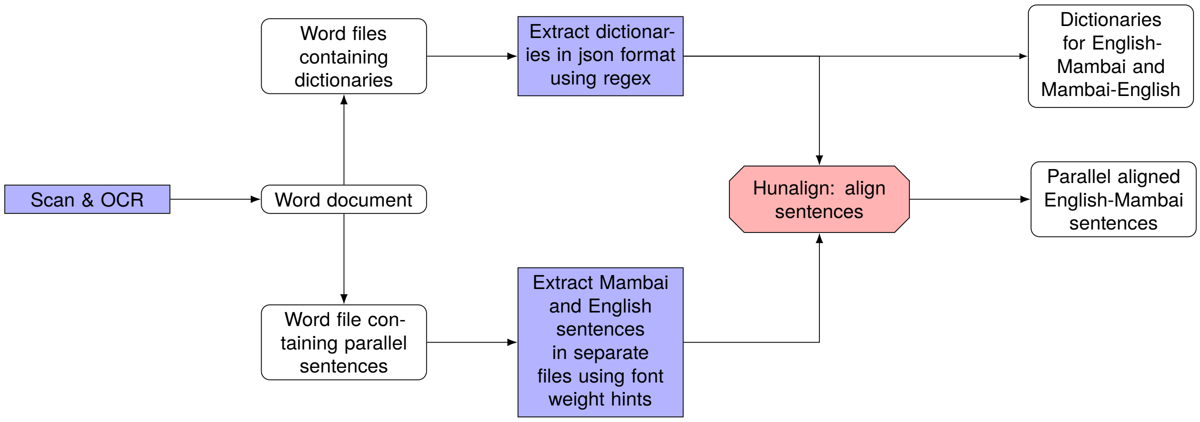


Figure 1: Overview of our process for extracting dictionaries and a parallel corpus from the Mambai Language Manual

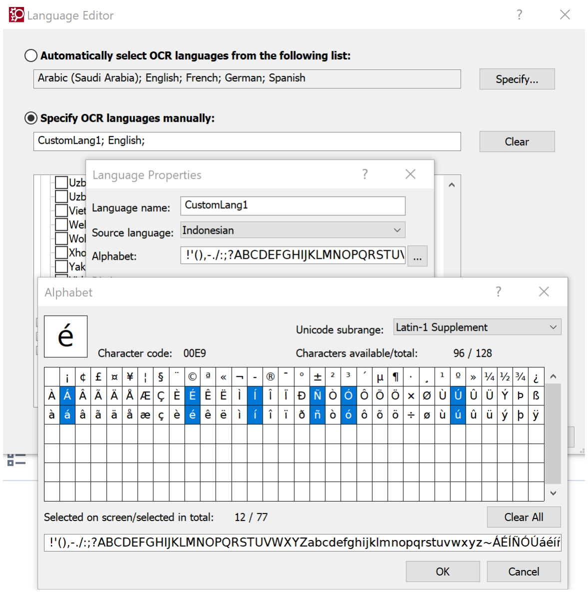


Figure 2: Mambai configuration in ABBYY FineReader 15.

(c) the section of the manual that contains the Mambai to English word dictionary (4,522 words).

3.3. Text Corpora

In this subsection, we present the process of our corpus construction: using the Word documents produced in Section 3.2, we create English-Mambai bilingual dictionaries in JSON format and a corpus of parallel English-Mambai sentences.

3.3.1. Dictionary extraction

For dictionary files, we mined triplets (entry, part of speech, translation) through the following process:

- using the python-docx library,[6] read the file by preserving font weight, and identify text in bold as the dictionary entry;
- use a regular expression to match the part of speech, if any;
- use the rest of the text as value corresponding to the entry;
- if one entry had multiple translations, denormalise them by splitting with ";" and ",".

This process outputs dictionaries in JSON format, one for the English to Mambai direction (1,790 entries), and one for the Mambai to English direction (1,592 entries). Where present, each entry also contains part of speech information, e.g.

```
{
  'entry': 'beik',
  'translation': 'silly',
  'part_of_speech': 'adj.'
}
```

3.3.2. Parallel sentence extraction

Since no embedding models or MT systems support Mambai, we were precluded from relying on sentence embeddings (Thompson and Koehn, 2019) or back-translations (Sennrich and Volk, 2011) to mine parallel sentences from extracted documents. Instead, we rely on a combination of Gale-Church sentence-length information (Gale and Church, 1993) and lexical similarity through the Hunalign[7] sentence aligner (Varga et al., 2007).

We identify Mambai sentences from their bold font-weight, English sentences from their normal font-weight, and section delimiters through text in upper case. For each section, we put the set of Mambai and English sentences in separate text

[6] https://python-docx.readthedocs.io/
[7] https://github.com/danielvarga/hunalign

files, which are fed to Hunalign, along with the bilingual dictionary extracted in Section 3.3.1. Hunalign outputs a series of tab-delimited aligned sentence pairs, with an alignment score for each pair. After manual review of a subset of 100 sentences, we find that setting a score threshold of 0.2 corresponds to keeping a high number of well-aligned sentences, while removing poorly aligned ones. After filtering out sentence pairs below this threshold, we land on 1,187 parallel sentences extracted from this phrase book, from a total of 1,275 potential bitexts.

Since sentences come from a language education manual, they tend to be relatively short, with an average of 5.05 words per sentence in Mambai, and 5.66 words per sentence in English. Some sentences have alternative words in parentheses, which we leave in place, e.g.:

```
"Baléb pôs masmidar lao xa (kafé).",
"Don't put sugar in my tea (coffee)."
```

4. Mambai Translation through Retrieval-Augmented LLM Prompting

After all required data is ready, we now turn to the machine translation part. The general process for translation is illustrated in Figure 3.

4.1. Rationale

Adelani et al. (2022) found that a couple thousand high-quality sentences can substantially increase low-resource MT performance, giving us hope that a language manual with a similar order of magnitude of data could be enough to produce moderate-quality translations.

Working with LLM prompting gives us a flexible format to incorporate both the parallel sentence corpus and the dictionary entries. Further, having access to a phrase book offers substantial domain coverage, in comparison with corpora purely from the religious domain, which are often the only option for low-resource languages (Haddow et al., 2022; Walter, 2016).

Here we work on English to Mambai translation, aiming to address the following research questions:

- Given an English sentence, how can a corpus of bilingual sentences, and a bilingual word dictionary, be incorporated in an LLM prompt to maximise translation accuracy?
- Which LLMs (open-source or proprietary) show the best results for translating into a low-resource language, and what is the observed variance between them?
- How does translation accuracy vary across test sets?

4.2. Methodology

4.2.1. Data setup

Our bilingual corpus of 1,187 parallel Mambai-English sentences is randomly split into 119 (10%) sentences used for testing translation, and 1,068 (90%) sentences for potential use in the prompt, after retrieval selection. Since our objective is to translate full sentences, not individual words, all 1,790 words in the Mambai dictionary are used in prompting.

We also assess translation system quality by providing a different test corpus of 50 sentences translated from English to Mambai by a native speaker of Mambai. This small corpus has relatively simple but slightly longer sentences, with 9 words per sentence on average. The English source sentences were designed to cover a broad range of domains, such as daily life activities, education, health and well-being, family relationships, religion, politics, weather, employment, food and agriculture, technology, personal characteristics, and Timor-Leste specific historical events.

By using the two test sets, we aim to evaluate robustness to variance between domains, as well as estimate risks of overfitting that come from using a test corpus that comes from the same material as the data for prompting. Expected variance between test sets comes from their different authors, their different years of publication (2001 vs 2024), and potentially by them covering different domains.

4.2.2. Prompt

We make use of the best performing prompt template from Peng et al. (2023), to which we add dictionary entries for words found in the sentence, landing on the following prompt template:

```
You are a translator for the Mambai
    language, originally from Timor-
    Leste.

# Example sentences

English: {Sent_eng_1}
Mambai: {Sent_mgm_1}

English: ...
Mambai: ...

# Dictionary entries

English: {Word_eng_1}
Mambai: {Word_mgm_1}

English: ...
Mambai: ...

Please provide the translation for the
    following sentence. Do not provide
    any explanations or text apart from
    the translation.

English: {input}
Mambai:
```

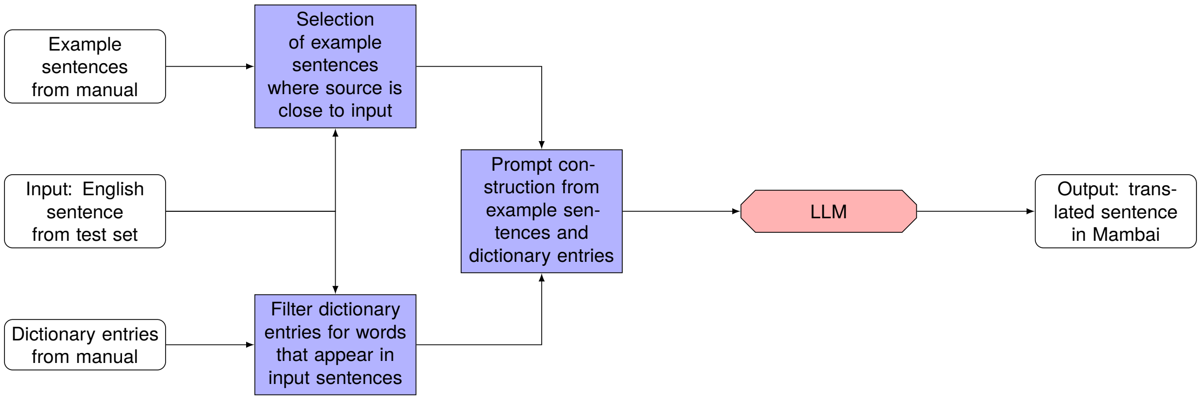


Figure 3: Overview of our process for translating English sentences to Mambai using both dictionary entries and sentence pairs in few-shot LLM prompting.

4.2.3. Models

We experiment with three models: **Mixtral** as it is the open-source model with the highest MT-bench score (Jiang et al., 2024), **LlaMa 70b** (Touvron et al., 2023) as it has a permissive license and has shown high zero-shot translation performance (Xu et al., 2024a), and **GPT-4**, which, despite being proprietary, has very high zero-shot translation performance (Xu et al., 2024a).

For each model, we experiment with the following setups:

- `UseDict` (either `True` or `False`): For each word that appears in the source language input (English), if this word is present in the English-Mambai dictionary, we include its dictionary translation in the prompt;
- N_{TFIDF}: Number of sentence pairs retrieved through TF-IDF, where the English sentences are ranked according to TF-IDF similarity to the input. The rationale here is that less frequent words can be harder to translate, therefore should be surfaced in the prompt more often. $N_{\text{TFIDF}} \in \{0, 5, 10\}$
- N_{embed}: Number of sentence pairs retrieved through LASER semantic embeddings (Touvron et al., 2023), where the English sentences in training set are first ranked using cosine similarity to the input. $N_{\text{embed}} \in \{0, 5, 10\}$, similar to Zhang et al., 2023a; Vilar et al., 2023; Hendy et al., 2023.

For each combination of the above features, we measure the BLEU and Chrf++ scores on both test sets, one from the language manual, and one manually translated by a native speaker.

4.3. Translation Results

Our experiment results for test sentences from the manual are provided in Table 1, and Table 2 provides the results for the test set collected from a native speaker.

To summarise, we make the following observations:

(1) Translation accuracy varies widely between both test sets. While we get an accuracy of up to 23.5 BLEU (41.9 ChrF++) for the test set that comes from the language manual, we could not reach a BLEU higher than 4.4 (33.1 ChrF++) for the test set from the native speaker. More analysis is needed to understand this discrepancy, but it sends a strong signal about the risks of overfitting by using a test set that comes from the same material as the examples used in prompting. In particular, we think our result might partially invalidate (Tanzer et al., 2024), which similarly attempts to translate into a very low-resource language using prompting from a single grammar book, but used exclusively sentences from the grammar book in the test set.

(2) Dictionary entries help improve translation quality. When including dictionary entries in the prompt, filtering on words that appear in the source text, we found that translation quality improved significantly. This is true across all experiments when keeping other hyperparameters constant, with an average improvement of 3.25 BLEU points and 2.7 ChrF++ points.

(3) A blend of sentences retrieved through semantic embeddings and through TF-IDF yields the highest translation accuracy. When working with a random split of sentences from the language manual in particular, a blend of 5 sentences retrieved through TF-IDF and 5 sentences retrieved through semantic embeddings outperforms 10 sentences retrieved exclusively through one of these features. This holds true for all three LLMs tested in this project.

(4) GPT-4 consistently outperforms other LLMs. GPT-4 yields both the highest translation score overall, and the higher translation score for every single experiment, when compared with LlaMa 70b and Mixtral 8x7B while keeping N_{TFIDF} and N_{embed} constant.

4.4. Error analysis

We find that the large gap in performance across test sets is mostly due to differences in translation output, rather than differences in the source English text (Table 3):

1. Using TF-IDF representations of English sentences, we computed the cosine similarity in the whole training set and the two tests sets, resulting in 0.021 for the manual test set and 0.017 for the native speaker test set, a relatively small difference. For the Mambai target reference, however, we get a 0.027 and 0.012 for the manual and native speaker's test sets, respectively, a much larger difference.

Model	N_{TFIDF}	N_{embed}	`UseDict`	**BLEU**	**ChrF**	**ChrF++**
gpt-4-turbo	0	0	FALSE	3.7	22.4	19.9
gpt-4-turbo	0	0	TRUE	6.9	25.3	24.7
gpt-4-turbo	10	0	FALSE	16.1	40.3	39.7
gpt-4-turbo	10	0	TRUE	20.9	41.8	41.6
gpt-4-turbo	0	10	FALSE	16.8	38.2	37.4
gpt-4-turbo	0	10	TRUE	18.3	39.6	39.5
gpt-4-turbo	5	5	FALSE	17.7	40.4	39.6
gpt-4-turbo	5	5	TRUE	**21.2**	**41.8**	**41.6**
Mixtral 8x7B	5	5	TRUE	9.0	30.9	30.4
LlaMa 70b	5	5	TRUE	12.3	32.3	31.8

Table 1: Experiment results for test set from the language manual. N_{TFIDF} and N_{embed} represent the number of sentence pairs retrieved through TF-IDF and semantic embeddings, respectively. `UseDict` indicates whether dictionary entries are included in the prompt. While different hyperparameter combinations were tested for all models, we only report on the best configuration for the less performant models (Mistral 8x7B and LlaMa 70b).

Model	N_{TFIDF}	N_{embed}	`UseDict`	**BLEU**	**ChrF**	**ChrF++**
gpt-4-turbo	0	0	TRUE	3	30.7	27.9
gpt-4-turbo	0	0	FALSE	0	30.8	26.9
gpt-4-turbo	10	0	TRUE	4	36.9	33.8
gpt-4-turbo	10	0	FALSE	0	33.4	29.9
gpt-4-turbo	0	10	TRUE	3.4	34.5	31.6
gpt-4-turbo	0	10	FALSE	0	31.4	27.8
gpt-4-turbo	5	5	TRUE	**4.4**	**35.9**	**33**
gpt-4-turbo	5	5	FALSE	0	33.7	29.9
Mixtral 8x7B	5	5	TRUE	3.5	26.8	24.6
LlaMa 70b	5	5	TRUE	0	27.7	24.7

Table 2: Experiment results for the minicorpus of translations collected from a native Mambai speaker. N_{TFIDF} and N_{embed} represent the number of sentence pairs retrieved through TF-IDF and semantic embeddings, respectively. `UseDict` indicates whether dictionary entries are included in the prompt. While different hyperparameter combinations were tested for all models, we only report on the best configuration for the less performant models (Mistral 8x7B and LlaMa 70b).

2. LASER Semantic similarity between each test set and the training set are roughly equivalent at 0.42 and 0.40 for the manual and native speaker's test sets, respectively, on the English source side.

Similarity	Lang	Method	Score
ManualTest x Train	eng	TF-IDF	0.021
NativeTest x Train	eng	TF-IDF	0.017
ManualTest x Train	mgm	TF-IDF	0.027
NativeTest x Train	mgm	TF-IDF	0.012
ManualTest x Train	eng	Semantic	0.42
NativeTest x Train	eng	Semantic	0.40

Table 3: Similarity scores using TF-IDF cosine similarity and LASER semantic cosine similarity between the two test sets and the training set for English (source, eng) and Mambai (target, mgm) sentences.

Through manual review of the translation differences in both test sets, we further identify the following potential causes for the large discrepancy in translation quality metrics:

(1) Literal vs figurative translation: As sentences in the language manual are made for learning, they tend to use more literal translations, which correspond to what LLMs produce. On the other hand, our test set translated by a native speaker often uses more idiosyncratic translation, further away from words used in from the source input.

(2) Language variation: The Mambai language has changed since 2001, when the Mambai Language Manual was published. In particular, we noted more usage of Portuguese and Tetun Dili words in our test set reference sentences, which might indicate that Mambai speakers mix more Tetun Dili and Portuguese in their Mambai since the two languages were chosen as official in the 2002 Constitution (Government of Timor-Leste, 2002).

(3) Spelling: Despite trying to stay close to spelling used in the Mambai Language Manual, we found that our test set at times uses different spelling than the language manual (e.g. less hyphenation, some letters missing). This reinforces our view that oral languages like Mambai are better covered by speech datasets.

5. Related Work

Traditionally, neural MT systems are trained on parallel corpora of aligned sentence pairs (Duong, 2017). Low-resource languages tend to have orders of magnitude less sentences available than higher-resource languages (Arivazhagan et al., 2019). To compensate for this lack of data, previous research found that low-resource MT accuracy can be improved through leveraging multilingual translation models that include better-resourced but related languages (Arivazhagan et al., 2019; Fan et al., 2020; Team et al., 2022). Other techniques include pretraining on monolingual data (Lample et al., 2018), the incorporation of audio data that shares an embedding space with text data (Communication et al., 2023), and the generation of synthetic parallel sentences (Edunov et al., 2018), including by leveraging bilingual dictionaries (Duan et al., 2020).

In parallel, large language models have shown increased ability to translate, at times surpassing specialised encoder-decoder MT systems (Robinson et al., 2023). Finding the right prompt recipe for increased MT accuracy using LLMs has been a topic of research (Zhang et al., 2023a; Li et al., 2022), with findings that few-shot prompting often improves MT accuracy (Zhang et al., 2023a), and that the type of sentences used as few-shot examples can have a large influence on accuracy (Moslem et al., 2023). Dynamic adaptation of the prompt by retrieving example sentences that are close to the input text (Kumar et al., 2023), or dictionary entries for words that appear in the source (Ghazvininejad et al., 2023) can further improve MT accuracy.

The applicability of common LLM prompting techniques when translating into very low-resource languages is unclear, given these languages might not be represented at all during LLM pretraining. Tanzer et al. (2024) partially addresses this issue by focusing on MT between English and Kalamang, an endangered Papuan language, using a single grammar book. Experimenting with different models (Claude 2, LlaMa, gpt-3.5, gpt-4), and different prompt setups (injecting sentences close to the input, dictionary entries, and the grammar explanations found in the book), they achieve up to 45.8 ChrF on the English to Kalamang direction. However, they work with a test set that is a random subset of sentences found in the book, raising issues around the applicability of their results to text translated by a different author, or to domains not covered in the grammar book.

Recognising the potential of LLMs for MT, and the importance of in-context examples used in prompting, our work experiments with retrieval-augmented LLM prompting for translation into a low-resource language. We test translation quality on both a subset of sentences coming from the language manual used as corpus, and a test set specially translated by a native Mambai speaker for this project.

Conclusion

In this paper, we introduced a novel corpus for the Mambai language, a language with around 200,000 native speakers that had virtually no NLP resources. Our corpus includes bilingual dictionaries in both directions for English-Mambai, a set of 1,187 parallel sentences from a language manual published in 2001, and a set of 50 parallel sentences translated by a native Mambai speaker. Our experiments on few-shot LLM prompting for English to Mambai translation showed that moderate MT quality can be achieved for test sentences very close to the original corpus, but MT quality decreases significantly for sentences that come from a separate corpus, thus highlighting the need for using test sets that do not come from the same material as original examples used in prompting. We think LLMs offer a flexible approach for integrating scarce resources in different formats (dictionary entries, parallel sentences), and few-shot prompting shows potential in improving low-resource MT using general purpose LLMs.

Limitations

The sentences used in both training set (from the Mambai Language Manual) and test sets tend to be rather short and simple, which raises questions around translation quality for longer sentences, or for technical domains that get little coverage in our corpus (e.g. health or legal text).

Mambai has no standard orthography. Even though the native Mambai speaker we collaborated with tried to follow spelling close to that used in the language manual, we expect that variances in spelling still negatively impacted the test BLEU score. This stresses the need for heightened focus on audio for primarily spoken languages like Mambai (Chrupała, 2023).

While we were able to gather a test set from a native Mambai speaker, they did not evaluate translation quality for MT-translated text; instead we relied solely on automated MT metrics. While BLEU tends to be a reliable measure of MT quality for morphologically simple languages like Mambai (Reiter, 2018), we would have preferred to dig deeper into the shortcomings of our LLM-generated translations.

Lastly, Mambai has a simple grammar and morphology, which might make it particularly prone to MT quality improvement using few-shot prompting. Therefore, our results might not translate well on more morphologically complex languages.

Future Work

This work focused solely on Mambai, without leveraging resources from related languages that have more resources, such as Tetun Dili, Portuguese, or Indonesian. In future work, we would like to investigate the addition of Tetun Dili sentences to the prompt, especially for domain-specific text that might be very poorly covered by our small Mambai corpus, but that could be covered by a larger Tetun Dili corpus.

In terms of finding the right recipe for prompting, future endeavours could use a more systematic approach, similar to Kumar et al. (2023) which uses a regression model for example selection. Additionally, more retrieval techniques could be tested, e.g. bag of words, or even ChrF similarity between the input and English source side.

In this paper, we used general purpose LLMs that likely saw little to no Mambai text during pretraining. We think future work could experiment with continuous pretraining on Mambai, or languages related to Mambai, before prompting, similar to approaches in Xu et al. (2024b) and Alves et al. (2024).

Acknowledgements

We thank Pr. Geoffrey Hull (Macquarie University), author of the Mambai Language Manual, for authorising usage of his work as part of this research. Pr. Hull remains the holder of copyright protecting this intellectual property.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Berlie. 2008. Notes on east timor: Languages and education. *Asian Journal of Social Science*, 36(3-4):629–637.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Grzegorz Chrupała. 2023. Putting natural in natural language processing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences.

Long Duong. 2017. Natural language processing for resource-poor languages. *University of Melbourne*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Helem Andressa de Oliveira Fogaça. 2013. *Estudo fonético e fonológico do Mambae de Same: uma língua de Timor-Leste*. Ph.D. thesis, University of Brasilia.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation.

Government of Timor-Leste. 2002. Constitution of the democratic republic of timor-leste.

Kirsty Sword Gusmão. 2023. The Key to Quality Inclusive Education in Timor-Leste's Third Decade as an Independent Nation. [Accessed 27-02-2024].

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Damien Kingsbury. 2010. National identity in timor-leste: challenges and opportunities. *South East Asia Research*, 18(1):133–159.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. Ctqscorer: Combining multiple features for in-context example selection for machine translation.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland. Association for Computational Linguistics.

Rahul Mehta and Vasudeva Varma. 2023. LLM-RM at SemEval-2023 task 2: Multilingual complex NER using XLM-RoBERTa. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 453–456, Toronto, Canada. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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)*. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Timor-Leste General Directorate of Statistics. 2015. 2015 population and housing census.

Timor-Leste General Directorate of Statistics. 2022. 2022 population and housing census.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. *Parallel corpora for medium density languages*, page 247–258. John Benjamins Publishing Company.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Dr. Stephen L. Walter. 2016. The embli endline evaluation study.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. 2021. Language models are good translators.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Language Resource References

Gowda, Thamme and Zhang, Zhao and Mattmann, Chris and May, Jonathan. 2021. *Many-to-English Machine Translation Tools, Data, and Pretrained Models*. Association for Computational Linguistics.

Hull, Geoffrey. 2001. *Mambai Language Manual: Ainaro Dialect*. Sebastião Aparício da Silva Project.

NLLB Team and Marta R. Costa-jussà and James Cross and Onur Çelebi and Maha Elbayad and Kenneth Heafield and Kevin Heffernan and Elahe Kalbassi and Janice Lam and Daniel Licht and Jean Maillard and Anna Sun and Skyler Wang and Guillaume Wenzek and Al Youngblood and Bapi Akula and Loic Barrault and Gabriel Mejia Gonzalez and Prangthip Hansanti and John Hoffman and Semarley Jarrett and Kaushik Ram Sadagopan and Dirk Rowe and Shannon Spruit and Chau Tran and Pierre Andrews and Necip Fazil Ayan and Shruti Bhosale and Sergey Edunov and Angela Fan and Cynthia Gao and Vedanuj Goswami and Francisco Guzmán and Philipp Koehn and Alexandre Mourachko and Christophe Ropers and Safiyyah Saleem and Holger Schwenk and Jeff Wang. 2022. *No Language Left Behind: Scaling Human-Centered Machine Translation*.

Improved Neural Word Segmentation for Standard Tibetan

Collin Brown
Indiana University
coljbrow@iu.edu

Abstract

As Tibetan is traditionally not written with word delimiters, various means of word segmentation are necessary to prepare data for downstream tasks. Neural word segmentation has proven a successful means of parsing Tibetan text, but current performance lags behind that of neural word segmenters in other languages, such as Chinese or Japanese, and even behind languages with relatively similar orthographic structures, such as Vietnamese or Thai. We apply methods that have proven useful for these latter two languages toward the development of a neural word segmenter with the goal of raising the peak performance of Tibetan neural word segmentation to a level comparable to that reached for orthographically similar languages.

Keywords: Tibetan, Word Segmentation

1. Introduction

Tibetan is a language—or rather, a number of languages and dialects of varying degrees of mutual-intelligibility—spoken in Tibet, a region overlapping a number of provinces in modern-day China including the Tibetan Autonomous Region, Sichuan, and Qinghai. Diaspora communities reside also in India, Nepal, and Bhutan; and a substantial, if smaller, number live also in Switzerland, Canada, the United Kingdom, and the United States (among many other countries).

Tibetan belongs to the Sino-Tibetan language family and is traditionally placed in the Tibeto-Burman branch, though the phylogeny of the family remains hotly contested. The Tibetan family can be further divided into various dialect and language groups, including Central (or Ü-Tsang) with approximately 1.2 million speakers, Amdo with 2.5 million, and Khams with 2 million, among others (Eberhard et al., 2024). However, Standard Tibetan generally serves as a lingua franca among them; thus, expanding the resources available to the language provides benefits not only to native speakers but also to the broader Tibetan community, whatever their regional or dialectal background. By improving word segmentation for Tibetan, we hope to facilitate the creation of further tools—word prediction, sentiment analysis, etc—which might make the language easier for its speakers to use in the digital domain, easing linguistic pressures that motivate them to switch to languages with more support, such as English, Mandarin Chinese, or Hindi.

Many Asian scripts are not written with spaces between words, and this obviously presents certain problems when one wishes to engage in most computational tasks, the models for which tend to operate on words rather than characters. Standard Chinese, another such space-less language, has the benefit of using characters, each of which are semantically heavy; character-embeddings allow for the training of highly accurate models. While some languages, such as Korean, have broadly adopted the practice of placing spaces between words, many orthographies descended from either the Indic or Sinitic traditions continue to go without them. Furthermore, some languages make use of orthographic features that make word segmentation an easier task; Japanese in particular uses multiple different scripts, and the transitions between these often serve as strong indicators of word boundaries. However, we do not have such luxuries with Tibetan which only explicitly marks syllable and sentence boundaries. While researchers have reached an accuracy of upwards of 98% for Japanese texts (Kitagawa and Komachi, 2018) and 97% for Chinese (Cai et al., 2017), Tibetan lags behind. Duanzhu et al (2020)

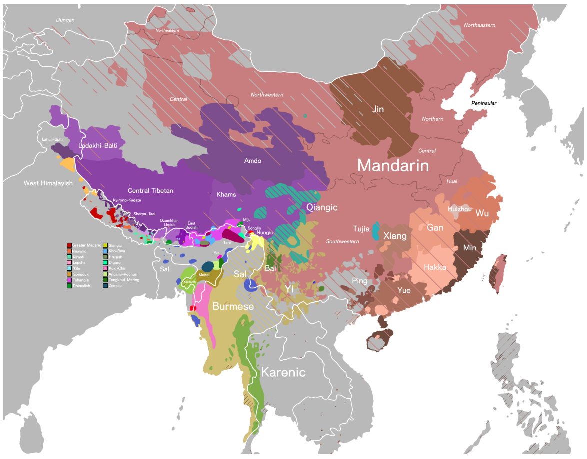


Figure 1: The Tibetic family appears here in shades of purple in the northwest quadrant. GalaxMaps, CC BY-SA 4.0, via Wikimedia Commons.

2nd EURALI Workshop, pages 12–17

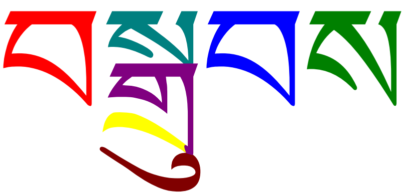
report a binary accuracy of 93.4% with an f1-score of 94.11% and a recall of 94.2%; Wang & Yang (2018), a f1-score of 94.1% and a recall of 93.89%; and Li et al (2022) an f1-score of 92.31% (Duanzhu et al., 2021; Wang and Yang, 2018; Li et al., 2022).

While the success of Chinese and Japanese can to some degree be attributed to the vast resources available for these two languages, neural word segmentation research for smaller (though by no means, small) languages such as Vietnamese and Thai have reached an accuracy of around 96% or higher (Zheng and Zheng, 2022).

2. Background

Phonetically, the Tibetan syllable is of only moderate complexity, but the language's standard orthography preserves the highly complex syllable structures of the ninth and tenth centuries. While spoken syllables in Lhasa Tibetan may begin and end with at most a single consonant respectively, they may be written with upwards of four initial consonants and two final consonants. Furthermore, vowels are not written as distinct letters but instead added as diacritics above the "head" letter, or the letter whose phonetic value serves as the basis for the onset of the syllable.

The maximally complex Tibetan syllable is composed of a prescript letter, a head letter, a postscript and a post-postscript letter. All but the head letter are (usually) composed of a single, simple letter, but the head letter can itself be composed of a superscript, a root, upwards of two subscripts, and a vowel diacritic.



བ (b), ས (s), ག (g), ར (r), ◌ུ (u), བ (b), and ས (s)

Figure 2: A maximally complex Tibetan syllable. The past tense form of the word, སྒྲུབ (sgrub), meaning "to complete".

As can be seen in figure 2, the unique complexity of Tibetan syllables allows them to carry a relatively high degree of semantic value; thus, they can serve as stronger indicators of word boundaries than syllables in more orthographically shallow languages.

There are thirty standard letters that may serve as the root of the head of a syllable. Onto these, four subscripts may be attached—◌ྱ (y), ◌ྲ (r), ◌ླ (l), ◌ྭ (w)—but only certain head-subscript combinations are allowed. In addition, ◌ྲ (r) and ◌ྭ (w) may appear together on the same root, meaning that the total number of root-subscript combinations comes out to fifty-five unique arrangements. Onto these, one may attach three superscripts—ར (r), ལ (l), ས (s)—and again these are only allowed in particular arrangements, meaning that the total number of head letters which feature a unique superscript, root, and subscript is only thirteen. Adding these, as well as the unique superscript-root combinations, to our running total gives us one-hundred-and-one unique head letters. Each of these may take up to one vowel diacritic, of which there are four, yielding six-hundred-and-six unique head letters. These diacritics are ◌ི (i), ◌ུ (u), ◌ེ (e), and ◌ོ (o).

Onto these, one may add some combination of prescript and postscript letters. There are ten postscript letters and two post-postscript letters (though really only one, as the other has been dropped in most writing). The post-postscript letter may only appear after four of the postscript letters, meaning there are a total of fourteen possible postscript combinations. Confusingly, one of these postscript letters, འ may also carry a vowel diacritic, though it is usually limited to འི (i) or འུ (u); however, འེ (e) and འོ (o) do appear, albeit rarely. Thus, we have eighteen possible postscript combinations.

Despite there being a maximum of one prescript letter, calculating the number of possible combinations is less straightforward given that there are more restrictions on which letters may appear in certain positions. There are five prefixes—ག (g), ད (d), བ (b), མ (m), and འ (a)—and calculating the total number of unique prescript-head letter combinations created by them is quite difficult given their distribution. Disregarding super- and sub-scripts, as well as vowels, there are a total of fifty-three unique prescript-head letter combinations. If we are liberal with our estimates, we would say that the number of unique prescript-head letter combinations (including all our superscript, subscript, and vowel combinations) comes out to around three thousand unique combinations. Assuming that many of these do not appear in the actual written language, we might lower this down to only a couple thousand unique combinations, onto which we would then necessarily add our various postscript letters, bringing out estimated total number of unique syllables into the tens of thousands.

This number is highly misleading, as we find out when we compile a dictionary of all the syllables that appear in any particular Tibetan corpus. The true number of unique syllables to be found in actual texts is considerably lower, usually in the sub-ten-thousand range, and if we filter out those that appear less than five times—as we do in our model—we arrive at much more modest numbers, usually between two and five thousand unique syllables, depending on the size and variety of the corpus.

In any case, the semantic load of the Tibetan syllable, as well as the fact that the vast majority of word boundaries are also syllable boundaries, allows us to use syllable embeddings as a heuristic by which to train our model.

In standard, written text, all syllables are delimited by a unique punctuation mark known as the *tseg*, written '་'. This allows us to easily parse through a text and separate out each syllable, whereas many other languages that make use of syllable embeddings for neural word segmentation—such as Thai or Khmer—must engage in more complex syllable-identifying methods beforehand. While certain questions do arise about what constitutes a word-boundary in Tibetan, for our purposes we may treat word boundaries as a subset of syllable boundaries. Specifically, the genitive and agentive cases sometimes take the form of postscripts on the final syllable of words ending in vowels; in the spoken language, they are realized via umlaut or lengthening of this final vowel, and so we will treat them as part of the word rather than separate particles.

3. Corpus

While the most extensive corpora available for Tibetan are limited to Classical Tibetan, we were able to make use of the UVA Tibetan Spoken Corpus (Germano et al., 2017) which, while a couple decades old, represents the most easily accessible corpus of pre-segmented text available. This corpus was compiled by the *Tibetan and Himalayan Digital Library* project which is affiliated with the University of Virginia and reflects the colloquial language of people living in Tibet rather than the highly formal, literary language often found in religious and official texts. As we intend to apply this word segmentation model towards the development of tools aimed at making Tibetan more accessible in the digital realm, it was important that the corpus reflect the kinds of language used by everyday people.

The corpus indicates word boundaries with a space, meaning that with minimal processing we can clean the corpus of non-Tibetan text, and divide it into syllables, storing each with some indication as to whether or not it is word-final. With 807,033 total syllables, we can take those which appear with a frequency of at least 5, resulting in 2584 embeddings. In total, about 18.52% of the syllables in this corpus are non-word-final (meaning they don't mark a word boundary). This is quite a bit smaller than the corpus used to train the AttaCut model, from Chormai et al (Chormai et al., 2019), which featured 2.56 million syllables. Similarly, whereas Duanzhu et al (2020) employ a corpus with 160,000 sentences, ours features only 70,000 (Duanzhu et al., 2021).

Unsurprisingly, the most common syllables found in the corpus include པ, an incredibly common nominalizer and derivational suffix; ལ, the oblique case particle; དང, a conjunction and comitative / associative particle; དེ, the medial demonstrative; and the various case and TAM endings and particles that compose Tibetan's robust nominal and verbal systems.

4. Methods

A somewhat recent and effective method for neural segmentation of Thai words is the use of syllable embeddings as input features. Training a neural network to identify word boundaries based on syllable embeddings rather than characters has proven quite effective, as evidenced by the AttaCut model developed by Chormai et al (Chormai et al., 2019). Because Tibetan explicitly marks syllable boundaries, and because of the orthographic depth of the language—with a spelling system that preserves pronunciations from antiquity—we determined that it might be particularly useful in improving performance.

In a manner similar to the AttaCut model, Nguyen 2019 makes use of BiLSTM to generate embeddings for the syllables present in a Vietnamese corpus and uses these to train a model to determine word-boundaries with a 98% accuracy. However, this model's success comes in part due to its use of a rule-based word segmenter, RDRsegmenter, as outlined in (Nguyen et al., 2018), in conjunction with its neural method. RDRsegmenter produces a set of word-boundary tags whose embeddings are concatenated with each syllable's embedding to produce those that are used to train the final model (Nguyen, 2019; Nguyen et al., 2018). Similar methods have proven effective for Chinese since each character corresponds, with some exceptions, to one syllable and one morpheme (Qian and Liu, 2012). It should be the case that Tibetan, which adheres less strictly to this one-syllable-one-morpheme structure, can still benefit from the application of this method.

Some combination of syllable embeddings, character embeddings, and word-boundary embeddings generated by a rule-based model have proven useful for word segmentation in many of the languages of East and Southeast Asia, including Khmer, Chinese, and even Classical Tibetan, for which there exists more readily available corpora owing to the digitization of many Buddhists texts (Buoy et al., 2020). Given the intense conservatism of Tibetan orthography it may be possible to supplement a corpus of modern, standard Tibetan with texts from Classical Tibetan; however, initial tests yielded no benefits. More research is required to determine if this is a viable route for improvement.

A last note worth considering is the presence of non-standard text within the corpus. Certain sequences, especially in older texts, lack the syllable delimiter, complicating the pre-processing necessary for our model. In future research, it may be worth considering the implementation of a syllable segmenter which would insert syllable boundaries where they are stylistically omitted from text (or left out in error). Furthermore, it would be necessary to operate on a sub-syllable basis if one wished to separate certain instances of various cases which modify words at such a level. For example, when a noun's final syllable lacks any post-script letter, the genitive case takes the form, -འི (-i), which is given no special treatment here but which may, in other applications, necessitate further delimiting.

5. Implementation

By generating syllable embeddings via a Word2Vec model, we are able to train a model to predict the probability that a given syllable—the center of our context window—is non-word-final. Word2Vec is used in order to ensure manageable model size. We limited our syllable embeddings to only those syllables which appeared at least five times in the corpus; this may present an issue when our model is faced with a much larger corpus with many more unique syllables, some of which may appear semi-frequently, as well as when presented with regional or alternative spellings. To help with this and to push our performance past 96% accuracy, it may prove useful to implement a rule-based segmenter as done in Nguyen 2019, whose predictions should improve the accuracy of our neural segmenter when coupled—or rather concatenated—with our syllable embeddings. Furthermore, we hope to find a larger corpus on which to train our model in order to reduce the number of out-of-vocabulary syllables our model must cope with.

Early attempts at Tibetan word segmentation drew on MaxMatch algorithms and rudimentary statistical models, but with the proliferation of neural networks throughout natural language processing, the task has largely adopted such methods.

Hyperparameters	
Embedding Size	400
Learning Rate	1e-5
No. Layers	5
Window Size	3
Batch Size	64
Epochs	20
Accuracy:	96.87%

Table 1: The above hyperparameters are shown to approach those optimal for the model.

Drawing on a similar method to Liu et al (2015), another implementation of neural Tibetan word segmentation, we implement our model as a binary decision task, with the model labelling each syllable as either word-final or non-word-final (Liu et al., 2015). Syllable delimiters and any word delimiters are removed. Unlike Duanzhu et al (2021), we treat each syllable as an irreducible unit; they implement character embeddings in addition to syllable embeddings, which proves useful for certain purposes such as morphological analysis but introduces more opportunities for error (Duanzhu et al., 2021). We opt for a more straightforward model, considering only discrete syllables within a context window, and maintain a simple binary output rather than the more complex tag sets used in some implementations, such as Liu et al (2015) and Wang & Yang (2018) (Liu et al., 2015; Wang and Yang, 2018).

Currently, the vast majority of corpora are available not in Modern Tibetan but instead in Classical Tibetan, due to the many Buddhist texts that have been digitized from that period. Compiling a larger corpus in Modern Tibetan would provide our model with more data and reduce the instances of out-of-dictionary syllables. Initial tests involving the training of a model on The Annotated Corpus of Classical Tibetan (ACTib) (Meleen and Roux, 2020), followed by fine-tuning on a Standard Tibetan corpus, proved unsuccessful in yielding benefits compared to training solely on Standard Tibetan.

As we can see from figure 3, binary accuracy is not improved by expansion of the window size beyond one syllable on either side of the target. We might have assumed that a broader window would allow the model to differentiate between the occurrence of certain common syllables in various contexts, especially in words with more than two or three syllables, but this does not seem to be the case. Rather, as figure 4 reveals, performance is much more contingent on embedding size. This is somewhat unsurprising; in a language such as Tibetan where such units often correspond with morphemes, much meaning may be packed in.

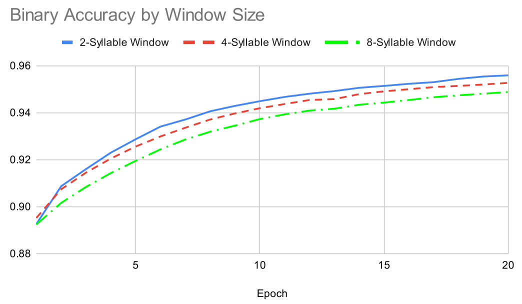


Figure 3: The window size of the model yields the best performance when only accounting for one syllable on either side of the target. Greater widths yield worse results overall, indicating that a local, relatively simplistic system can account for most word boundaries.

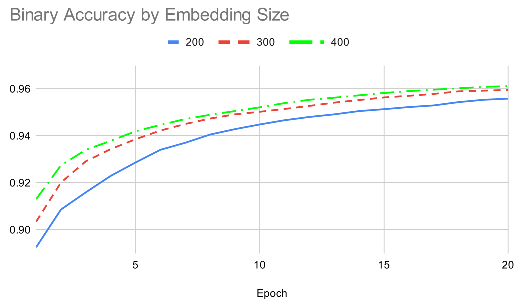


Figure 4: Larger vector sizes yields greater performance, but this diminishes above a value of 300. With model size in mind, we determine that values above 400 do not yield sufficient returns.

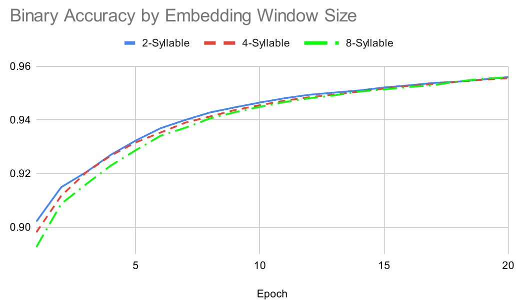


Figure 5: By changing the window we use when training our embeddings, we find a similar, albeit less pronounced, effect as with the model's window. Here, a 3-syllable window is considered optimal.

6. Results

While unable to achieve a level of performance found in more resource-rich languages, we are able to match that of languages with similar orthographic traditions. Furthermore, our findings indicate a great margin for improvement via the application of rule-based heuristics and larger corpus sizes.

If it were the case that the model primarily considered the target syllable, we would expect a 1-syllable window to outperform the 3-syllable window; however, this 3-syllable window outperforms both the 1- and 5-syllable window (and any greater number), indicating that it is the immediate, local context (and not any more-distant relation) that can account for most word boundaries. Augmenting this local window with more information (such as a rule-based heuristic) may yield further benefits.

The relatively great impact of embedding size does reflect the semantic weight of Tibetan syllables; their complexity provides information about word boundaries not contained in the orthographic units of even structurally similar languages. This is undoubtedly influenced as well by the syllable-to-morpheme ratio of Tibetan which (like many neighboring languages) tends to approach 1.0.

7. Conclusion

By implementing methods that have proven successful for neural word segmentation in orthographically similar languages, such as Thai and Vietnamese, we have been able to achieve a level of performance approaching the most performant word segmenters for Standard Tibetan, though further exploration may yield enough improvements so as to surpass the current peak performance and bring Tibetan word segmentation up to a comparable level as has been achieved for these other languages. Currently, we are limited by the availability of large corpora in Standard Tibetan; the acquisition of more data in addition to the refinement of existing methods and their augmentation with novel heuristics would, in benefitting neural word segmentation, provide downstream benefits for all varieties of natural language processing tasks.

8. Bibliographical References

Rina Buoy, Nguonly Taing, and Sokchea Kor. 2020. Khmer word segmentation using bilstm networks.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 608–615, Vancouver, Canada. Association for Computational Linguistics.

Pattarawat Chormai, Ponrawee Prasertsom, and Attapol Rutherford. 2019. Attacut: A fast and accurate neural thai word segmenter.

Sangjie Duanzhu, Cizhen Jiacuo, and Cairang Jia. 2021. Revisiting tibetan word segmentation with neural networks. In *Chinese Lexical Semantics*, pages 515–524, Cham. Springer International Publishing.

David Eberhard, Gary Simons, and Charles Fennig. 2024. Ethnologue: Languages of the world.

David Germano, Edward Garrett, and Stephen Weinberger. 2017. Uva tibetan spoken corpus.

Yoshiaki Kitagawa and Mamoru Komachi. 2018. Long short-term memory for Japanese word segmentation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Yan Li, Xiaomin Li, Yiru Wang, Hui Lv, Fenfang Li, and La Duo. 2022. Character-based joint word segmentation and part-of-speech tagging for tibetan based on deep learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Huidan Liu, Congjun Long, Minghua Nuo, and Jian Wu. 2015. Tibetan word segmentation as subsyllable tagging with syllable's part-of-speech property. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 189–201, Cham. Springer International Publishing.

Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He. 2011. Tibetan word segmentation as syllable tagging using conditional random field. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 168–177, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Marieke Meleen and Élie Roux. 2020. The annotated corpus of classical tibetan (actib).

Dat Quoc Nguyen. 2019. A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 28–34, Sydney, Australia. Australasian Language Technology Association.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A fast and accurate Vietnamese word segmenter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xian Qian and Yang Liu. 2012. Joint chinese word segmentation, pos tagging and parsing.

Lili Wang and Hongwu Yang. 2018. Tibetan word segmentation method based on bilstm$_crfmodel$. In *2018 International Conference on Asian Language Processing (IALP)*, pages 297–302.

Kexiao Zheng and Wenkui Zheng. 2022. Deep neural networks algorithm for vietnamese word segmentation. *Scientific Programming*, (8187680).

Open Text Collections as a Resource for Doing NLP with Eurasian Languages

Sebastian Nordhoff, Christian Döhler, Mandana Seyfeddinipur
Berlin-Brandenburg Academy of Sciences and Humanities
nordhoff@bbaw.de, doehler@bbaw.de, seyfeddinipur@bbaw.de

Abstract

The Open Text Collections project establishes a high-quality publication channel for interlinear glossed text from endangered languages. Text collections are made available in an open interoperable format and as a more traditional book publication. The project addresses a variety of audiences, eg. community members, typological linguists, anthropologists, and NLP practitioners.

Keywords: text collection, interlinear glossed text, language resources

1. Introduction

Franz Boas established the "Boasian Trilogy" in language documentation and description (Himmelmann, 1998), consisting of a grammatical description, a dictionary, and a text collection. All three levels of description are necessary to get a comprehensive overview of a language, and more importantly they complement each other. Linguists working in any field will often find themselves going back and forth between all three components. While we have good outlets for grammars (eg. Comprehensive Grammar Library[1]) and dictionaries (eg. Dictionaria[2]), such is not the case for text collections. This means that only few of them are published, and even fewer follow the FAIR principles of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016).

The project Open Text Collections (henceforth OTC)[3] remedies this by making high quality text collections from endangered languages available in an open interoperable format. Next to providing pdfs and/or printed books to researchers and to the language communities themselves, this setup makes the data available in CLDF format (Forkel et al., 2018) for downstream use in NLP applications.

Most reference grammars published today are the result of a language documentation project, often part of authors' dissertation projects. These grammars should be data-driven and accompanied by a corpus in order to facilitate the verification or falsification of the analysis (Mosel, 2012). While countless hours are invested into the structuring and glossing of texts, in many cases, however, these texts are not made available in a reusable way. Linguists tend to have them somewhere on their hard drive, or uploaded to an archive, but there is no generally established way of publishing them, at least not in a format which would feed further research downstream (eg. linguistic typology, corpus-based language description, or NLP). This means that these valuable results of language documentation often fail to be discovered.

OTC establishes a quality venue for publishing text collections, following the setup created by Language Science Press. The platform is community-driven and aims at being attractive to both data producers (ie. language documenters) as well as data users (ie. language communities, typologists, NLP practitioners). For data producers, the platform sets up guidelines for quality control, rigorous peer review, and top-notch publishing (pdf and print-on-demand), making sure that the time invested in a text collection will not harm job prospects. For data consumers, different output format are available to suit different needs: printed books without interlinearization for the language communities; pdfs/books with interlinearization and a search interface for typologists (prototype available at `https://imtvault.org`), and all the data in CLDF format for NLP practitioners. By making reuse easy, the research output will spread more widely, which in turn is very attractive for the data producers.

As of today, there are 5 regional boards and 45 proposed text collections. This paper showcases the platform, its motivations, and its benefits for data producers and consumers.

2. Content coverage

Text collections are an old publication format, which has its origin in history, human geography, and social anthropology. In modern linguistics, the study of texts has given rise to entire subfields, for example corpus linguistics, and it is now standard prac-

[1] `https://langsci-press.org/catalog/series/cogl`
[2] `https://dictionaria.clld.org`
[3] `https://opentextcollections.github.io/`

2nd EURALI Workshop, pages 18–23

tice to add a few sample texts to grammatical descriptions. In some cases, grammar authors have published collections of texts as separate monographs in book form. For example, Jeffrey Heath's descriptive trilogy of the Australian language Nunggubuyu consists of a text collection (1980), a dictionary (1982), and a grammar (1984).

But what is the difference between a text corpus and a text collection? What is the difference between an archive deposit and a text collection? A language corpus of one of the major languages is technologically way more advanced than what is feasible for low-resource languages, where, very often, there is only one researcher working on a language. Moreover, corpus linguistics aims for representativeness, for a broad coverage of different criteria: genre, spoken or written style, topic, speaker background. This sets the bar too high for a language documentation project. On the other hand, an archive collection from a documentation project generally has a focus on natural, unedited, spoken language. It includes audio-visual recordings of speech events of various genres. For the OTC project, we endorse a notion of text as "written oral literature".

Moreover, archives tend to have a kind of "Russian doll" structure (Evans and Dench, 2006, 25) with a small core of well-analysed material, a medium number of translated texts in the middle and a huge amount of raw data with no significant transcription or translation at the outside. This small core of well-analysed texts potentially falls within the scope of the OTC project, but the archives in their entirety have a much larger scope.

The OTC project is located between corpora and archive collections, and the intended output differs from both in various ways. Therefore, the project has to find its own definition of "text collection". To this end, we have defined the following criteria to gauge submissions:

Curation: The submission has made a careful selection of texts from a language (eg. from a documentation project) and provides them as a coherent whole. A text collection may be structured by variety, topic or genre. This is different from a full corpus or a deposit in a language archive, in that selectivity and content coherence are ranked higher than quantity and representativeness.

Contextualization: The submission has a prose introduction, which gives geographical, anthropological, historical and linguistic context. This includes an introduction to the speech community, the language, the recording methods, the individual narrators, etc. Contextualization should go beyond the metadata as can be found in a language archive. Such contextualisation gives full credit to the original authors (narrators/speakers) because, after all, these texts are much more than just data points. Moreover, contextualization is demanded by researchers from many fields, for example anthropology, oral history, sociolinguistics or comparative narratology.

Ethics: The submission ensures that as much input is collected from key stakeholders as possible, especially on the topics of cultural sensitivities, access control, publishing licenses, and intellectual property. In most cases, the researcher submitting a text collection to OTC will consult the language community and/or the individual speakers on these points, but in cases of legacy material this can include the heirs of the speakers, or the heirs of the collector.

Editing: The submission has adapted the source material to be understandable outside of the immediate context (time and place) of narration, and the changes applied to the original source are documented and justified. Contributors may choose to edit out false starts, pauses, self-corrections, etc., but the criteria for doing so should be stated explicitly. OTC endorses a notion of "text" that is closer to "written oral literature" than to the close transcriptions that are useful for detailed analysis of speech phenomena.

Transparency: The submission has good provenance, which includes well-structured metadata, but also links to the original recordings deposited in an archive or scans in the case of legacy material. Furthermore, all decisions and steps in the editing process are documented.

Accessibility: The text collection will be available under an open and interoperable format following the FAIR standards of findability, accessibility, interoperability, and reusability.

Glossed: The submission has been fully interlinearized and glossed, following the Leipzig Glossing Rules.

3. Social setup

OTC is based on a bottom-up, scholar-led, community-driven structure. The platform is provided by the Berlin-Brandenburg Academy of Sciences and Humanities in co-operation with the publishing structure of Language Science Press.

Interested researchers can form a regional board to cover a given area. Currently, there are 5 such areas (Africa, Caucasus, Eurasia, Papunesia, South

Figure 1: Expressions of interest for languages of Eurasia in OTC

Phylum	# Languages
Burmo-Qiangic	1
Indo-Aryan	2
Iranian	1
Macro-Tani	1
Nakh-Dagestanian	7
Tai-Kadai	1
Uralic	3
total	16

Table 1: Expressions of interest per phylum, Eurasia only.

America). The regional boards organise a rigorous peer review process that ensures high-quality results. Peer review is organised as a two-step process. An initial proposal will contain the linguistic, anthropological and philological context, accompanied by one sample file. The proposal is peer evaluated by the regional editors. If the proposed is judged positively on merits of focus, coherence, adequacy, ethics and technical quality of the sample file, the compiler is invited to submit the full collection. The full collection will undergo peer review, with one text being selected for in-depth review, while from the remaining texts, only a subset of randomly drawn sentences will be highlighted for review. This ensures both depth and breadth of reviewing without overburdening the reviewers. Text collections can number several hundred pages, which would be very time-consuming to review one by one. Consistency and adherence to guidelines will be checked computationally.

More areas, or regional boards, than the initial five can be added, but have to undergo vetting by the existing regional boards. It is envisioned, for instance, to split the rather large area of "Eurasia" into several subareas of a more manageable size. Figure 1 gives an overview of the collections which have been proposed to OTC for the languages of Eurasia.

4. Geographical coverage

For languages of Eurasia, the relevant regional boards are OTC Caucasus and OTC Eurasia. At the time of writing, there are 16 collections which have been proposed to OTC, whose affiliation is given in Table 1.

5. Technology

OTC can ingest several types of file formats commonly used in language documentation formats. These are converted to a common backend in CLDF format, from which a variety of output formats can be generated.

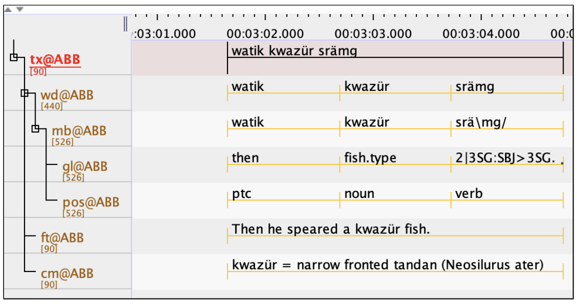


Figure 2: The sample text in ELAN

5.1. Ingestion

There are a number of different language documentation projects, which typically submit their work to one of the DELAMAN[4] archives, eg. AILLA,[5] ELAR,[6] PARADISEC[7] or TLA.[8] The most commonly used programs to produce interlinear glossed text (IGT) are ELAN and FLEx.

ELAN is a program, shown in Figure 2, which allows users to annotate multimedia on different "tiers" (Wittenburg et al., 2006). Different speakers will have different tiers, and tiers can be of different types, eg. transcription, translation, and glosses. Relations between tiers are explicit. Users have a lot of freedom about which tiers to define and what features to assign to them, leading to a vast heterogeneity of tier types (von Prince and Nordhoff, 2020; Nordhoff, 2020). ELAN uses an XML format as its backend. The library *eldpy* reads ELAN files and applies a number of heuristics to find the most probable tiers for transcription, translations, glosses. Criteria evaluated are: the name of the

[4]https://www.delaman.org
[5]https://ailla.utexas.org/
[6]https://www.elararchive.org
[7]https://www.paradisec.org.au
[8]https://archive.mpi.nl/tla

1	ID	Primary_Text	Analyzed_Word	Gloss	Translated_Text
42	a2279	watik kwazür srämg	watik kwazür srä\mg/	then fish.type 2|3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/shoot	Then he speared a kwazür fish.
43	a2281	kwazür ysme nge fäth srärmir	kwazür ys =me nge fäth srä\rmir/	fish.type spike INS child DIM 2|3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/p	With the Kwazür spines
44	a2283	etha ys kwazür ane mane yaththgr ane ysme	etha ys kwazür ane mane ya\thth/gr ane ys =me y	three spike fish.type DEM which 3SG.MASC:IO:NPST:STAT/be.stickin	With those three spines on the kwazür

Figure 3: The sample text in CLDF (.csv) format.

tier ('ft' is typically indicative of "free translation", 'ge' is "gloss english" etc), the relation to other tiers ("symbolic association" is either a translation or a gloss), and the language of the tier (translation tiers should pass a language detection test for English; transcription tiers should fail such a test). Based on these criteria, content is extracted and stored as the CLDF fields "Primary_Data", "Analyzed_Text", "Glosses", and "Translation" (see Figure 3).

FLEx is another program which is often used in language documentation projects. It allows the linguist to tokenize and gloss a transcribed text with the help of a lexicon. The lexicon grows as more and more texts are ingested. FLEx also uses an XML backend. The CLDF library *cldflex* (Matter, 2024) can be used to extract the relevant content and store it as CLDF. By and large, FLEx shows a lot less heterogeneity than ELAN.

tex and xlsx are other formats which are structured enough to provide import routines. The *langsci-gb4e* package for the LaTeX typesetting language is commonly used in grammar writing, and the content can easily be extracted with *linglit*, as has been shown for IMTVault (https://imtvault.org). These two latter formats are less prevalent than ELAN or FLEx, but still frequent enough to warrant import routines.

5.2. Backend

OTC stores the interlinear glossed text in the Cross-Linguistic Data format (CLDF, [9] (Forkel et al., 2018), Figure 3), a format which is an emerging standard for research data in linguistic typology and beyond and which can easily be ingested into CLLD (cross-linguistic linked data) applications. CLDF provides several components, of which the component "examples"[10] is the most pertinent for OTC. The relevant columns are Primary_Data, Analyzed_Text, Glosses, and Translation, complemented by a column for Glottocode,[11] and a column for comments. The CLDF format is extensible, meaning that additional columns can easily be added, but no promise is made that the content therein can be consumed.

[9] https://cldf.clld.org
[10] https://github.com/cldf/cldf/tree/master/components/examples
[11] https://glottolog.org

The creation and refinement of the text collection is done on GitHub, with releases being automatically archived on Zenodo[12] using the GitHub-Zendodo bridge.

5.3. Output formats

There are three main target groups for OTC content: NLP practitioners, linguists, and speaker communities. For NLP practitioners, a csv dump is made available (cf. Figure 3), next to a rendering in JSON-LD. Linguists can use the csv dump for quantitative research or an ElasticSearch HTML frontend for qualitative explorations, based on work done for IMTVault (Nordhoff and Krämer, 2022). The text is also made available as a pdf with interlinearized examples (Figure 4). Language communities finally can use the pdfs generated from the backend with a two-column layout with vernacular on the left and translation on the right (Figure 5). Both pdf formats are fed into the print-on-demand pipelines established by Language Science Press. These printed books are then available world wide via the usual distribution channels (eg Amazon, local bookstores, Verzeichnis lieferbarer Bücher etc.)

6. Use downstream

A number of recent studies have shown the usefulness of well-structured textual data for NLP approaches. Most of them focus on ways to overcome bottlenecks in the production of IGT, for example segmentation and glossing (McMillan-Major 2020, Barriga Martínez et al. 2021, Liu et al. 2021, Moeller and Hulden 2021). Two example studies of NLP approaches are explained in more detail here.

[12] https://zenodo.org/communities/otc/records?q=&l=list&p=1&s=10&sort=newest

(41) *kukufia mane sfrärm kofär ane gäwkarä sukogrm*
kukufia mane sf\rä/rm kofä =r ane gäw =karä
PN which 3SG.MASC:SBJ:PST:DUR/be fish PURP DEM fish.spear PROP
su\ko/grm
3SG.MASC:SBJ:PST:DUR:STAT/be.standing
'Kukufia stood there at the front with his harpoon, looking for fish.'

(42) *watik kwazür srämg*
watik kwazür srä\mg/
then fish.type 2 | 3SG:SBJ>3SG.MASC:OBJ:IRR:PFV/shoot
'Then he speared a kwazür fish.'[11]

[11] kwazür = narrow fronted tandan (Neosilurus ater)

Figure 4: The sample text in "scientific" format

8.3 *Kukufia*

Kukufia is a narrative lasting roughly 6min. It was the first recording made during the documentation project. It was recorded by Christian Döhler on September 5th 2010 only in audio format. The speaker is Abia Bai, and the recording took place inside his house at Rouku.

nzone yf rä abia. nzä worä rokuma. nzone ŋafe bäi. trikasi bänema kwa natrikwé. kabe tnz yf sfrärm kukufia. kukufia mane yara masun swamnzrm. nafane ŋare edawä. nä kayé kabe zä swamnzrm we rokun. näbi ŋarekarä fi sfrärm. fi zefara bi farsir. karesa zfth kar yf rä. watik karesa zfthen fi bä bsfrärm. nagayé zbo thgathinzako. madma kafarwä a srak nge katanwä.	My name is Abia. I am from *Rouku*. My father was Bäi. I'm going to tell you a story about what's-his-name. The short man's name was Kukufia. Kukufia lived in *Masu*. He had two wives. There was another man who lived here in *Rouku*. He had one wife. He set off to cut down a sago palm. The name of that place is *Karesa Zfth*. While he was there at *Karesa Zfth*, he left his two children here, the older girl and younger boy.
nä kayé kukufia zenfara kofär. ŋarsfo zärsöfätha gardame rafisir kofä thoraksir. nafane gäw kofä rusima. ane entharukwr gardame krentharuf krenfar. ŋanrafinzr e mnzärfr. kar yf rä ŋars rokurokun. wati garda fä sanzina foba krenfar. zänfrefa sränrn. nafane yf zunbräknwrm ”kukufia kukufia kukufia”. wati katan nagayé fäthane ŋafe frükaren krakaristh. ”ngth kabe yanor”. fi mnzen boba thfrnm. etfth mnzen kafar mnzen. watik kukufia yanyak. kräs ”ey bä mane ethkgr mnzen?”. ”bä nä mane zbo nthkgr?”. keke katakatané nä zayafath. yakasi keke. yanyak kwot we mnz zräkwr. neba zräkwr. nagayé fäth kranmätrth. madma kafarwä katan srak fäth. wati thmesa bobo ŋars rokurokufo. wati foba zetharufath. gardame katan emothf sfrafinzrm. nafangth thden sfrärm gardan.	One day, Kukufia set off to go fishing. He went down to the river to paddle his canoe and look for fish. He had his harpoon to spear fish. He put the things in the canoe. He got into the canoe and set off. He paddled all the way to the *Mnzärfr*. This is a place on the riverbank. Then he left the canoe there and started to walk up here. When he came up, he shouted. He called his name ”Kukufia Kukufia Kukufia”. In the absence of their father, the little children heard this. ”Little brother, there's a man shouting!” The two were there in the house, in the sleeping house, in the big house. Then Kukufia approached. He asked them, ”Hey, who are you in the house? Who are you in there?” The little ones didn't answer. They gave no answer. He approached the house and knocked hard. He knocked on the other side. The little children came out. The big girl and the small boy. He took them to the riverbank there. Then they got into the canoe. The little girl paddled the canoe. Her little brother sat in the centre of the canoe.

Figure 5: The sample text in “community” format

For Kalamang, an endangered Papuan language, Tanzer et al. (2024) have tested the translation capabilities of language models versus humans by feeding them the grammatical description, an approach they call MTOB (Machine Translation from One Book), and then comparing their translations from Kalamang to English and vice versa. Their study shows that humans are more successful at present, but they also show several points for improving the models.

For Japhug, an small language of Southern China, and Tsez, a small language spoken in the Caucasus, Okabe and Yvon (2023) have experimented with Bayesian models for simultaneously segmenting utterances into words and morphemes. They have tested two models to simultaneously segment into words and morphemes: one segmenting in parallel and the other in a hierarchical manner. They show that in the unsupervised condition the hierarchical model produces higher accuracy. What's more is that the study makes a number of suggestions to improve the results, eg. by incorporating contextual word models or adding further levels of supervision like phonology.

Such examples show that NLP research, however preliminary, when applied to low-resource languages, can help both the linguists working in language documentation and description and the language communities in participating in the development of large language models, thereby, increasing the relevance of small languages and overcoming the digital divide.

7. Conclusion

The Open Text Collections project remedies the lack of recognized publication venues for text collections of under-resourced languages and thereby pushes further the efforts to make lesser-resourced language content available in digital formats. In order to overcome the digital divide, the project wants to provide existing structured data to speaker communities and academics alike, in a formats suitable for the respective groups. Furthermore, the project provides researchers the prestige they deserve (ie. a peer-reviewed book publication) for creating interlinear glossed texts. Finally, the project provides a source of data for NLP research and facilitates further typological research. There are currently 16 text collections being prepared for languages of Eurasia, and as the project grows, more data from the less-resourced languages of the continent will become available as data sources for NLP research and community purposes alike.

8. Bibliographical References

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Nicholas Evans and Alan Charles Dench. 2006. Introduction: Catching language. In Felix K Ameka, Alan Charles Dench, and Nicholas Evans, editors, *Catching Language: The Standing Challenge of Grammar Writing*, pages 1–40. Mouton de Gruyter, Berlin, New York.

Robert Forkel, Johann-Mattis List, Simon Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Sci Data*, 5.

Jeffrey Heath. 1980. *Nunggubuyu myths and ethnographic texts*. Australian Institute of Aboriginal Studies, Canberra.

Jeffrey Heath. 1982. *Nunggubuyu dictionary*. Australian Institute of Aboriginal Studies, Canberra.

Jeffrey Heath. 1984. *Functional grammar of Nunggubuyu*. Australian Institute of Aboriginal Studies, Canberra.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological Segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Florian Matter. 2024. Cldflex. URL: https://github.com/fmatter/cldflex.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. *Society for Computation in Linguistics*, 3:338–349.

Sarah Moeller and Mans Hulden. 2021. Integrating automated segmentation and glossing into documentary and descriptive linguistics. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 86–95. Association for Computational Linguistics.

Ulrike Mosel. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. *Language Documentation & Conservation Special Publication*, 4:235–250.

Sebastian Nordhoff. 2020. Modelling and annotating interlinear glossed text from 280 different endangered languages as Linked Data with LIGT. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 93–104, Barcelona, Spain. Association for Computational Linguistics.

Sebastian Nordhoff and Thomas Krämer. 2022. IMTVault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France.

Shu Okabe and François Yvon. 2023. Joint word and morpheme segmentation with Bayesian non-parametric models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 640–654, Dubrovnik, Croatia. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. ArXiv preprint, arXiv:2309.16575.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

The Extraction and Fine-grained Classification of Written Cantonese Materials through Linguistic Feature Detection

[1]Chaak-ming Lau, [2]Mingfei Lau, [1]Ann Wai Huen To
[1]The Education University of Hong Kong, [2]CanCLID
lchaakming@eduhk.hk, laubonghaudoi@icloud.com, towaihuenann@gmail.com

Abstract

This paper presents a linguistically-informed, non-machine-learning tool for classifying Written Cantonese, Standard Written Chinese, and the intermediate varieties used by Cantonese-speaking users from Hong Kong, which are often grouped into a single "Traditional Chinese" label. Our approach addresses the lack of textual materials for Cantonese NLP, a consequence of a lower sociolinguistic status of Written Cantonese and the interchangeable use of these varieties by users without sufficient language labeling. The tool utilizes key lexical markers identified from past linguistic research to determine whether a segment is Cantonese, Standard Written Chinese, mixed or unmarked. The task is reduced into string operations to allow for a flexible and efficient extraction of high-quality Cantonese data from large datasets mixed with Standard Written Chinese. This implementation ensures that the tool can process large amounts of data at a low cost by bypassing model-inferencing, which is particularly significant for marginalized languages. The tool also aims to provide a baseline measure for future classification systems, and the approach may be applicable to other low-resource regional or diglossic languages.

Keywords: Language Classifier, Cantonese, Diglossia

1. Introduction

Cantonese, a regional language prevalent in Hong Kong and parts of southern China, presents unique challenges and opportunities for the advancement of minority language resource development. Despite being a vibrant language with over 7 million users in Hong Kong (Census and Statistics Department, 2022; Bacon-Shone et al., 2015) and at least 40 million in nearby regions (Qu, 2021), it is currently considered a low-resource language (Joshi et al., 2020), notwithstanding its significant user base and clear economic demand.

The progress of Cantonese NLP has been disproportionately impeded due to the lack of appropriate written materials, a situation tied to the region's complex linguistic landscape. Like many low-resource languages with robust speaker communities, researchers have access to speakers and spoken materials but transcribed, written or labeled resources remain scarce. This scarcity is intensified by the diglossic situation in Hong Kong (Leung and Li, 2020), where most publicly available texts are written in Standard Written Chinese rather than Cantonese, or occasionally a blend of both. This situation is further complicated by copyright restrictions and the ineffectiveness of tools designed for Standard Chinese in accurately processing Cantonese.

The increasing need to compile resources for pre-training language models and generating automatic speech recognition training data is evident. An earlier version of this tool was first used as an efficient auto-classifier to mine Cantonese content from the vast amount of web data which contains a low percentage of Cantonese content. This paper further develops this method into a robust strategy that is devised based on past linguistic research. This paper first discusses a linguistic analysis of the "writing modes" involved in this classification task (§2), provides a linguistically-motivated task description (§3), and then presents a two-level rule-based implementation (§4) and an evaluation (§5) of the current library.

2. Cantonese and SWC

2.1. Contrasting the two varieties

The two main varieties under question are Cantonese (BCP 47: ***yue***) and Hong Kong Standard Written Chinese (SWC, BCP 47: ***zh-hk***). Both varieties are typically written in the Traditional Han Script (繁體中文). The former is usually used in speech, but it does cross the line occasionally: there is a higher chance of seeing Cantonese in informal writing, whereas the use of SWC is dominant in formal occasions. This is a case of diglossia (Ferguson, 1959), which refers to the use of two distinct varieties with different social statuses ("H" versus "L") and used in different social settings. The Hong Kong variant of SWC, often considered the "H" variety, is generally compatible with Mandarin and comprehensible to Chinese speakers outside Hong Kong. Cantonese dominates spoken communication, but is considered to be the "L" variety here. Its written form is unintelligible to non-users.

2nd EURALI Workshop, pages 24–29

Despite the apparent similarity between SWC and Mandarin, the two are significantly different in the Hong Kong context, as the former inherits some Cantonese lexical items and occasionally does not conform to Mandarin usage. Putonghua, the Mandarin-based national language of China, is seldom used among Hong Kong locals (See Li 2017, Leung and Li 2020 and Lai 2013), and therefore SWC is sometimes written by users who have zero Mandarin knowledge. Hong Kong SWC is filled with Cantonese elements in writing, which are analyzed as deviations from the Mandarin standard by some scholars (Shi et al., 2014; Tin, 2020), and simply a different register (or version) of Cantonese by others (Bauer, 1988; Snow, 2004, 2008).

Here is an example showing the difference between the two, and why this is not just a Cantonese versus Mandarin classification problem (See Lau 2024 for a full discussion). These sentences are modified from widely circulated examples found in teacher training materials in Hong Kong, which serve to illustrate the multiple writing norms used in Hong Kong. SWC words not accepted in spoken Cantonese are underlined. Cantonese elements that are SWC-violating are enclosed in boxes. Other elements without any special formatting are shared between SWC and Cantonese. LSHK Jyutping romanization is added on top of the characters.

(1) SWC	taa1 他	wo4 和	dai6dai6 弟弟	zo6 坐	haau6ce1 校車	soeng5hok6 上學
(2a) Can1	keoi5 佢	tung4 同	sai3lou2 細佬	daap3 搭	haau6ce1 校車	faan1hok6 返學
(2b) Can2	keoi5 佢	tung4 同	dai4dai2 弟弟	co5 坐	haau6ce1 校車	faan1hok6 返學
(3) Mixed	keoi5 佢	wo4 和	dai4dai2 弟弟	co5 坐	haau6ce1 校車	faan1hok6 返學
"*He and his younger brother go to school by school bus.*"						
(4) Unmarked	dai4dai2 弟弟	co5 坐	haau6ce1 校車			
"*Younger brother takes the school bus.*"						

Table 1: The spectrum between Cantonese and SWC

The SWC sentence (1) represents the norm taught in schools, distinct from everyday speech in (2a), mainly in terms of word choice. Despite this, they are pronounced in Cantonese using a nearly identical set of grapheme-to-phoneme conversion rules, rendering the sentence comprehensible, albeit unnatural-sounding, in spoken Cantonese. Some words are shared between SWC and Cantonese, for example the word for 'school bus' is shared, and the SWC words 'brother' and 'sit' are also legitimate in Cantonese. Sentence (2a) can be adjusted to resemble SWC more closely, as shown in (2b), without undermining its validity as a well-formed Cantonese sentence. Texts that mix the two, as in sentence (3), also exist. This sentence is not accepted in speech, nor is it recognized as SWC. This type of blending, or translanguaging, is commonplace in some use cases, e.g. texting. Conversely, there are sentences that are acceptable in both SWC and Cantonese, as shown in (4). This is a short sentence that does not contain any marked feature that will violate the convention of either SWC or Cantonese, and therefore usable in both forms.

The example above highlights the similarity between writing norms in Hong Kong, indicating that the non-binary nature of the problem. It is feasible, and indeed prevalent, for sentences or fragments to possess multiple statuses. This necessitates a carefully defined set of labels to better encapsulate the classification task.

2.2. The two varieties in computational linguistic literature

The classification of CJK languages has been of interest to the community. Work includes Xu et al. (2017); Huang and Lee (2008); Lu et al. (2020). However, most classification attempts focused on the major varieties and usually not the finer-grained distinctions, which is most needed in a minoritized language context. Cantonese and SWC have also been discussed in the literature on machine translation (Wong and Tsai, 2022). Previous work often presupposed a clear demarcation between the two, and resulted in a conversion between extreme points on the spectrum.

The Cantonese and SWC distinction, as illustrated above, with varying levels of social acceptance, is not as straightforward.

The ***NLLB*** (No Language Left Behind) project (Costa-jussà et al., 2022) developed a classifier designed to classify closely related languages. While it significantly contributed to the detection of sub-Saharan varieties, it struggled to accurately distinguish between Yue Chinese (***yue***, which is taken as Cantonese here) and Hong Kong Chinese (***zh-hk***), with results falling at chance level (p.33, figure 9). Upon further examination, this issue stems from the underlying FLORES dataset, which incorrectly labeled all SWC data as *yue*.

FastLangID[1] is a tool built on the original fastText model, emphasizing accurate classification between Asian languages. It supports three Chinese locales: Simplified Chinese (***zh-hans***), Traditional Chinese (***zh-hant***), and Yue Chinese (***zh-yue***). There is also a separate code for Cantonese

[1]https://github.com/ffreemt/fast-langid

(***yue***). The results from this library do not match the expectations of the task.

From this brief review, it is clear that the classification between Cantonese and SWC requires further scrutiny. This issue extends to many other underrepresented varieties. A bottom-up approach captures the differences between existing datasets, but determining where to draw the line (during data collection or labeling) requires top-down judgments from linguistic literature. This will be discussed in the subsequent section.

3. Linguistically-motivated task definition

Due to the noted inadequacy of a bottom-up approach, this section reviews the linguistic literature to reach a more accessible definition for the labeling of these closely related varieties. The challenge lies in determining a meaningful way for distinguishing between Cantonese and SWC.

Criterion 1: Text Comprehensibility Shi et al. (2014) base their classification on text comprehensibility by native, monolingual Mandarin speakers, suggesting that a text containing 50% or more incomprehensible Cantonese elements qualifies as Cantonese writing (p.6). This is, however, a negative definition that relies on the linguistic intuition of an external group of users, not Hong Kong users. For the classification task, the definition of SWC should capture the localized idealization of what the standard is like by Cantonese speakers, with some tolerance of local words. On the other hand, Cantonese is characterized by its authenticity as judged by its users, not by the existence of words unique to Cantonese, but by not using words that sound odd (i.e. violate the requirements).

Criterion 2: Distribution of Cantonese Elements Snow (2004) offers an in-depth analysis of the distribution of Cantonese and Standard Chinese elements in broadly-defined Cantonese writings, distinguishing six sub-types of Cantonese text based on how Cantonese is inserted. His work notably identifies intermediate mixing patterns (Random mixing, Patterned mixing, SWC narration with Cantonese dialogues), which are distinct document types requiring classification.

This paper uses the latter criterion as the basis for the classification task.

3.1. Language Labels for the Task

The two major categories in this task, Cantonese and SWC, and other related, intermediate varieties, are defined linguistically below based on the division of labor observed by speakers from Hong Kong.

<table>
<tr><th></th><th>Example words</th></tr>
<tr><td>Cantonese feature</td><td>[嘅嗰啲咗佢喺咁嘢冇啫哋畀...
唔 [係得會好識使洗駛...</td></tr>
<tr><td>Cantonese exclude</td><td>(關係 | 吱唔 | 咿唔 | ...</td></tr>
<tr><td>SWC feature</td><td>[這哪唄咱啥甭那是的...</td></tr>
<tr><td>SWC exclude</td><td>是 [否日次非但旦]| ...
[目綠藍紅中] 的 | 的 [士確式] ...</td></tr>
</table>

Table 2: A subset of items used for classification.

Cantonese A text that conforms to Cantonese speech in a non-verbatim reading process. Following this requirement, the use of SWC-marked elements will be a violation. Text under this category can be used in conversation.

SWC The school-taught written Chinese form, which is similar to Mandarin in many aspects but is read out in Cantonese. The writing process can be described as a replacement of words in Cantonese speech to eliminate disallowed elements (Lau, 2024).

Mixed A piece of text that contains random use of Cantonese and SWC elements, characterized by violation of both Cantonese and SWC requirements. For longer texts, there are two finer-grained labels: *"CantoneseInSWC"* and *"MixedInSWC"*, which refer to patterned insertion of Cantonese or Cantonese/SWC mixed segments in dialogues or quotes, while keeping SWC as the main language for the narrative.

Unmarked A string that does not show any features that clearly violate either Cantonese or SWC requirements.

3.2. Classification Approach

The core of the classification is keyword or key-string based, which is a variant of the bag-of-word strategy, but with units larger than words. This is also similar to the strategies used in LIWC (Pennebaker), widely used in social sciences research. Here is an abridged list[2] with features of Cantonese that clearly violate SWC, and vice versa. These features can be expressed in terms of lexical violations, which can be understood as elements that must not appear in the idealized varieties.

The features listed above are all lexical items. There are grammatical elements that Cantonese allows whereas SWC bans, such as the classifier-noun structure in the subject position (e.g. 隻狗 "CL-dog", 個袋 "CL-bag"). Such detection requires sentential parsing and may not bring significant

[2]A full list can be found in the project's public repository.

gain in document classification accuracy, and was therefore not implemented.

A segment is considered markedly Cantonese if it contains some Cantonese features, and does not contain SWC elements that violate the norm for Cantonese. A document is Cantonese if its constituent segments are either markedly Cantonese or Unmarked.

4. Implementation

Our proposed method has been implemented in Python and made publically available[3].

By default, regardless of the length of the document, classification will be done to the incoming string and a 4-way classification will be returned. This can be used for a short segment (e.g. a couple of sentences), or a longer document.

In the implementation, we first defined a list of Cantonese and SWC features in regular expressions (exemplified in Table 2) and the following variables:

1. $canto$: (# of Cantonese_Feature – Cantonese_Exclude) / Total_Features.

2. swc: (# of SWC_Feature – SWC_Exclude) / Total_Features.

3. $tolerance$: Highest acceptable percentage for a Neutral sentence, defaults to 0.01

4. $presence$: The threshold indicating "significant presence" of a variety, defaults to 0.03

5. $prevalence$: The difference between the ratio of two varieties that shall be counted as an overwhelming presence, defaults to 0.9

For each input segment, the number of Cantonese and SWC features are obtained by regex matches and classified into four classes based on the logic below:

For more accurate classification, two additional parameters can be set.

1. seg This option delimits all lines with clear punctuation marks (full stops, question marks, etc.) to obtain individual sentences. With multiple sentences, we can determine the category of the document more accurately. If a main category (either Cantonese or SWC) plus Unmarked sentences accounts for 95% of all segments, this will be returned as the label. If there is no clear winner, it will be returned as a Mixed document.

[3]https://github.com/CanCLID/cantonesedetect

Algorithm 1 Logic for Segment Judgment

if $canto + swc = 0$ AND $swc < tolerance$ AND $canto < tolerance$ **then**
  Unmarked
else
  if $(canto - swc) > prevalence$ AND $swc < presence$ **then**
    Cantonese
  else if $(swc - canto) > prevalence$ AND $canto < presence$ **then**
    SWC
  else
    Mixed
  end if
end if

2. $quotes$ This option divides the document into two parts: all text enclosed in a pair of quotation marks (quotes) and other text surrounding the quotes (matrix), the two sets will be sent to the classifier separately. This mode is particularly useful for the sub-categorization of Mixed writing, which is often done in a patterned manner, such as the use of Cantonese dialogues in an otherwise SWC text.

5. Evaluation

We constructed a test dataset with 420 sentences collected from published materials and social media from Hong Kong.

Table 3 shows some examples of this dataset. We first calculated the 4-way classification accuracy of our classifier, then we defined Cantonese as the positive label, thus the correct detection of Cantonese sentences as True Positives, and then calculated the confusion matrix and get the Precision and Recall results. Our experiments show that the 4-way classification accuracy can consistently remain 90%+.

5.1. Effectiveness of the tool

As mentioned above, the classifier in our experiments is implemented in a balanced way so that it doesn't put emphasis on any one of the 4 classes. However, since the original design goal was to extract Cantonese data from a large base of Chinese texts, we value its precision over recall, i.e. we prefer missing Cantonese sentences to misclassifying non-Cantonese sentences as Cantonese. Results of our evaluation are shown in Table 4. For other use cases where recall or overall accuracy is emphasized, one can adjust the classifier by adding/deleting the hard-coded linguistic feature list. For example, some elements like 和 ("and", as opposed to 同 in Cantonese) can be added as

Label	Number	Sentence examples
SWC	181	但這不應成為通車的阻礙 推廣心理和精神健康的重要性
Cantonese	59	就可以換購泰國直送嘅百分之百鮮芒果雪條 幫你輕鬆搵出全港最抵嘅貸款，甚至免息買二手車
Mixed	4	但長遠來講，都係申請息口較低的貸款比較划算 選定了心儀嘅機構先查詢個人實際年利率，咁會比較明智
Unmarked	176	如果你選擇租貸，就要預繳幾期供款 最低實際年利率：百分之五點一九
Total	**420**	

Table 3: Example sentences of our test dataset

a SWC feature for more aggressive filtering.

		Prediction	
		Cantonese	Non-Cantonese
Label	Cantonese	57	2
	Non-Canto	1	360
Precision			**0.983**
Recall			0.966
4-class accuracy			**0.967**

Table 4: Results of our approach on the test set

Our approach proved significantly better than existing methods, and is the first solution to effectively extract large-scale written Cantonese data for Large Language Model (LLM) and other downstream applications. Our approach reached 98.3%+ precision on our test dataset, which guarantees the extraction outputs are predominantly Cantonese.

On an AMD Ryzen 7 5800H CPU, our current implementation took 0.10 seconds to finish the classification of 420 sentences, compared to fastText's 0.48s with the lid.176.bin model.

Note that the current implementation is not fully optimized, and can be done so by implementing the strategy used in fastText.

5.2. Limitations

We acknowledge certain constraints of our language classification tool, listed as follows:

- **Precision:** The current implementation does not consider grammatical constructions, collocation and frequency. Adding more violation rules will give a higher precision.
- **Recall:** The tool may reject valid Cantonese or SWC expressions due to the use of certain strings in proper names that are not enclosed in quotes.
- **Workflow:** Codepoint-based filtering can be applied before determining the finer-grained distinctions.
- **Other varieties:** Currently the tool only classifies different genres used in Hong Kong, and does not take into account other forms of written Chinese varieties.

Despite these limitations, our tool demonstrates reasonable accuracy for the task. For our original use case which operates at the document level, multiple sentences form the basis of judgment. This ensures a fairly reliable classification. For a more purpose-general classifier, additional strategies can be added to further improve the tool's accuracy. This will be left for future work.

6. Conclusion

This paper discusses a classifier for Cantonese, primarily aimed at extracting relevant materials for training and beyond. While more sophisticated statistical or machine learning-based approaches could be employed, our rule-based approach utilizing simple string matching, has proven to be simple and high-performing.

A key insight of this solution is to approach language classification from research findings on vernacular writing, making a clear definition of language varieties. It is hoped that linguistically-motivated approaches will be considered in future task definitions for the classification of written forms of under-resourced languages.

7. Acknowledgments

We would like to express our sincere gratitude to the two anonymous reviewers for their valuable feedback, and members of the TypeDuck team at EdUHK for their continuous support.

8. Bibliographical References

John Bacon-Shone, Kingsley Bolton, and Kang Kwong Luke. 2015. *Language use, proficiency and attitudes in Hong Kong*. Social Sciences Research Centre, the University of Hong Kong, Hong Kong.

Robert S. Bauer. 1988. Written cantonese of hong kong. *Cahiers de Linguistique Asie Orientale*, 17(2):245 – 293.

Census and HKSAR Statistics Department. 2022. Main tables, 2021 population census.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 404–410, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mee Ling Lai. 2013. The linguistics landscape of Hong Kong after the change of sovereignty. *International Journal of Multilingualism*, 10(3):251–272.

Chaak Ming Lau. 2024. Ideologically driven divergence in cantonese vernacular writing practices. In Jean-François Dupré, editor, *The Politics of Language in Hong Kong*. Routledge.

Wai Mun Leung and David Chor Shing Li. 2020. *兩文三語: 香港語文教育政策研究 [Biliteracy and Trilingualism: Language Education Policy Research in Hong Kong]*. City University of Hong Kong Press.

David Chor Shing Li. 2017. *Challenges in Acquiring Standard Written Chinese and Putonghua*, pages 71–107. Springer International Publishing, Cham.

Xugang Lu, Peng Shen, Yu Tsao, and Hisashi Kawai. 2020. Unsupervised neural adaptation model based on optimal transport for spoken language identification. *CoRR*, abs/2012.13152.

James W Pennebaker. Linguistic inquiry and word count: Liwc 2001.

Shao Bing Qu. 2021. *粵港澳大灣區語言生活狀況報告（2021）[Report on the Status of Language Life in the Guangdong-Hong Kong-Macao Greater Bay Area (2021)]*. The Commercial Press.

Dingxu Shi, Jingmin Shao, and Zhiyu Zhu. 2014. *港式中文與標準中文的比較（第二版）[Hong Kong Written Chinese and Standard Chinese: A comparison] (2nd ed.)*. Hong Kong Educational Publishing Co.

Don Snow. 2004. *Cantonese as written language the growth of a written Chinese vernacular*. Hong Kong University Press.

Don Snow. 2008. Cantonese as written standard? *Journal of Asian Pacific Communication*, 18(2):190–208.

Siu-lam [田小琳] Tin. 2020. *香港語言文字面面觀 [Aspects of the language use in Hong Kong]*. Joint Publishing HK.

Ka Ming Wong and Richard Tzong-Han Tsai. 2022. Mixed embedding of xlm for unsupervised cantonese-chinese neural machine translation (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):13081–13082.

Fan Xu, Mingwen Wang, and Maoxi Li. 2017. Sentence-level dialects identification in the greater china region. *CoRR*, abs/1701.01908.

Neural Mining of Persian Short Argumentative Texts

Mohammad Yeghaneh Abkenar[1,2], Manfred Stede[1]
[1]University of Potsdam
Applied Computational Linguistics
[2]Bundesdruckerei Gruppe GmbH Berlin
yeghanehabkenar@uni-potsdam.de, stede@uni-potsdam.de

Abstract

Argumentation mining (AM) is concerned with extracting arguments from texts and classifying the elements (e.g., claim and premise) and relations between them, as well as creating an argumentative structure. A significant hurdle to research in this area for the Persian language is the lack of annotated corpora. This paper introduces the first argument-annotated corpus in Persian and thereby the possibility of expanding argumentation mining to this language. The starting point is the English argumentative microtext corpus part 1 (AMT) (Peldszus and Stede, 2015), and we built the Persian variant by machine translation and careful post-editing of the output. We call this corpus *Persian argumentative microtext* (PAMT). Moreover, we present the first results for Argumentative Discourse Unit (ADU) classification for Persian, which is considered to be one of the main subtasks of argumentation mining. We determine the ADUs and their types (claim vs. premise) by two methods: (i) span categorization using the deep learning model of spaCy Version 3.0 (a CNN model on top of Bloom embedding with attention), and (ii) a neural sequence tagger. The results that we obtain with the second approach are comparable to those achieved on the same subtask in AMT and its other translations.

Keywords: Argumentation Mining, Persian Argumentative Corpus, Persian Language Resource

1. Introduction

One of the most essential requirements for developing Natural Language Processing (NLP) solutions for any language is data in that language. Based on the findings of (Paolillo and Das, 2006), out of over 7,000 languages spoken globally, approximately 20 of them have text corpora containing hundreds of millions of words. The language having the most data is by far English, followed by Chinese and Spanish. Japanese as well as Western-European languages are other languages with sizable datasets. The bulk of the languages spoken in Asia and Africa, on the other hand, do not have the training data needed to create reliable, cutting-edge NLP systems. *Low-resource languages* are characterized by being less explored, lacking in resources, being underrepresented in computational tools, and by a lack of annotated data (Singh, 2008). While Persian may not be fully classified as a low-resource language in theoretical terms, according to Joshi et al. (2020), it falls within the category of "The Underdogs" (level 4) in the language race. This designation implies that for Persian there is a significant amount of unlabeled data, but when compared to languages such as English, Spanish, German, Japanese, and French, which belong to level 1, "The Winners," Persian has a smaller amount of annotated data available (Joshi et al., 2020). As noted by Shamsfard (2019), reputable datasets for training and testing Persian systems for important NLP tasks are lacking, although the language is spoken by around 110 million people. Hence, the scarcity of resources and annotated data makes Persian an interesting candidate for research focused on addressing the needs of language with limited resources.

In recent years, progress in the wider field of natural language processing (NLP) such as pre-trained transformer-based models (Devlin et al., 2018) in combination with the increasing availabilty of data of different types has created great potential for almost every area in NLP, including argumentation mining (Stede and Schneider, 2018; Lawrence and Reed, 2020). Argumentation Mining (AM), and specifically the problem of finding argumentation structures in text, has received much attention in the past ten years, but with the research mainly focusing on English.

Broadly, AM can be seen as an extension of sentiment analysis. While sentiment analysis is about "what people think about an entity X", AM extends this to "why people think Y about X", thus uncovering more complex argumentation processes rather than just opinions and sentiments.

Aside from academic interest, AM attracts attention due to its wide range of applications, such as exemplified in the IBM Debater Project.[1] Further, argumentation mining can be used for a variety of important applications such as:

- Decision assistance, using AM in decision making on a controversial issue

[1]https://research.ibm.com/haifa/dept/vst/debater.shtml

2nd EURALI Workshop, pages 30–35

- Product reviews, where AM tools can be applied to product reviews, for instance, to understand why customers value a product.
- Writing support, to assess the quality of argumentative text and provide feedback to authors

Unlike most NLP problems, AM is not a single, straightforward task but a constellation of subtasks. In this paper, we focus on Argumentative Discourse Unit (ADU) classification, which is defined by Hidey et al. (2017, p. 14) as follows:

- *Claim* (Conclusion): A statement articulating the speaker's perspective on a particular issue. It can include predictions, interpretations, evaluations, and expressions of agreement or disagreement with others' assertions.
- *Premise* (Evidence): a statement put forth by the speaker to reinforce a claim, aiming to convince the audience of the claim's validity. While premises can convey opinions, their primary purpose is not to introduce a new viewpoint but rather to support or attack one already expressed by another proposition.

Identifying these components is consistent with the standard definition of an argument, as stated by (Van Eemeren et al., 2004), which requires at least one claim and one statement of evidence, referred to as a premise.

A major contribution of this paper is the free availability of the first annotated Persian corpus for argumentation mining, based on a corpus of short English and German texts introduced by (Peldszus and Stede, 2015). Additionally, we present the first model for argumentation mining for Persian short argumentative texts. Our results on ADU classification can be considered as the first results on this task in Persian. They indicate that sequence tagging models, which have been used for other languages, can also be considered a useful approach for this task in Persian.

2. Related Work

To the best of our knowledge, there are currently no argumentative corpora and results for argumentation mining in Persian, but there is some research on argumentation mining for other non-English languages. Aker and Zhang (2017) created the first annotated Chinese corpus using existing English corpora and manually matched claims and premises with parallel Chinese texts. (Namor et al., 2019) presented an early model for AM for Italian short argumentative texts. By adapting the model created by (Peldszus and Stede, 2015) to Italian and semi-automatically interpreting the original English corpus, they constructed a corpus of Italian microtexts. They utilized two phases for translation: in the first phase, they automatically translated the entire corpus using the DeepL translator service, known for its high-quality translations. In the second phase, they did manual post-editing. They reported results on all four original subtasks of AM according to (Peldszus and Stede, 2015), namely classifying attachment (at), central claim (cc), role (ro) and function(fu). Similarly, (Fishcheva and Kotelnikov, 2019) provided a Russian-language corpus for AM, which is based on machine translation of the Persuasive Essays corpus (Stab et al., 2014) and the AMT corpus. They investigated specifically the subtask of ADU role classification as "proponent" or "opponent".

3. Corpus

3.1. Original Corpus: AMT

The AMT corpus (part 1) consists of 112 short argumentative texts. 22 texts were written by the authors as "proof of concept" of the idea, and 90 texts were collected in a controlled text production experiment in which students wrote short texts, according to suggested length and rhetorical complexity (Peldszus and Stede, 2015).

All texts have been originally written in German and then were professionally (manually) translated into English. Although the texts are short, they are also 'complete', and the argument structure is generally quite clear. The annotation scheme for the AMT corpus has been constructed on the basis of Freeman's approach (Freeman, 2011). Essentially, the ways in which premises and claims are modeled corresponds to a hypothetical dialectical exchange between a proponent and an opponent. We show an example of an annotated text from the AMT corpus in Figure 1.

In the IAA study reported by Peldszus and Stede (2015), three annotators agreed on the complete task (in accordance with the annotation guidelines) with a Fleiss k=0.83 score, and with significantly larger agreement on the fundamental difference between support and attack relations.

The original AMT corpus comes in an XML format. We have extracted texts and their labels using regular expressions (regex) and other extraction packages such as beautifulsoup.[2] Overall, this first part of the AMT contains 112 claims (one for each text), and 464 premises.[3]

[2]BeautifulSoup

[3]Both parts of the English corpus, as well as annotation guidelines and further information, can be found here: argmicro

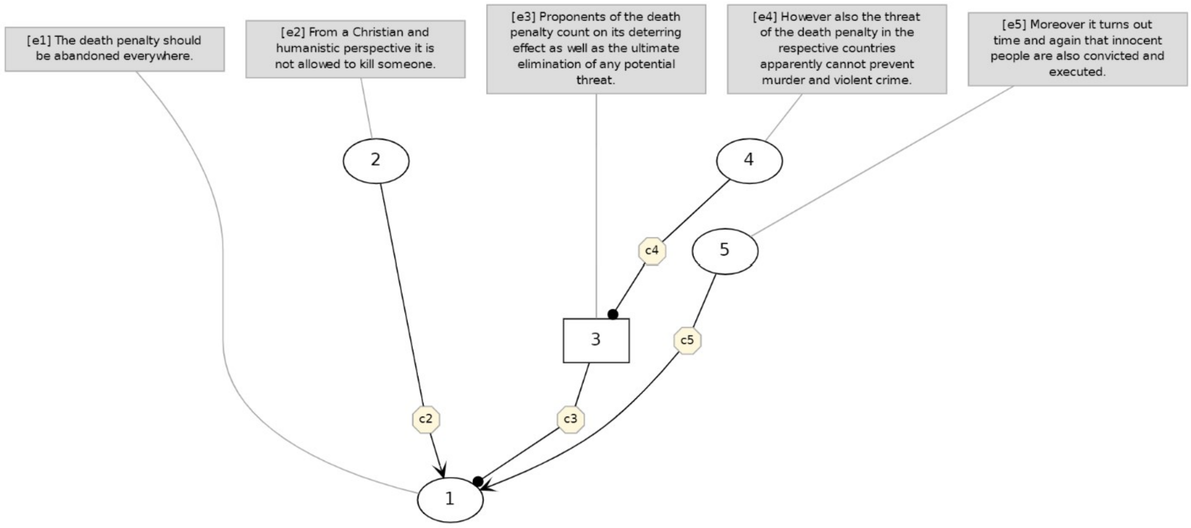


Figure 1: Example text from AMT corpus part 1 (argument 26) and its argumentation structure: Text segmented into ADUs; proponent and opponent role nodes (ellipses versus rectangles); supporting and attacking relations (arrow head, circle head). The first ADU (e1) is the claim, the others are premises.

3.2. Persian Corpus: PAMT

We created PAMT by translating the English AMT (part 1) and mapping all layers of annotations from AMT to PAMT.

Translation. This process was divided into two steps, automatic translation and manual post-editing. For translation, we used the Google translate application programming interface (API).[4] Then, translations were carefully proofread using an XML editor and a customised version of the annotation tool Prodigy,[5] which is a scriptable annotation tool of spaCy.[6]

A maxim for post-editing was to keep the original sentence order and structures of the English texts as parallel as possible to the Persian, so that mapping sentence- and clause-level annotations will be facilitated. English names (such as names of streets, countries, etc.) were translated to Persian. The post-editing was done by an English translator who is an expert in English literature and fluent in Persian.

In the resulting PAMT, the majority of texts consist of four, five, or six segments (ADUs), with an average of 5.1 segments. On average, each text has 3.7 sentences, with an average of 89.5 tokens per text. All other statistics are consistent with those reported in the original paper (Peldszus and Stede, 2015).

Annotation. While in our current work, we focus on classifying only the ADU types (claim, premise), we also mapped the relation annotations (support and two types of attack) from AMT to PAMT. Thus, the Persian corpus provides the same tree structures as those that are illustrated in Figure 1.

The annotated corpus and accompanying code is freely available.[7]

4. Experiments and Results

Since AMT and PAMT texts are short, they do not contain non-argumentative material. Therefore, ADU annotation covers the texts completely, so that the task of identifying claims and premises reduced to a binary classification. (This is in contrast to longer texts such as those in the Persuasive Essays corpus by Stab and Gurevych (2014), which can contain non-argumentative sentences.)

We experiment with two separate approaches, span categorization and neural sequence tagging. For the first approach, we divided the corpus into 90 texts for training and 22 texts for evaluation. For the second approach, we used 3-fold cross-validation. In order to prepare the corpus for the classification tasks, we used spacy and hazm[8] for tokenization and adding part of speech (POS) labels.

Span Categorization. As our first approach, we view the task as a span categorization problem. We used spaCy, an open-source library for NLP.

[4] https://pypi.org/project/googletrans/
[5] https://prodi.gy/
[6] https://spacy.io/
[7] https://github.com/myeghaneh/PAMT
[8] https://github.com/roshan-research/hazm

Recent improvements in spaCy Version 3.0 and Prodigy allow us to label spans even when they are potentially overlapping and nested (though this does not occur in our corpus). Specifically, we use spaCy's *SpanCategorizer* with a CNN model on top of Bloom embeddings with attention.

Neural Sequence Tagger. Following the approach of Chernodub et al. (2019) and Abkenar et al. (2021), we implemented a neural sequence tagger with the Flair NLP framework[9] to identify argumentative units and classify them as claim or premises in PAMT. For sequence labeling tasks, the calculated character-based embeddings are passed into a bidirectional long-short-term memory conditional random field. The tagger learns to assign *B*-$\{C|P\}$ and *I*-$\{C|P\}$ tags to tokens, representing the beginning or the "interior" of claim and premise, respectively. We did a few experiments on different Persian Word embeddings, and we chose Persian FastText embeddings trained over crawls as pre-trained language models (fa-crawl) (Akbik et al., 2019).

We trained on-the-fly in every training mini-batch. This means that the embeddings are not stored in memory. The advantage is that this keeps the memory footprint low. A sample output is shown in Figure 2 with colored labels for the two types of ADUs.

Span Categorization	P	R	F1
PREMISE	0.535	0.523	0.529
CLAIM	0.571	0.545	0.558

Table 1: Class-specific results of ADU classification for PAMT by span categorization.

Sequence Tagging	P	R	F1
PREMISE	0.737	0.304	0.410
CLAIM	0.618	0.734	0.670

Table 2: Class-specific results of ADU classification for PAMT by Sequence tagging using 3-fold cross-validation.

Results. Tables 1 and 2 show a comparison of the class-specific results for our best performing models on PAMT by the two approaches. Overall F1 values are given in Table 3: using span categorization we achieve a micro F1-Score 0.55 for claim vs. premise. Applying the neural sequence tagger with Farsi embeddings yields 0.64 micro F1-Score. These results are, to best of our knowledge, the first that have been reported for this ADU classification task on Persian. In Table 3, we also show the corresponding result reported by Abkenar et al. (2021) for the English AMT corpus.

Method	F1
Persian SpanCategorizer	0.550
Persian NeuralSequneceTagger	**0.636**
Engilsh NeuralSequneceTagger	**0.718**

Table 3: Comparison of PAMT model performance (micro F1-Score) for ADU classification (claim vs. premise) to the result on English AMT by Abkenar et al. (2021).

5. Conclusion and Outlook

Based on the English Argumentative Microtext Corpus, we have produced the first Persian argument-annotated corpus and make it available to a general audience. The AMT corpus was systematically translated into Persian using machine translation (Google Translate API), post-processing, and post-editing of the AMT. Additionally, we projected the entire annotation layer of AMT onto PAMT, making it available for further analyses. Second, we investigated the problem of classifying Argumentative Discourse Units (ADUs) into two classes, "Premise" and "Claim", in Persian. The best performance in Persian was achieved by the Neural Sequence Tagger, with a micro F1-score of 0.64. In comparison to results from experiments with the Italian corpus(Namor et al., 2019), the results were somewhat lower, possibly due to the smaller Persian model in Flair or to differences between the languages. The results of the Neural Sequence Tagger were notably better than those of the SpanCategorizer.

For further research, we plan to conduct more experiments by introducing a corpus similar to the Persuasive Essay Corpus (PEC) (Stab and Gurevych, 2014) in Persian, and using both corpora for cross-domain train/test experiments.

6. Ethics and Limitations

Given our restricted resources for conducting independent studies, our focus was exclusively on Persian, without consideration for other languages spoken in Iran, such as Kurdish, Laki, (Ahmadi et al., 2023b) Baluchi, (Kargaran et al., 2023) and Gilaki (Ahmadi et al., 2023a), which are often deemed low-resource languages. We aspire to broaden our research scope to encompass these languages in the future and encourage collaboration with scientists from these language areas interested in similar topics.

Our study was constrained by a relatively small corpus size, but we prioritized translation quality.

[9] https://github.com/flairnlp/flair

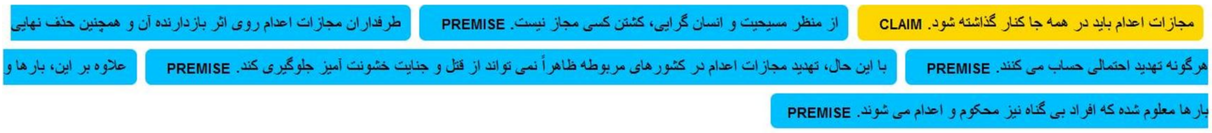


Figure 2: An example text (argument 26) about topic "Introduction of Capital Punishment" in Persian corpus (PAMT) with the prediction of claim from premise by our model.

To address this, we plan to expand the corpus in future versions and incorporate larger datasets. Additionally, our focus solely on ADU classification represents a limitation. Future research will encompass other subtasks within argumentation mining, broadening our findings.

7. Acknowledgement

We would like to thank Mahtab Dadarsefat for her exceptional translation and post-editing contributions, as well as to Sara Shahmohammadi, for her insightful comments on the translation process. Additionally, we thank our anonymous reviewers for the constructive comments, and we extend our heartfelt gratitude to EURALI 2024 for their dedication to promoting diversity and inclusion in language technology by providing resources and tools for lesser-resourced languages.

8. Bibliographical References

Mohammad Yeghaneh Abkenar, Manfred Stede, and Stephan Oepen. 2021. Neural argumentation mining on essays and microtexts with contextualized word embeddings.

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023a. Pali: A language identification benchmark for perso-arabic scripts. *arXiv preprint arXiv:2304.01322*.

Sina Ahmadi, Zahra Azin, Sara Belelli, and Antonios Anastasopoulos. 2023b. Approaches to corpus creation for low-resource language technology: the case of southern kurdish and laki. *arXiv preprint arXiv:2304.01319*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. 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.

Ahmet Aker and Huangpan Zhang. 2017. Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining*, pages 67–72.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

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*.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-lingual argumentation mining for russian texts. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 134–144. Springer.

James B Freeman. 2011. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. *arXiv preprint arXiv:2310.16248*.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Ivan Namor, Pietro Totis, Samuele Garda, and Manfred Stede. 2019. Mining italian short argumentative texts. In *CLiC-it*.

John C Paolillo and Anupam Das. 2006. Evaluating language statistics: The ethnologue and beyond. *Contract report for UNESCO Institute for Statistics*.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International Conference Language Technologies for All (LT4All)*.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *ArgNLP*, pages 21–25.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Frans H Van Eemeren, Robert Grootendorst, and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Endangered Language Preservation: A Model for Automatic Speech Recognition Based on Khroskyabs Data

Ruiyao Li, Yunfan Lai
Trinity College Dublin, Trinity College Dublin
College Green, Dublin 2, Ireland, College Green, Dublin 2, Ireland
lir4@tcd.ie, yunfan.lai@tcd.ie

Abstract

This is a report on an Automatic Speech Recognition (ASR) experiment conducted using our Khroskyabs data. With the impact of information technology development and globalization challenges on linguistic diversity, this study focuses on the preservation crisis of the endangered Khroskyabs language, a language falling under the Gyalrongic language group (Glottocode: guan1266). We used Automatic Speech Recognition technology and the Wav2Vec2 model to transcribe the Khroskyabs language. Despite challenges such as data scarcity and the language's complex morphology, preliminary results show promising character accuracy from the model. Additionally, the linguist also has given relatively high evaluations to the transcription results of our model. Therefore, the experimental and evaluation results demonstrate the high practicality of our model. At the same time, the results also reveal issues with high word error rates, so we plan to augment our existing dataset with additional Khroskyabs data in our further studies. This study provides insights and methodologies for using Automatic Speech Recognition to transcribe and protect Khroskyabs, and we hope that this can contribute to the preservation efforts of other endangered languages.

Keywords: ASR, Gyalrong, Khroskyabs

1. Introduction

According to Moseley (2010), in recent decades, alongside the development of information technology, there has been a gradual reduction in the diversity of human languages. Particularly with the challenges of globalization, the preservation of many Asian languages, such as the Khroskyabs language, is facing a crisis. Therefore, we hope to apply automatic speech recognition tools to transcribe some traditional stories in Khroskyabs into IPA, thereby protecting the language and culture by preserving these traditional stories in Khroskyabs.

1.1. Endangered Language Preservation

Khroskyabs is a language in Gyalrongic language group spoken in western Sichuan, China. Currently, there are about 9000 native speakers of Khroskyabs. The transmission of Khroskyabs relies entirely on speech, as it lacks a writing system. It is Gong (2017) indicates that the Gyalrongic language group is classified as endangered, gradually heading towards extinction under the pressure of Sichuanese Mandarin and the Amdo Tibetan. Our fieldwork on the Khroskyabs language also observed that the local people, due to pursuing education and work opportunities outside, have become less proficient in speaking the Khroskyabs language compared to earlier generations. Additionally, there are no specialized schools teaching the Khroskyabs language. Furthermore, the lack of a written system for Khroskyabs exacerbates its preservation challenges.

The preservation of the Khroskyabs language is important. Due to its long-standing use in secluded mountainous regions, minimally affected by external linguistic influences, Khroskyabs, just like many other Gyalrongic languages, has retained a substantial amount of ancient Sino-Tibetan features (Gong, 2017). It holds significant importance in Sino-Tibetan historical linguistics, as it preserves the complex consonant clusters and verb morphology in proto-Sino-Tibetan. Additionally, the Khroskyabs language is a highly morphologically rich language, characterized by numerous verb affixes and root alternations. These features of it are beneficial for the study of Sino-Tibetan historical linguistics, underscoring the urgent need for attention to its endangered status. Beyond its scholarly value, preserving this language also supports the cultural identity and heritage of its speakers, promoting inclusion and underscoring the importance of linguistic diversity. These considerations drive our pursuit of new preservation methods, including the application of automatic speech recognition tools, to protect the Khroskyabs language for the benefit of both academia and its native speaker communities.

1.2. Method

This section outlines the methodology employed in our study, focusing on the selection of the

2nd EURALI Workshop, pages 36–40

Khroskyabs language as our subject and the implementation of the Wav2Vec2 model for automatic speech recognition.

1.2.1. The Source of the Data

The language we have chosen is Khroskyabs, which belongs to the western branch of the Gyalrongic language group (Sun, 2000a,b; Huang, 2001; Lai, 2017). Khroskyabs is among the less spoken languages within this group.

Protecting endangered languages faces a significant challenge: transcription. Linguists may spend up to half an hour transcribing just one minute of audio. Therefore, using automatic speech recognition can expedite and streamline the transcription process for endangered languages, enabling us to efficiently document and preserve them. However, in the process of automatic speech recognition, a large amount of input data is required to train the model. Compared to the data for many endangered languages (Guillaume et al., 2022), the dataset for Khroskyabs is notably larger (Lai, Yunfan, unpublished). These data include recordings of local elders telling traditional stories in Khroskyabs and transcriptions by the linguist, ensuring transcription accuracy. Because the model cannot recognize the punctuation, we removed all punctuation marks. In this experimental training of our model, we only used one hour of Khroskyabs data to assess the model's utility when faced with languages lacking ample annotated data. The previous data format was .txt, but in our training, we required the data format to be .eaf, which necessitated re-splitting the audio and inputting transcriptions in ELAN. Therefore, moving forward, we plan to augment the amount of Khroskyabs data to enhance the model's accuracy after putting more date into ELAN. Afterwards, the dataset will be uploaded to Pangloss to make it publicly available.

1.2.2. The Model Selection

Currently, there are several automatic speech recognition tools available, and for low-resource languages, there are some data augmentation techniques that can help improve ASR systems (Bartelds et al., 2023). For our project, we have selected the XLS-R-Wav2Vec2 model fine-tuned for low-resource languages (O'Neill et al., 2023). This model has shown promising results in the context of Newar and Dzardzongke languages spoken in Nepal (O'Neill et al., 2023).

The Wav2Vec2 model employs multitask learning to optimize both its audio feature extractor and language model components, thereby enhancing its performance on low-resource languages. Importantly, the model supports transfer learning, allowing knowledge transfer from a related high-resource language model to improve the training process and performance of the low-resource language model.

In this study, we will demonstrate the development of an automatic speech recognition model for Khroskyabs using the model. For fine-tuning the model, several hyperparameters were configured to optimize the training process. The training used a per-device train batch size of 8, combined with gradient accumulation steps set to 2. The model was set to train for a total of 50 epochs. Additionally, a learning rate of 3e-4 was chosen.

In Section 2, we will present the experimental results concerning Khroskyabs transcription. In Section 3.1, we will discuss the challenges faced and potential future improvements. Lastly, we will have a conclusion in Section 4.

2. Evaluation of the Results Using Khroskyabs

In this section, we will showcase the model trained using Khroskyabs data as the foundation and discuss the outcomes of our training.

2.1. Experimental Results

In our experiment, we used one hour of Khroskyabs data. The Khroskyabs dataset comprises six audio recordings, each featuring different speakers, thereby adding a challenge to the model training process.

The quality of our automatic speech recognition system is evaluated using two metrics: character error rate (CER) and word error rate (WER). Both metrics quantify the disparity between the recognized text and the original text, with character error rate focusing on character-level errors and word error rate on word-level errors. These are two classic metrics used to evaluate automatic speech recognition systems.

The Figure 1 illustrates the average word error rate at each step of the training process across iterations, ranging from 100 to 1400, when training with one hour of Khroskyabs data.

From here, it can be observed that after 100 to 600 iterations of training, the results were far from satisfactory, with the word error rate approaching nearly one hundred percent. However, after further training, particularly at 1200 iterations, the word error rate decreased to eighty-seven percent.

Although the results above may not be entirely satisfactory, we can also observe the median character and word error rates for each checkpoint, as depicted in Figure 2.

From this table, it can be observed that at the first checkpoint, the median character error rate

Step	Training Loss	Validation Loss	Wer
100	6.402600	3.848435	1.000000
200	3.320600	3.311335	1.000000
300	3.281700	3.279070	1.000000
400	3.212000	3.196068	1.000000
500	2.755600	2.177797	1.010610
600	1.357800	1.550854	1.031830
700	0.870800	1.424223	0.893899
800	0.617200	1.517462	0.920424
900	0.508900	1.593333	0.888594
1000	0.400400	1.622587	0.875332
1100	0.296900	1.762700	0.877984
1200	0.282100	1.816226	0.875332
1300	0.225100	1.911786	0.899204
1400	0.212300	1.914231	0.899204

Figure 1: Word error rate across iterations

```
01/400 Median CER:     1.0
01/400 Median WER:     1.0

01/800 Median CER:     0.213
01/800 Median WER:     0.75

01/1200 Median CER:    0.192
01/1200 Median WER:    0.667
```

Figure 2: The median character and word error rates at each checkpoint

is 1.0, and the median word error rate is also 1.0. However, by the third checkpoint, the median character error rate further decreases to 0.192, while the median WER decreases to 0.667. These results indicate that as training progresses, both the character-level and word-level error rates of the model gradually decrease. This outcome is more satisfactory, particularly considering that the model is trained on only one hour of language data, with a character error rate of 0.19 already being low.

In evaluating the performance of our model, we also have box plots to visually analyze the character error rate and word error rate of the Wav2Vec2 model, as shown in Figure 4.

We can see from the box plots that the character error rate results demonstrated satisfying performance, with a median close to 0 and a very compact interquartile range. It suggests that the majority of characters were accurately recognized. Although there were a few minor outliers, they had minimal impact on the overall performance. In contrast, the word error rate median was relatively higher, indicating that recognition errors at the word level were more common and dispersed, and the distribution of word error rate included a

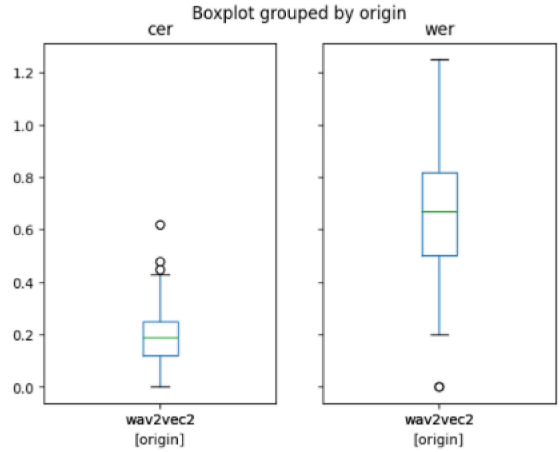


Figure 3: Boxplots of character and word error rates

significant outlier. Overall, these findings suggest that our model exhibits greater stability and accuracy in character recognition.

2.2. Evaluation by the Linguist

In addition to using classical character and word error rates as our model evaluation metrics, we also sought the opinion of the linguist regarding the transcription quality. In Figure 2.2, we compare the transcription results produced by our model with those transcribed by the linguist.

Figure 2.2 displays the transcription results for three randomly selected sentences. The top row shows the transcriptions provided by the linguist for the recordings, while the bottom row presents the transcriptions generated by our trained automatic speech recognition model. Portions highlighted in red indicate errors in our model's transcription, while parts within parentheses denote omissions in our model's transcription. It can be observed the model demonstrates good accuracy in character recognition, with occasional errors in discerning vowels, distinguishing between voiced and voiceless consonants, and occasionally omitting some consonants and tones.

From the results shown in this figure, the number of corrections required to achieve transcription quality appears to be lower than the quantity indicated by the character error rate we obtained earlier. Discrepancies between assessments of classic evaluation methods and assessments by linguists are also mentioned in Guillaume et al. (2022). The linguist, who is also the annotator of the training data and specializes in Khroskyabs, has also provided a positive evaluation of the model's accuracy. This suggests that the practical utility of the model we trained may be higher than what is evaluated by character and word er-

1. ǣɕə ŋəɲɟə̂jə tædpɑ̂fstænpɑtəɣə ætə̂jə jəʁûpʰræsce zjǣŋ vî nærôdpɑ rəŋó təɣî
   ǣɕə ŋəɲɟə̂jə tædpɑ̂fstænbɑtəɣə ætə̂(jə) jóɣpʰrǽsceə zjǣŋ vî nærű̂dpɑ rəŋóə
2. ɲəmɑ̂ nɑ̂riɕəvæ mdǣʁʑə gərgôɣ ndʑêjtə lodzê pʰɑ́dtəmɲədtəɣu vîskə
   ɲəmɑ̂ nɑ̂riɕəvæ mdǣʁʑəgərgô (ɣ)o(n)dʑêjtə lodzê pʰɑ́təɣmɲəddtə(ɣu) vîskə
3. jóŋ ŋəɲɟə̂jə jóŋ χtsʰətsʰə̂ jóŋ mêr nærjé cə̂ tʰjǣ ŋókɑ næsɑ̂ɲ næpʰrǽɲrætə χtsʰətsʰə̂ ɕəntɕʰê nɑví rəŋó
   joɑ́ɣ ŋəɲɟə̂jə jɑ́ŋ χtsətsʰə̂ntə jóŋmé(r)nærjé cə tʰjǣ ŋókɑrætəɕə næsɑ̂(ɲ)næpʰrǽɲnrætə̂ χtsʰətsʰə̂ ɕəntɕʰê nɑví rəŋó

Figure 4: Comparison between the transcription result of the linguist and the model

ror rates.

3. Reflections And Further Studies

Now, we can observe that the model we trained has demonstrated a satisfactory level of accuracy in transcribing Khroskyabs. In this section, we will critically reflect on our approach and propose some possible further studies.

3.1. Reflections on the Model

Although our model has demonstrated a low character error rate, our results also reveal a higher word error rate, which is likely associated with the complex morphology of Khroskyabs. This indicates that our model currently lacks the capability to accurately capture word boundaries and has not fully adapted to the unique phonological and morphological characteristics of Khroskyabs. It shows the complexities of transcribing low-resource languages, where limited data availability and linguistic diversity pose significant challenges.

Furthermore, the model we developed struggles with accurately transcribing Chinese loanwords. In our data, Khroskyabs is transcribed using the International Phonetic Alphabet, while Chinese loanwords are transcribed using the Pinyin system. This has led to a higher error rate in processing Chinese loanwords. Additionally, the limited occurrence of Chinese loanwords in speech exacerbates the model's challenges in handling them.

3.2. Further Studies

To address the issue of a high word error rate, we plan to augment our existing dataset with additional Khroskyabs data. During this round of training, we used one hour of Khroskyabs data, and we aim to double this amount by incorporating an additional hour of data. This expansion is expected to enrich our dataset, providing a broader linguistic base that could enhance the model's understanding of the complex morphology.

To address the challenge of low transcription accuracy for Chinese loanwords, we plan to revise the original data, retranscribing all the Chinese loans and replacing Pinyin with IPA. Also, we plan to increase the presence of Chinese loanwords in our training dataset, which could potentially improve the model's proficiency in accurately processing these loanwords.

4. Conclusion

In the experiment, we demonstrated the transcription of endangered languages such as Khroskyabs using automatic speech recognition technology and the Wav2Vec2 model. Our results, after many training iterations, showed a median word error rate of 0.67 and a character error rate of 0.19. These results indicate an optimistic outcome in character accuracy and have been highly rated by linguists. However, we still face challenges, notably the high word error rate, likely due to the model's insufficient morphological understanding of the language. In the future, we plan to incorporate more data to enhance the model's transcription accuracy.

5. Bibliographical References

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

Xun Gong. 2017. The Morphology of the Gyalrongic language group and Old Chinese. *Ancient Scripts and Historical Phonology of Chinese: Fudan Journal of Chinese Civilization Studies*, pages 134–156.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Bufan Huang. 2001. Research on the Language Belonging of Guanyinqiao. *Language and Linguistics*, 2(1):69–92.

Yunfan Lai. 2017. *Grammaire du Khroskyabs de Wobzi*. Ph.D. thesis, Université Sorbonne Paris Cité.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Alexander O'Neill, Marieke Meelen, Rolando Coto-Solano, Sonam Phuntsog, and Charles Ramble. 2023. Language Preservation through ASR.

Jackson T-S Sun. 2000a. Parallelisms in the verb morphology of Sidaba rGyalrong and Lavrung in rGyalrongic. *Language and linguistics*, 1(1):161–190.

Jackson T-S Sun. 2000b. Stem alternations in Puxi verb inflection: Toward validating the rGyalrongic subgroup in Qiangic. *Language and linguistics*, 1(2):211–232.

6. Language Resource References

Lai, Yunfan. unpublished. *Siyuewu Khroskyabs texts*. Unpublished field notes.

This Word Mean What: Constructing a Singlish Dictionary with ChatGPT

Chow Siew Yeng, Chang-Uk Shin, Francis Bond[1]
[1]Department of Asian Studies, Palacký University, Olomouc
siewyeng001@e.ntu.edu.sg, papower2@gmail.com, bond@ieee.org

Abstract

Despite the magnitude of recent progress in natural language processing and multilingual language modeling research, the vast majority of NLP research is focused on English and other major languages. This is because recent NLP research is mainly data-driven, and there is more data for resource-rich languages. In particular, Large Language Models (LLM) make use of large unlabeled datasets, a resource that many languages do not have. In this project, we built a new, open-sourced dictionary of Singlish, a contact variety that contains features from English and other local languages and is syntactically, phonologically and lexically distinct from Standard English (Tan, 2010). First, a list of Singlish words was extracted from various online sources. Then using an open Chat-GPT LLM API, the description, including the defintion, part of speech, pronunciation and examples was produced. These were then refined through post processing carried out by a native speaker. The dictionary currently has 1,783 entries and is published under the CC-BY-SA license. The project was carried out with the intention of facilitating future Singlish research and other applications as the accumulation and management of language resources will be of great help in promoting research on the language in the future.

Keywords: Singlish dictionary, ChatGPT, Data Generation

1. Introduction

1.1. Purpose

In recent NLP research, studies that require a considerable amount of language and computational resources have become increasingly mainstream (Touvron et al., 2023). While BERT, a transformer encoder-based LLM released in 2018 (Devlin et al., 2019), learned from around 3.3 billion words, state-of-the-art models now use more than ten times that amount. For example, OpenAI's GPT3 and subsequent models have been trained on datasets of at least 500 billion tokens. In addition, the modeling performance of LLM seems to improve in proportion to the amount of data sets, the number of model layers, and the number of parameters (Kalyan et al., 2021).

However, most of these language resources are only available in the most popular languages. In particular, the largest amount of data is available for standard English, and the proportion of data available for languages such as Singlish is much smaller. This project seeks to fill this gap in the current state of language resources available. In the case of Singlish, due to its non-official status, there are no large dictionaries. Therefore, this project, inspired by the various language resource creating efforts through LLM (de Schryver, 2023; Elsner and Needle, 2023) in the last year, creates a dictionary through a pretrained LLM (ChatGPT) and manual editing in order to promote future research on Singlish. The generated responses by the LLM were used to create a rough draft of the dictionary, speeding up the process as compared to writing a dictionary from scratch. While ChatGPT is not fine-tuned to Singlish, there are no LLMs trained specially for Singlish that are able to generate responses like ChatGPT.[1] Even though this effort is not sufficient for an LLM training, building this open-sourced dictionary that will allow for contributions from the public is a step in the right direction.

This dictionary (with the exception of its examples) is written in standard English and aims to describe the Singlish phrases and their usage to its readers. Each entry in the constructed dictionary contains the word or phrase's definition, example sentence, pronunciation, part of speech and alternate spellings and language of origin if applicable.

The results of this study are released under the open license 'CC-BY-SA' and are expected to enable further opportunities for future Singlish studies.

1.2. Singlish

Singlish is a contact language whose emergence can be attributed to the diversity of languages spoken in Singapore (Soh et al., 2022) such as Hokkien, Bazaar Malay, Cantonese (which only exerted more influence in recent years: Lim, 2011) and English. It has English as its superstrate language and has its lexicon, syntax and even prosody influenced by substrate languages like

[1]There is SingBert which is a fine-tuned version of BERT on Singlish but it does not generate responses like ChatGPT, and is trained on a small and noisy corpus.

Baba Malay (Lim, 2011). In particular, a prominent feature that separates it from standard English is its extensive use of particles, on which, numerous studies been conducted (Wong, 2005; Leimgruber, 2016). Singlish syntactic features include optional inflection of verbs (for third person singular subjects), optional articles and the lack of plural marking (Chow and Bond, 2022). While certain syntactic features are common in Singlish, the extent to which they are used by individual speakers varies with factors such as other languages they speak. Another feature that the variety has is tone. The tonality of Singlish has been extensively covered by Lim (2011). Particles, in particular, are distinguished through tone. Hence, it is an important feature and is taken into account in this dictionary. The local variety is important to the Singaporean identity and is used heavily in daily conversations (Li, 2021) despite the government's efforts to completely replace it with standard English (Cavallaro and Ng, 2009).

Due to the diversity in the sources for its lexicon and the informal nature of the language, its vocabulary is not well captured in standard dictionaries. Moreover, in many cases, the origin language and standard spelling (if it exists) of a word in Singlish may not be known to the average speaker.

(1)	*Dey,*	*wǒ mén*	*paktor*	*always*	*makan*	*at*	
	T	Md	C	E	M	E	
	Hey	we	date		eat		
	kopitiam	*one.*					
	M+H	S					
	coffee shop	PART					

'Hey when we date we always eat at the coffee shop (one).'

For instance, a typical Singlish utterance is shown in 1 (Cheng, 2021)[2] where the second line represents the origin language of the term (T-Tamil, Md-Mandarin, C-Cantonese, E-English, M-Malay, H-Hokkien/Hakka, S-Singlish). A single utterance could easily involve words from multiple languages but is spoken, simply, as Singlish. It is not code-switching, because a speaker of Singlish typically will not speak all or even any of the non-English languages.

Hence, a Singlish dictionary functions to collect into the lexicon words originating across various languages and unique expressions that are used in this variety and explain them to the readers.

[2]While Cheng (2021) labelled 'one"s origin to be unknown, in this paper, it is instead labelled as 'S' for Singlish as its usage has Chinese origins, but has taken its form in an English word. More on the origins of 'one' can be found in Wong (2005).

1.3. Existing Resources

Although Singlish is predominantly a spoken language, there are some resources in the forms of corpora and dictionaries. The different corpora of Singlish and Singapore English illustrate the usage of language in different mediums and time periods.

The NUS SMS corpus (Chen and Kan, 2015) is a corpus of 67,093 text messages focusing on English and Mandarin Chinese. The data was crowd-sourced from Amazon's Mechanical Turk, ShortTask, ZhuBaJie and NUS students. Although this was not a Singlish centred project, 46.9% of the English SMS were contributed by people from Singapore (Chen and Kan, 2012, p. 18). With the sheer amount of data, this corpus contains a significant collection of Singapore English. It is a public corpus and can be used freely with citation.

The International Corpus of English, the Singapore Corpus (ICE-Singapore) is a record of spoken and written English text in Singapore with many subcategories including telephone calls, broadcast interviews, academic writing and creative writing. The corpus contains a lot of text data but as the data is collected across many different domains, Singlish is mainly found in private dialogues (categorised with the tag 'S1') and makes up only a small proportion of the total data.

The Red Dot Baby Talk wordlist is a list of words used in the Red Dot Baby Talk Quiz (Woon and Styles, 2021), a quiz made to help document 'baby talk' in Singaporean children and the age of acquisition of these words. The list consists over a hundred words and because of the nature of the quiz, are mostly basic words that are likely to be in the lexicon for young children or babies, including onomatopoeia. This list is published under the CC by NC 4.0 license and can be used for non-commercial purposes.

There have also been other non-official dictionaries for Singlish, each with distinct properties.

In 2001, the Coxford Singlish Dictionary (Goh, 2002) was first published. Its name is a play on the Oxford dictionary and it is, according to Huddart (2014, pp. 75), 'an amalgamation of satirical comment on Singaporean society and a source of linguistic data'. The dictionary is written in a playful tone and contains, in addition to Singlish words, Singlish pronunciations of words in standard English. It has 809 Singlish words and phrases.

The Dictionary of Singlish and Singapore English (Lee, 2004) (henceforth DSSE) is a substantial collection of over 1,000 entries providing their origin (or speculated origins), their meanings and real examples of usage. Visitors to the website are free to suggest contributions to the dictionary through an online form. Unfortunately, the dictionary was last updated in 2016 and the data on the dictionary does not have an open licence.

Singlish Dictionary (hereafter singlish.net) is an online dictionary compiled from May 2017 to August 2018. Despite being active for a relatively short period of time, there are around 140 entries including phrases and common acronyms used in Singlish. The content of the website is shared "AS IS" with no warranties, and confers no rights.

Wiktionary is a freely available international dictionary that contains words in various languages. A portion of the words are tagged with the categories of Singlish and Singapore English and the entries for these words contain their pronunciations, alternative forms and etymologies. This dictionary is licensed under CC-BY-SA.

As of 2022, there were 27 words marked as Singlish in the Oxford English Dictionary (OED): very incomplete coverage.[3]

The various resources differ in terms of the licensing rights and comprehensiveness of their lexicon. Both the Coxford dictionary and DSSE have more entries, but are not open source or actively maintained. singlish.net and the OED are neither large nor open. On the other hand, while Wiktionary is open, it has relatively fewer Singlish entries. Currently, there is no Singlish dictionary that is open-source, has a relatively large lexicon and takes note of particle tones. Therefore, this project will build a new, large, open lexicon to support future research through fully open-source data with a new methodology.

2. Methodology

2.1. Outline

The process of the project can be divided into the following steps:

1. **Collect a list of Singlish words** from DSSE, Red Dot Baby Talk (Woon and Styles, 2021) and Wiktionary (Wiktionary.org).

2. **Generate draft entries** for these words using ChatGPT (OpenAI, 2023).

3. **Manually check the entries** by the principal annotator through cross validation with existing sources and native speaker intuition to improve the entries (e.g., changing to standard spelling if it exists: *chochok* to *cocok*).

   At the same time, 4 other annotators were recruited for the purpose of getting an agreement score for the tags chosen for each of the responses as part of evaluation.

4. **Normalize the pronunciations** of the words in X-SAMPA notation.[4]

2.2. Details

To elaborate on the process, firstly, a script was written to scrape the lexical entries from DSSE and wiktionary. The words from Red Dot Baby Talk were taken from the source data (a spreadsheet). We did not take the categorised examples and descriptions in order to keep our license open. These lists were then merged.

Through the ChatGPT API, a draft of description, pronunciation, POS and alternate spellings was then produced for the collected words. The following prompt was given:

> You are writing a Singlish dictionary. Now, please write a dictionary entry of 'WORD'. With a full description of the word, its POS, pronunciation and alternate spellings if there are.

In the prompt, 'WORD' was replaced with each word in the word list. An example of a query and response pair is shown below for the entry of 'bo hiu'.

Query: "You are writing a Singlish dictionary. Now, please write a dictionary entry of 'bo hiu'. With a full description of the word, its POS, pronunciation and alternate spellings if there are."

Response: "Sure, here's a dictionary entry for 'bo hiu' in Singlish:
Word: Bo hiu
Pronunciation: boh hiu
Definition: Bo hiu is a Singlish phrase that translates to "don't care" or "not bothered". It is often used to indicate a lack of interest, indifference or apathy towards a person or situation. It can also be used to express nonchalance or to shrug off something that might be seen as important or serious.

Example:

- A: "Eh, why didn't you answer my call last night?"

- B: "Bo hiu lah, I was sleeping."

Overall, 'bo hiu' is a common Singlish phrase that is used to express a casual attitude towards something or someone.

---

[3] https://www.smartsinga.com/facts-about-singapore/, accessed 2023-10-23

[4] X-SAMPA is a computer-readable representation of the IPA that uses only ASCII. It was proposed as an extension of SAMPA (Wells, 1995).

The prompts were submitted from May to September 2023. All the responses were then collated into a single document for ease of evaluation. After that, the entries were manually checked by the principal annotator (the first author). In cases where their intuition was insufficient to make a judgement, ChatGPT's response was cross-validated against other sources such as the previously mentioned dictionaries. After comparing against other sources and native knowledge of the variety, each entry was marked with a tag of 'Yes', 'No', 'Not sure', 'Partial-pron/etymology' or 'Partial-example/spelling, etc.', indicating the acceptability of the description given by ChatGPT and if inadequate, the nature of its inadequacy. The full list of tags and their examples can be found in Table 1 under Section 3. In some cases where the tag was a form of 'Partial' or 'No', an updated description was then written if the principal annotator's knowledge was enough; other dictionaries were not referred to in such cases. This precaution was taken to avoid any form of copying from other resources that are not specified to be free to use. In cases where the description was tagged 'Yes' or 'Not sure', the original descriptions were kept[5] but the entries differ in the dictionary through an additional tag which indicates whether they have been verified.

Another possibility was for an entry to be merely an alternative spelling of another. In that case, only one description was preserved and the other similar entries were then annotated with the 'Redirect to' tag, with an indication of the chosen entry with the description. For instance, 'balukoo' and 'baluku' (both meaning 'bruise'(noun)) were in the original word list but 'baluku' was chosen as the dictionary entry and 'balukoo' was then redirected to the other spelling. Each final description was also split into its components such as definition, part of speech and example.

At the same time, 4 other annotators were recruited and initially grouped into 2 groups. Annotators in group 1 were each tasked with tagging the first 50 entries in the list of ChatGPT's responses provided on separate google sheets while annotators in group 2 were tasked with the next 50 entries. The final tags chosen by all the annotators (including the principal annotator) are compared in Section 3.

Lastly, pronunciations were added to the entries in X-SAMPA. Although ChatGPT provided pronunciations of the words in a majority of its answers, they were written in a variety of formats, from IPA to phonetic spelling and as compared to the meanings of the words, are more often inaccurate. Hence, they were all re-annotated and standardised in X-SAMPA.

On average, the descriptions from ChatGPT were evaluated and annotated at a rough rate of 30 words/hour including cases where the tag 'No' was given and no corrected description is given. This rate varied significantly based on the accuracy of ChatGPT's descriptions which affected the need for updated definitions and examples.

2.3. Annotators

The 5 annotators involved in this project range from 25 to 30 years old. They are all Singaporean Chinese native speakers of Singlish and 3 out of the 5 have gone through some form of military service.[6] During the initial split, it was ensured that both groups had members with military service experience. All annotators are unpaid volunteers.

3. Results

3.1. Data

At this point in time, the dictionary has 1,783 words/phrases in its lexicon, including some which have been redirected to an entry of an alternate spelling. 138 of those entries originated from the Red Dot Baby Talk word list, 1,201 entries came from DSSE and the other 462 entries from Wiktionary. The duplicates in the entries have been deleted in the portion of the data that has been annotated hence the total is less than the sum of the entries in the resources. This collection of Singlish vocabulary with definitions is currently the largest Singlish dictionary in existence.

3.2. Detailed Annotation

Out of the 579 entries that have been manually annotated by the principal annotator, 46.2% of ChatGPT's descriptions were completely satisfactory (tagged with 'Yes'), 26.8% were unsatisfactory (tagged with 'No'), 4.9% were partially correct but unsatisfactory information on pronunciation or etymology, 8.0% were partially correct but unsatisfactory in areas such as example, spelling or etc. and 6.8% were tagged 'Redirect-to', which means that the word is an alternative spelling of another one (which has been chosen as the entry) listed on the vocabulary. The last 7.3% were tagged 'Not sure' which means that the annotator did not know the word well enough to evaluate the correctness of the description and, just in the detailed annotation, that the description given by ChatGPT was similar enough to that in the online resource but also,

[5] Small edits such as removing the mention that the phrase 'is a Singlish phrase' were made.

[6] This may be a factor in their knowledge of the Singlish words collated considering that many Singlish words are more often used in the military context.

in most cases, contained something additional that has not been verified. An example of each of these tags is shown in Table 1.

The description for ‘leh’ was wrong as it is not used to seek confirmation. The rest of the responses gave accurate definitions of the word but included additional less accurate information, with most of the inaccuracies lying in the description of the pronunciation. For example, both ‘a’s in ‘atas’ should be pronounced as /a/. On the other hand, in the entry ‘Ah Long’, its etymology as well as alternate spellings are wrong. In the case of ‘blank’, the description largely matches the native speaker’s knowledge and the other sources, but a particular part - ‘a lack of emotion’ - has yet to be verified and hence, it was tagged with ‘Not sure’. The last row of Table 1 shows ‘balukoo’ which is redirected to another entry ‘baluku’. In cases like the last, the accuracy of the description is not taken into account in the tag chosen.[7]

A sample of updated entries after this process of detailed annotation (excluding their tags and pronunciations) is shown below.

- ***leh*** (Particle: -) Leh is a Singlish particle.

  1. The high level tone leh (tone 1) is used to ask questions. It has a similar meaning to ”what about” except it occurs after the noun.

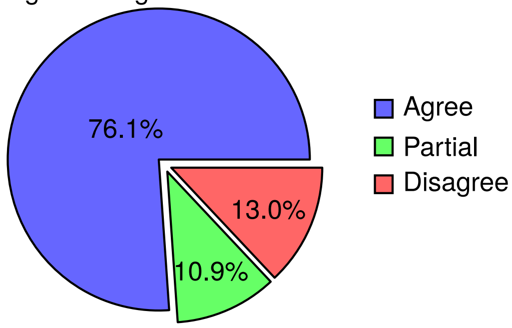
  2. The mid level tone leh (tone 3) is used usually with the intention of persuasion and to sound more convincing. — *1. We’re about to go already. Your friend leh1? 2. Eh, don’t like that leh3. We need you here to play mahjong. (translation: Don’t be like that. We need you here to play mahjong)*

- ***atas*** (Adj: Malay) Atas means ‘haughty’ or ‘snobbish’, often used to describe someone who is acting or behaving in an elitist or pretentious manner. Additionally, it can also refer to something that is high-end or posh, such as a luxurious restaurant or an expensive brand. Note: Atas is derived from the Malay language, where it means ‘upper class’ or ‘high society’. In Singlish, it has taken on a slightly negative connotation due to its association with snobbery and elitism. — *Wow, your friend is so atas. She only wants to eat at Michelin-starred restaurants!*

- ***Ah Long*** (N: Hokkien) Ah long is a term used to refer to loan sharks. They are typically unlicensed moneylenders who charge very high interest rates and use harassment, intimidation, and violence to collect payments from borrowers who are unable to pay off their debts — *I heard he borrowed money from an Ah Long to pay off his gambling debts.*

[7]The description for ‘balukoo’ is actually inaccurate.

Figure 1: Agreement scores after streamline



3.3. Inter-annotator Agreement

During the annotation process, some annotators were particularly enthusiastic and tagged more than the required amount. Thus, at the end, a total of 334 entries were tagged by at least 2 annotators and their agreement scores were calculated. With the predefined categories, it was found that 40.0% of the annotations had complete agreement (all annotators assigned the same category for the entry), while 13.9% had partial agreement (at least 2 but not all annotators agreed) and in 46.1% of entries, there was no agreement.[8] The agreement score is low but this can largely be attributed to the granularity of the tags. For instance, a majority of disagreements was associated with the ‘Partial’ tags including example, spelling, pronunciation. A breakdown of the distribution in scores and entries is shown in Table 2

In addition, of the annotations with complete agreement, 51.1% were annotated as ‘Yes’, 48.9% as ‘No’ and there were no agreement for ‘Partial’ tags. Annotations without complete agreement were relatively evenly spread: 39.7% as ‘Yes’, 19.8% as ‘No’, and 40.5% as ‘Partial’. The annotators annotated differently for the descriptions generated by LLM despite all being native speakers of the variety. We posit that the high level of disagreement is due to the fact that the responses contain a lot of information, and that the speakers, having their own idiolects, may interpret the expressions differently.

As part of further processing the data, we streamlined the categories into three: ‘Useful definition’ (comprising ‘Yes’ and both ‘Partial’s), ‘Not sure’ and ‘Wrong definition’. This categorization yielded 76.1% complete agreement, with 74.3% of those agreeing that ChatGPT’s outputs gave the correct definition. More detailed dissections of the scores are shown in Figure 1 and Figure 2.

[8]In this calculation, we treated ‘Not sure’ like ‘NA’ and did not count those annotations. As such, the agreement scores were only tabulated for entries which had at least two tags excluding ‘Not sure’ and this was 230 entries.

Word	Response	Tag
chai poh	Certainly, here’s a dictionary entry for “chai poh”: Word: Chai Poh Pronunciation: chai-poh / tʃʌɪ p̚ɒ̄ Alternate spellings: Cai Poh, Chye Poh Description: Chai poh is a term used in Singlish to refer to preserved radish that’s commonly used in local dishes such as fried carrot cake, chee cheong fun, and oyster omelette. The radish is finely sliced, mixed with salt and spices, and then left to ferment for a couple of weeks, which gives it a pungent aroma and a slightly salty taste. Some people also enjoy eating chai poh on its own as a snack. Example: “Can you please add extra chai poh to my fried carrot cake?”...	Yes
leh	Sure, I’d be happy to help! Leh (pronounced ”luh” or ”leh” with a rising tone) is a common Singlish particle that is used to soften a request, statement, or command. It typically follows a sentence and is added to the end to express a mild suggestion or to seek agreement or confirmation. It can also be used to express surprise or disbelief. Here are some examples of how ”leh” is used in sentences: ”Can you pass me the salt, leh?” ...	No
atas	Sure, here’s the dictionary entry for “atas”: - Word: Atas - Pronunciation: a-tos (with emphasis on the first syllable) Definition: Atas is a Singlish term that means “haughty” or “snobbish,” often used to describe someone who is acting or behaving in an elitist or pretentious manner. Additionally, it can also refer to something that is high-end or posh, such as a luxurious restaurant or an expensive brand. Example of usage: “Wow, your friend is so atas. She only wants to eat at Michelin-starred restaurants!”...	Partial-pron/ etymology
Ah Long	- Pronunciation: ah long Alternate spellings: Ah Beng, Ah Long, Ah Beng Long Definition:Ah Long is a term used in Singlish to refer to an illegal moneylender or loan shark. It is derived from the Hokkien dialect, with ”Ah” being a common prefix used to address someone, and ”Long” meaning a dragon. Together, the term suggests that these moneylenders are often ruthless and possess an intimidating presence, just like a dragon. Ah Longs typically operate through illegal channels, offering high-interest loans to individuals who are in need of immediate cash. They generally target vulnerable borrowers, often using intimidation tactics or violence to enforce repayment. Example: ”I heard he borrowed money from an Ah Long to pay off his gambling debts.”...	Partial-example/ spelling, etc.
blank	Sure, here’s the dictionary entry for ‘blank’: Word: Blank Pronunciation: /blæŋk/ Definition: An expression used to describe a state of mental confusion or inability to think of a response to a question or situation. It can also be used to describe a lack of emotion or enthusiasm towards something. Example: “Huh, sorry ah, I go blank already. What was the question again ah?” (Translation: “Sorry, I’m feeling confused and I can’t remember the question. Can you repeat it?”)...	Not sure
balukoo	Balukoo is a Singlish slang word that means ‘nonsense’ or ‘rubbish’. It is commonly used to describe something that is untrue or foolish. It is pronounced as buh-loo-koo with emphasis on the first syllable. Example: “Don’t listen to him, what he said is just balukoo.”	Redirect to

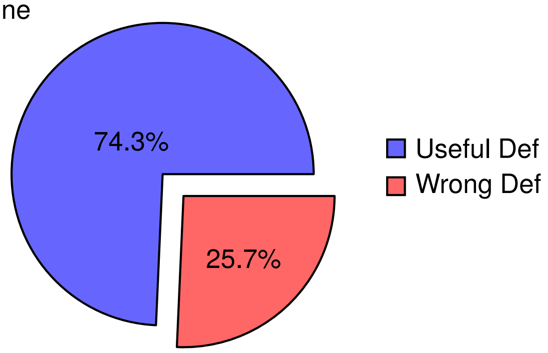
Table 1: Responses with tags

The responses here are shortened versions that only give up to the first example.

Table 2: Inter-Annotator Agreement

Entries	# Annotators	Agreement Scores		
		Complete	**Partial**	**Disagree**
1-53	4	4	24	25
54-100	3	13	26	7
101-334	2	65	-	170

Figure 2: Breakdown of agreed tags after streamline



4. Discussion

The scores in Figure 1 and Figure 2 indicate that for over half the samples, all annotators agreed that ChatGPT's responses were useful in terms of generating definitions for the entries. The significant difference between the proportion of 'Yes' annotated answers (which ranged around 25-60%) and the proportion of answers with just the correct definition shows that ChatGPT often outputs inaccuracies when providing additional information.

The fact that the model produces responses with at least a correct definition a majority of the time makes it helpful as a starting point for generating a dictionary though these outputs have to be checked before the dictionary can be reliable.

Although the initial granularity of the annotation tags resulted in a more rigorous annotation process and a lower agreement score, the complexity, together with how responses were not overwhelmingly of one level of correctness forces the participants to read each response carefully. It is intended to produce a more informative and accurate response. As a result, we were able to infer with confidence that approximately half of the samples have an agreed upon correct definition provided in the response. This translates into a significantly lowered cost for creating a new dictionary.

As part of the process of checking, ChatGPT's results were compared with DSSE, the Red Dot baby Talk and multiple online resources. We found no examples of existing text being reproduced exactly. In general, the examples given seem very different and although the same words occur in both definitions in a few entries, they are not close enough to be considered copying. For exampleconsider the following

- **ChatGPT** Agak agak is a Singlish phrase that derives its roots from the Malay language. It is typically used to express a rough estimation or a guess. The phrase can be translated to mean "roughly" or "approximately" in English. It is commonly employed in everyday conversations to describe a haphazard estimation of measurements, time, or quantities when precise information is not available.
- **DSSE** agak /ah-gah(k), ɑgɑ(k̚)/ n. & v. [Mal., conjecture, guessing; agak-agak approximately, as far as one can guess, more or less] Also agak-agak. A n. A guess, an estimation: Agaration. B v. Guess, estimate.

5. Current State

Currently, 579 of the entries have been annotated with tags and 399 of the entries are complete with POS, description, example, origin (if known[9]) and alternate spellings. The remaining 68% (1,204/1,783 entries) contain the unchecked ChatGPT descriptions as of now. The state of the entry (verified or unverified) will be displayed in the dictionary through a *verified* tag. We are going through the entries at the rate of around 100 per month.[10]

As noted in Lim (2011)'s paper, tone plays an important role in Singlish, especially for sentence final particles. As such, the tonal descriptions for the pragmatic particles[11] were added by hand. The tonal numbers were given through an approximate matching with the first six tones used in jyutping with the addition of the quick falling tone used by the particle *lah*. In that case, the diacritic used for the Mandarin fourth tone (làh) was used.

[9] This relies solely on the principal annotator's knowledge

[10] The difference between annotation and completion numbers is caused by the annotator's incomplete knowledge of the vocabulary. For instance, they may be unsure about the correctness of the response or they may know it is wrong but cannot describe the term well.

[11] The tonal particles in the dictionary are *ah, hah, hor, leh, lah, lor, mah, meh, sia, what, wor*

The breakdown of words according to their POS and Origin is given in Table 5.

POS	Count	%
Noun	253	64.7%
Verb	55	14.0%
Adjective	48	12.3%
Interjection	12	3.1%
Other	23	5.9%

Origin	Count	%
Malay	97	32.5%
English	77	25.8%
Hokkien	72	24.2%
Cantonese	17	5.7%
Other	35	11.8%

Table 3: POS / Origin of Annotated Words

6. Future Steps

The current dictionary has certain limitations and potential for further expansion.

We will tap into other Singlish language resources to continue NLP research, and strive to increase the dictionary's coverage and utility. A method that is being considered is the augmentation of the dictionary automatically. For every description given by ChatGPT, an automated process can be created that checks through every word inside it. Words in the description that are neither in standard English dictionaries nor already inside this Singlish dictionary can then be fed into the prompt mentioned earlier, continuously expanding the Singlish dictionary.

Subsequently, the Singlish words in the corpora such as the NUS SMS corpus can also be semi-automatically fed into ChatGPT to further expand the dictionary. These sentences can then be used as examples in existing or future entries. However, due to the format of the text messages, a certain amount of additional processing might be needed to ensure that the new entries and examples obtained are of a suitable format for the dictionary.

As an extension of this study, we could also compare the results of multiple prompts. This serves to increase our understanding of how useful ChatGPT is in this task, and also perhaps, generate more satisfactory responses. Another approach in this sense is to run multiple LLMs to generate multiple descriptive hypotheses that can be used together to build one complete entry.

The issue of variation could also be further explored. Variation in the annotator's knowledge of Singlish may account for the high level of disagreement. Singlish is not taught formally and the lexicon used varies largely depending on factors like home language, social circle, age, etc. It is possible to recruit annotators with more similar intuitions e.g., through selecting ones from a specific background (or perhaps even choosing only those who have a similar confidence/description of their own Singlish). However, we do not wish to pick one 'standard' variety. Instead, another approach could be to accept the description or a slight variation of it if at least one annotator puts 'yes', and note that it may not be universally accepted.

Finally, we would like to link the entries to wordnet senses, so that they can easily be translated into other languages through the Open Multilingual Wordnet (Fellbaum, 1998; Bond and Foster, 2013).

7. Conclusion

This study shows the relative effectiveness of using an LLM to create more resources for a low-resource language. For Singlish, while the accuracy of the descriptions are far from perfect, more than half were deemed by the principal annotator to be accurate and comprehensive, and the generation of the entries in general provided a baseline that facilitated the building of the dictionary. As compared to having to write all the entries manually, or even through crowd sourcing, this method is an efficient and low cost way of creating and expanding a dictionary. Overall, we have found ChatGPT a useful tool to make draft entries: around 50% of samples were usable as is. Correcting entries with some errors is still faster than writing descriptions from scratch, significantly reducing the amount of work.

The development of the largest Singlish open sourced dictionary and the first to provide a tonal description of particles in this project is a step towards collecting more Singlish data and improving the resource available for this variety. We expect to have checked and, as necessary, rewritten a majority in the dictionary by May 2024. Just like other open dictionaries such as Urban Dictionary have contributed towards training specialised embedding models for NLP (Wilson et al., 2020), we hope that this can contribute towards and encourage Singlish language research, especially since large-scale language resources are becoming increasing prevalent in the field of NLP. Although Singlish is used in everyday speech in Singapore, it does not enjoy the status of being an official language and there is no representative dictionary. To fill this void, through the use of LLM (ChatGPT) and other online resources, we have created a new, completely open-source Singlish dictionary.

The resource described in this study is published on Github with a CC-BY-SA license.

8. Acknowledgements

We would like to thank the reviewers for providing constructive feedback on how to improve our paper. Additionally, we would also like to thank Professor Suzy Styles for sharing her vocabulary lists.

9. Bibliographical References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362, Sofia.

Francesco Cavallaro and Bee Chin Ng. 2009. Between status and solidarity in Singapore. *World Englishes*, 28(2):143–159.

Tao Chen and Min-Yen Kan. 2012. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*.

Renae Cheng. 2021. 10 bizarre things singaporeans do that the rest of the world won't understand. [Online; posted 14-June-2021].

Siew Yeng Chow and Francis Bond. 2022. Singlish where got rules one? Constructing a computational grammar for Singlish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5243–5250.

Gilles-Maurice de Schryver. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4):355–387.

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

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Colin Goh. 2002. *The Coxford Singlish dictionary*, second edition. Angsana Books.

David Huddart. 2014. Declarations of Linguistic Independence: The Postcolonial Dictionary. In *Involuntary Associations*. Liverpool University Press.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing.

Jack Tsen Ta Lee. 2004. A Dictionary of Singlish and Singapore English.

Jakob RE Leimgruber. 2016. Bah in Singapore English. *World Englishes*, 35(1):78–97.

Zhuoyang Li. 2021. An analysis of the linguistic characteristics of Singlish. *Journal of Contemporary Educational Research*, 5(3).

Lisa Lim. 2011. Tone in Singlish: Substrate features from Sinitic and Malay. *Substrate Features in Creole Languages*, pages 271–288.

Ying Qi Soh, Junwen Lee, and Ying-Ying Tan. 2022. Ethnicity and Tone Production on Singlish Particles. *Languages*, 7(3):243.

Teresa Rebecca Tan. 2010. Singlish. *Singapore Infopedia*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

John C Wells. 1995. Computer-coding the IPA: a proposed extension of SAMPA. *University College of London*.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.

Jock Wong. 2005. "Why you so Singlish one?" A semantic and cultural interpretation of the Singapore English particle one. *Language in Society*, 34(2):239–275.

10. Language Resource References

Tao Chen and Min-Yen Kan. 2015. The National University of Singapore SMS Corpus.

ICE-Singapore. International corpus of english — singapore.

OpenAI. 2023. ChatGPT: Large-scale GPT-3.5 for Conversational AI.

Fei Ting Woon and Suzy Styles. 2021. Red-Dot Baby-Talk Quiz.

An Evaluation of Language Models for Hyper-partisan Ideology Detection in Persian Twitter

Sahar Omidi Shayegan[1,3], Isar Nejadgholi[2], Kellin Pelrine[1,3], Hao Yu[1,3], Sacha Levy[1,3], Zachary Yang[1,3], Jean-François Godbout[3,4], Reihaneh Rabbany[1,3]
[1]McGill University, [2]National Research Council Canada, [3]Mila - Quebec AI Institute, [4]Université de Montréal
{sahar.omidishayegan, kellin.pelrine, hao.yu2 , sacha.levy, zachary.yang}@mail.mcgill.ca
isar.nejadgholi@nrc-cnrc.gc.ca, jean-francois.godbout@umontreal.ca, reihaneh.rabbany@mcgill.ca

Abstract

Large Language Models (LLMs) are now capable of successfully identifying the political beliefs of English-speaking social media users from their posts. However, assessing how LLMs perform in non-English languages remains difficult. In this work, we contribute to this area of research by determining the extent to which LLMs can predict the political ideologies of users on Persian social media. We begin by discussing the challenges associated with defining political parties within the Persian context and propose a solution based on a technique designed for the detection of hyper-partisan ideologies on social media. We create a new benchmark and show the potential and limitations of both open-source and commercial LLMs in classifying the hyper-partisan ideologies of users. We compare these models with smaller fine-tuned ones, both on the Persian language (ParsBERT) and translated data (RoBERTa), and confirm that they considerably outperform generative LLMs in this task. We further demonstrate that the performance of the generative LLMs degrades when classifying users based on their tweets instead of their bios, even if tweets are added as additional information; whereas the smaller fine-tuned models are more robust and achieve similar performance for all input settings. This study represents a first step toward political ideology detection in Persian social media, with implications for future research to understand the dynamics of political conflicts in Iran.

Keywords: Computational Social Science, Persian Language, Ideology Prediction

1. Introduction

Political ideology detection using Twitter data (now X) has been extensively studied in the English language (e.g., Pelrine et al. (2023); Yu et al. (2023); Törnberg (2023); Barberá (2015); Pennacchiotti and Popescu (2011)). The few studies that focus on other languages are generally limited to Western democracies, where the analysis of political campaigns and elections on social media has been used to monitor shifts in public opinion and the interactions between different ideological groups (Rodríguez-García et al., 2022; Chen et al., 2017; Jiang et al., 2022). Therefore, there is a significant research gap in studies conducted in other languages and those focusing on different types of political systems . In this work, we address this important limitation by focusing on the case of Iran. Despite the pivotal role that this platform has played in influencing political narratives in this country (Khorramrouz et al., 2023; Kermani and Tafreshi, 2023), it remains difficult to understand how political conflicts unfold between supporters and opponents of the Iranian regime.

The task of delineating the ideological orientation of supporters and opponents to the Islamic Republic of Iran poses several challenges. Indeed, unlike democratic countries where political parties are well-defined, the main division in Iran is largely driven by political ideology, which is not channeled through organized and institutionalized partisan groups (Azadi and Mesgaran, 2021). Thus, in the absence of distinct political parties, our research focuses on the more direct computational task of categorizing distinct ideological markers, specifically, hyper-partisan users representing two extreme viewpoints: "Pro-Government", the government supporters committed to the principles of the Islamic Republic; and "Pro-Monarchy", those who favour the return of the former monarchical regime. We recognize that there are several other political ideologies in the Iranian political space, including secularists, reformists, women's rights activists and Kurdish activists. However, for this study, we have decided to group all of these remaining ideologies under the class of "Others". We acknowledge this limitation and leave the more in-depth analysis of this last category for future research.

Our analysis of hyper-partisan ideology prediction in Persian Twitter focuses on data collected during the Woman-Life-Freedom movement, from October 18th 2022 to January 11th 2023. This period saw a significant surge in Persian tweets, with users extensively employing political and ideological hashtags and key terms. The importance of this event makes this time frame crucial for understanding the dynamics of political conflicts in Iran. We first labelled the users in our data by relying on clear ideological stances declared in the Twitter bios of users. We refer to these users as hyper-partisan users. This approach allowed us to anchor our research on users from contrasting ideological

2nd *EURALI Workshop*, pages 51–62

backgrounds who are forthright about their beliefs, thereby ensuring minimal overlap of classes and mitigating the risk of mislabeling. Those without explicit indicators that failed to align with either one of these two groups were classified under an "Others" category, indicating a broader ideological spectrum. In this study, we specifically explored two tasks: 1) classifying hyper-partisan users based on the text in their bios and 2) classifying hyper-partisan users based on various combinations of the text found in their bios and in their tweets.

We, then, investigated the performance of different Large Language Models (LLMs) for identifying the above groups. Inspired by the widespread acclaim and proven efficacy of LLMs in diverse NLP tasks, including ideology prediction for English social media data (Tornberg, 2023; Yu et al., 2023), we evaluate these models for hyper-partisan ideology prediction within our labelled dataset. We begin with a comprehensive assessment of GPT-3.5, and then moved to other forms of LLMs, including open-source conversational LLMs such as Llama 2 Chat and WizardLM and smaller fine-tuned classifier models like RoBERTa (Liu et al., 2019a) and ParsBERT (Farahani et al., 2021).

Our results show that GPT-3.5 can classify bios with clear ideological markers reasonably well. However, this model is limited in the level of detail it can handle in the prompt and performs optimally only when all of its inputs are translated into English. Open-source conversational LLMs, such as WizardLM and Llama 2, achieve similar performances but also only when the data is translated into English. On the other hand, fine-tuned BERT-family LMs, both in Persian (ParsBERT) and English (RoBERTa), significantly outperform all generative LLMs. Overall, our results confirm that classifying tweets is a more challenging task for generative LLMs rather than classifying users based on the information found in their bios. More specifically, adding tweets to the input obfuscates the results of GPT-3.5, improves the classification performance of fine-tuned ParsBERT, and does not have a significant impact on fine-tuned RoBERTa. The main contributions of this paper are as follows:

- We evaluate various hyper-partisan ideology detection methods on Persian Twitter using different open and closed-source LLMs. Our work is a first step towards an area previously understudied despite Twitter's significance influence in Iran's political debates.
- Focusing on a period with a surge in Persian political tweets, we collect and label a new benchmark of Persian posts for this task and classify them into three main ideological groups: "Pro-Government", "Pro-Monarchy", and a third group, "Others", comprising various opposition factions.
- We present a comprehensive analysis of the potential and limitations of GPT-3.5 compared with other generative LLMs and fine-tuned classifiers. We also offer some insights into their efficacy in Persian and other low-resource language contexts.

2. Background and Related Work

This work is at the intersection of Persian NLP, and political ideology detection on social media. Here, we review the related work in each of these areas of research.

2.1. NLP in Persian

Although a large number of people speak Persian (there are approximately 110 million Persian speakers worldwide), very few language resources have so far been developed in this language. Shamsfard (2019) discuss the challenges of studying Persian and the reasons why it should be considered a low-resource language. They emphasize the need for effective solutions to leverage the potential of NLP techniques to create more resources for the automatic processing of Persian data.

There have also been efforts in creating foundational models in Persian, including ParsBERT (Farahani et al., 2021), GPT2-Persian (Khashei, 2021), ALBERT Persian (Farahani, 2020). Besides these general-purpose pre-trained Persian models, ARMAN has been specifically trained for text summarization in this language as well (Salemi et al., 2021). Furthermore, Persian is included in several multilingual pre-trained language models, including mBERT (Devlin et al., 2018), and XLMR (Conneau et al., 2020). Finally, Persian is also included in recently released generative AI models, such as LLaMA (Touvron et al., 2023a) and ChatGPT (Radford et al., 2021), but this language only represents a very small percentage of their training data. Indeed, while numerous studies have explored the application of generative LLMs in various of tasks beyond standard NLP benchmarks (Bandi et al., 2023; Ahuja et al., 2023; Weidinger et al., 2023; Bang et al., 2023), research such as Lai et al. (2023) and Zhu et al. (2023) has specifically evaluated the performance of the GPT-3.5 model in multilingual contexts, including Persian. However, their focus was not on political ideology detection on social media per se. Therefore, to the best of our knowledge, our study represents the first attempt to apply these more powerful models to this task in Persian.

2.2. Domain Background

Ideology Detection on Social Media: Ideology detection in online communities is a dynamic area of research that aims to classify and identify the partisanship or ideological leaning of of social media users (Pelrine et al., 2023; Pennacchiotti and Popescu, 2011; Chen et al., 2017). Thus far, most of this research has been focused on Twitter. For example, Yu et al. (2023) have examined how LLMs and smaller language models can be used to classify Twitter users according to their ideology. Their study involved three datasets, predominantly in English, related to the 2020 US election, the 2021 Canadian election, and COVID-19. They examined the capabilities of Llama 2, GPT-3.5, and RoBERTa, and found that RoBERTa outperformed the other two after fine-tuning. Additionally, they proposed to distinguish between "Explicit ideology" and "Implicit ideology". In this context, "Explicit" refers to classifying users based on their biographical information, which includes obvious ideological identifiers. On the other hand, "Implicit" involves predicting ideologies based on less explicit data, mainly a random set of users' tweets, which are less informative than the bio descriptions. Here, we employ a similar approach and consider classifying users based on their bios when they contain a strong indicator related to their tweets, which are sampled in different ways, as well as their combination.

Social Media and Ideology Analysis in Persian: We find several studies that have attempted to analyze social media activities on Persian Twitter. Notably, the work of Kermani and Tafreshi (2023) used the retweet graph, analyzed the political ideologies of Iranians during the 2017 presidential election and emphasized the significance of social media as a deliberative space for political discussions. Their results confirm that there were three communities active on Twitter during the election: reformists, conservatives, and diaspora. In another related work, Honari and Alinejad (2022), looked at bot activities on Twitter that supported controversial policies in Iran. Kermani and Hooman (2022) shed light on a significant feminist discourse among Iranian Twitter users during the summer of 2020. While the #MeToo movement emerged on this platform in Western countries in 2017, allowing millions of women to share their experiences of sexual abuse and harassment, Iranian users began discussing their own similar experiences on this platform three years later. The results of this study highlight the distinctions between Iranian feminism and its Western counterpart by highlighting the challenges of advocating for women's rights in Iranian society on social media.

Several studies also look at classifying Twitter users according to their ideological leanings, most notably to reveal the level of political change advocated by different political factions in Iran. For instance, Azadi and Mesgaran (2021) categorizes users into three distinct groups: "pro-regime", "dissidents", and "neutral individuals". Their work also focuses on two samples of Iranian Twitter users: the influencers and the ordinary people. They provide various statistical insights about these two samples, such as the age of their accounts, their time zone, and their interactions. They also classify some of the existing ideology clusters by focusing on their level of coordination and how much they are rooting for a regime change. In another work, Kermani (2023) confirm the extensive Twitter engagement of Iranian users in September 2022, despite all of the attempts by the government to impede online activism. Their analyses provide insights into the strategies used by pro-government agents to influence the debates and how the users overcame them. Finally, the work by Khorramrouz et al. (2023) examines the Mahsa Amini movement more specifically through the lens of gender equality. Their research reveals that the movement has intensified the polarization among Twitter users on this issue, with a more pronounced increase among those advocating for gender equality. Moreover, the authors categorize users into 'state-aligned' and 'pro-protest' groups, and argue that the pro-protest users align more closely with the baseline characteristics of Twitter users.

Overall, these studies help us identify the main ideological fault lines in the context of Iranian politics today. On the one hand, the main supporters of the government fall into the 'state-aligned' and 'pro-regime' categories of users. On the other hand, the dissidents encompass the 'monarchist', 'pro-protest', 'pro-women rights', and 'pro-minorities'. Since the other remaining dissidents users belong to a broad spectrum of (evolving and overlapping) ideologies without explicit markers, we have decided to group them in the "Others" category to minimize the risk of mislabeling. In this study, we only focus on the categories of "Pro-Government" and "Pro-Monarchy".

3. Dataset

Starting in September 2022, Persian Twitter users have been increasingly adding political hashtags to their tweets in response to political unrest in Iran. Using the Twitter Research API, we gathered our dataset by collecting real-time tweets between October 18th, 2022, and January 11th, 2023. Our data collection relied on a series of relevant political hashtags, which can be seen in Figure 1. We used 26 seed hashtags for this collection, both in Persian and English, which were identified by the authors who are familiar with the political context in Iran. A total of 231 million tweets were collected from 3.9 million users.

In the next step, the users were sorted by the

Language	Hashtags
Persian	مهسا_امینی، اعتصابات_سراسری، لبیک_یا_خامنه_ای، ایران_قوی
English	opiran, Mahsa_Amini, IranProtests2022

Figure 1: Examples of the hashtags used for crawling the tweets from Twitter.

Group	Keywords
Monarchy-supporters	👑، 🦁، پهلوی ، 🇮🇷
Government-supporters	سید علی ، شهادت ، #حاج_قاسم ، شهید ، ظهور

Figure 2: The indicator keywords used to find the most forthright supporters of groups.

number of times they were retweeted within the dataset, which ranks the more influential users first. Their Twitter biographical information was examined to find out if they were supporting one of the two extreme ideological views included in this study, "Pro-Government" and "Pro-Monarchy".

We define the "Pro-Government" group as users who support the 1979 Islamic Revolution and the Islamic Republic of Iran—the current government in power. The "Pro-Monarchy" group are the users who support the Pahlavi dynasty and the former Imperial State of Iran.

We selected 1000 accounts (500 for each category) using a simple keyword search in their bio information in order to sample users who are likely to belong to either one of the categories of interest. Figure 2 shows some of these indicator keywords. We then ***labeled each user manually*** into three categories: *"Pro-Monarchy"*, *"Pro-Government"*, or *"Others"*. This led to a list of 382 "Pro-Monarchy" users, 316 "Pro-Government" users, and 302 users that could not be classified in those two opposing categories. Furthermore, we filtered out about 10% of the users who had excessively strong ideological keywords on their biographies since we considered them too easy for the classification task.

After this filtering, we are left with 909 users. We split the final dataset to train, validate, and test with the ratio of 0.4, 0.1, and 0.5, respectively. This resulted in 363 train samples, 91 validation samples, and 454 test samples. All the reported results are on the test set. The data collection flow is illustrated in Figure 3.

In Table 1, we included several examples of tweets posted by users classified in each group. To protect the privacy of the Twitter users, we do not include any biographical information in this paper. We also paraphrased and translated the example of tweets shown in Table 1. These messages show that the tweets supporting the monarchy evoke historical symbols and slogans associated with the pre-revolutionary era, such as the "lion and sun" flag and references to "Javid Shah" (Long Live the King), which is directly associated with the Pahlavi dynasty. The government-supportive tweets use language and imagery that reinforce loyalty to the Islamic Republic and its religious leadership, as seen in hashtags such as "#Labbaik_Ya_Khamenei" (I am at your service, Khamenei). The mention of the chador (veil worn by women) also suggest support towards the current government's ideology in reaction to the Mahsa Amini protests. The tweets from other groups also articulate some opposition towards government repression.

4. Experiments

We test several LLMs and fine-tuned LMs to detect extreme political views of Persian Twitter users based on their biographical information and tweet content. This section is divided into three parts: 1. evaluation of GPT-3.5; 2. comparison with other LLMs; and 3. comparison with fine-tuned classifiers.

4.1. Evaluation of GPT-3.5

Here, we assess GPT-3.5, one of the most prominent conversational LLMs, which has gained a significant amount of attention since it was released to the public in 2022 (Radford et al., 2021). As a multi-lingual generative model (Brown et al., 2020), GPT-3.5 includes the Persian (Farsi) language, which constitutes 0.00856% of its training set, corresponding to a corpus totalling 16,731,301 words.[1] This model has been trained on large datasets of conversation data, including social media posts, customer support interactions, and chatbot logs (Dwivedi et al., 2023). It also employs Reinforcement Learning from Human Feedback (RLHF). With RLHF, the feedback obtained from human evaluators is used to train the model further to maximize the reward received when the generated text aligns with human expectations. (Lambert et al., 2022). For all our experiments, we use OpenAI's API with the `GPT-3.5-turbo` (September 15th) model with its temperature set at 0 to ensure reproducible results.

The prompts provided to the LLMs consist of two essential components: the task specification and the associated input data. A question is formulated by defining the specific task that the LLM is expected to perform and by providing the input data relevant for that task. Subsequently, the question and the input are concatenated into a single

[1] https://rb.gy/y2w1t

Table 1: English Translations of Example (Paraphrased) Tweets Across Groups

Group	Translation
Monarchy Supporters	A great slogan that emerged in the heart of Iran, Zahedan: #JavidShah #Mahsa_Amini
	During the Iranian freedom-loving march in Berlin, the lion and sun flag was raised. -Saturday, October 23, 2022 #Mahsa_Amini #IranRevolution
Government Supporters	The chador (veil) you have put on is around the enemy's neck. So hold your chador tighter! #Labbaik_Ya_Khamenei #End_of_Appeasement
	A student who was martyred due to knife attacks by street thugs and hooligans. #Labbaik_Ya_Khamenei #End_Immorality
Other Groups	We will not back down because of the blood you shed and the children you imprisoned. #Mahsa_Amini #Mehrsa_Mousavi
	We are the voice of years of coercion, suppression, and censorship. #Nation-wide_Strikes #OpIran #Mahsa_Amini

prompt, which is then presented to GPT-3.5 for processing. We explain the task design in detail below.

4.1.1. Designing task description

Language: We initially crafted task descriptions for GPT-3.5 in both English and Persian. These task descriptions were developed by the authors who are native speakers or fluent in both English and Persian. To expand our investigation further, we explored the results provided by the model when instructing GPT-3.5 to translate the Persian task description into English and then using that translated task description to execute the task. In this experiment, our goal is to compare prompts written in Persian with those translated into English or originally written in English.

Level of Details: In another set of experiments, we explored different levels of detail in the questions presented to the model. We came up with three settings: a generic question, a more detailed one, and one with an extensive explanation. The first prompt, 'Generic', only provides the labels "Pro-Government", "Pro-Monarchy" and "Others". The second prompt, 'Detailed', provides some context and defines what the two main classes represent, *"Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories."* The third prompt, 'Extensive', complements this with additional information on the "Others" group by adding, *"Group 3 encompasses all other political stances not falling into these two categories and includes secularists and reformists, women's rights activists and Kurdish activists."* The full list of prompts is provided in Table 2.

4.1.2. Input Design

Given that our dataset includes hyper-partisans who express their political leaning in their bio descriptions, we began the experiments by using the user's bio as input. We then revised our input to provide the model with a more comprehensive user context by adding the user's tweets as input. To select the tweets to be added to input, we experimented with different methods namely: 1. 'latest', which includes most recent tweets; 2. 'hashtag', which includes tweets with popular hashtag; and 3. 'retweet', which includes most popular tweets.

To select tweets related to popular hashtags, we arranged a list of hashtags used by the user throughout the dataset's time period, ranking them based on frequency of usage. We then chose one tweet associated with each hashtag, beginning with the most frequently used ones, depending on the desired number of input tweets. In cases where a user had few hashtags, we re-generated the list to meet the desired input quantity. If a user did not have enough popular hashtag tweets, we randomly selected additional tweets until we reached the required number.

To select the users' most popular tweets, we ranked each user's tweets by the number of retweets they received up to that point and selected input from the top of this list. It is important to note that our dataset was collected in real-time, so at the time we collected a particular tweet, it had no retweets. But because of the collective nature of this movement on Twitter, most of the tweets were already retweets of other users' tweets, and the Twitter API returned us the number of retweets the original tweet had up to that point in time. Therefore, in our experiments, by the *"number of retweets"*, we mean the number of retweets the original tweets had received. Finally, we combined the bio description with tweets selected with our best tweet sampling strategy and

Table 2: List of English Prompts for Hyper-partisan Ideology Detection Task

Input Type	Detail Level	Task Description
Bio	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description of a Twitter user which is translated to English and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Bio description: {input_text}*
Bio	Generic	*We are interested in studying political groups on Farsi Twitter. Your task is to analyze the bio description of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the idea of monarchy. Group 2 stands behind the Islamic Republic. Group 3 encompasses all other political stances not falling into the first two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Bio description: {input_text}*
Bio	Extensive	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories and includes secularists and reformists, women's rights activists, and Kurdish activists. Respond with '1', '2', or '3' with no other text or explanation.\n Bio description: {input_text}*
Tweets	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the tweets of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Tweets: {input_text}*
Bio and Tweets	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description and tweets of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation.\n Bio:{bio} \n Tweets:{tweets}*

gave it to GPT-3.5 as the input of the prompt.

4.2. Evaluation of Open-source models

We compared GPT-3.5 with two open-source LLMs for this task. These models are Llama 2 (70B Chat)[2] and WizardLM (70B)[3]. Llama 2 (Touvron et al., 2023b) is a collection of pre-trained generative text models developed by Meta. The scale of this model is from 7 billion to 70 billion parameters. Meta claims instruction-tuned Llama 2 Chat series is optimized for multi-round dialogue and outperforms open-source chat optimized models on most benchmarks. Meanwhile, building upon the original LLaMA (Touvron et al., 2023a) framework, WizardLM (Xu et al., 2023) elevates LLM with additional functionalities. This model is also fined-tuned for chat using AI-evolved instructions.

The first challenge we encountered with open-source models was defining their specific tasks, which required us to provide more detailed instructions on the desired format of the output. Indeed, our evaluation method was required to receive a numerical label for each user from the language model. However, these open-source models did not follow through and responded with sentences instead of numbers. We attempted to use regu-

[2]https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[3]https://huggingface.co/WizardLM/WizardLM-70B-V1.0

lar expressions to extract the labels from the text. However, this approach did not work because the sentences produced by the models were often too complex. We opted to use an alternative approach for fixing this issue that required passing the models' responses through an LLM once more to code the intended label suggested by the text. This step to generate a numerical label was done using GPT-3.5, which turned out to be remarkably effective.

Two experiments were conducted on these LLMs: the first one involved using the bio as input for the detailed task description; the second one was a detailed task description with the translation of bios to English as input. We translated the bios using GPT-3.5. For the translation task, we use the prompt "*Here is the bio of a user's Twitter account. Translate it into English. Please respond with only the translation and no further explanation. \n Twitter bio: {input_text}*"

Both of the models were run with the temperature set to 0 and with vLLM (Kwon et al., 2023), a framework which uses the PagedAttention to optimize the utilization of GPU memory and improve performance.

4.3. Evaluation of fine-tuned classifiers

Another method that we believe is useful in ideology prediction with textual data implies fine-tuning the classifiers. In this research, we employed two classifiers from the BERT family: ParsBERT for the Persian language and RoBERTa for the English language. We also ran preliminary experiments on multilingual BERT (m-BERT), but the results were much worse than those of RoBERTa and ParsBERT, so we did not continue with this model.

The two input designs that showed the most promising results on GPT-3.5 were given to classifiers separately. These input designs include: 1. bio description of the users; and 2. 20 tweets selected based on the most used hashtags of the users. We then used the data translated by GPT-3.5 to work with RoBERTa and fine-tuned the language models with the following hyper-parameters: ParsBERT with *batch size*=16, *learning rate*=0.0001, *warm-up steps*=0, *weight decay*=0.205, *epochs*=3, and RoBERTa with *batch size*=16, *learning rate*=0.00008, *warm-up steps*=100, *weight decay*=0.251, *epochs*=4.

5. Results and Discussions

Our initial experiments aimed to assess the performance of GPT-3.5 with various approaches. In our first experiments, we observed that GPT-3.5 failed in this task when prompted with a few-shot strategy. This explains why we decided to adopt a zero-shot prompting strategy for the rest of the paper.

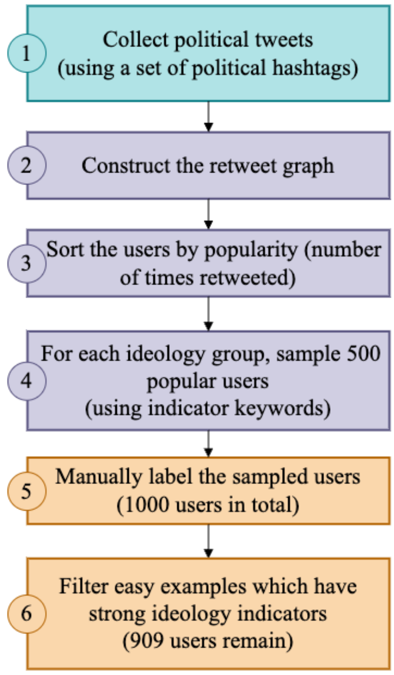


Figure 3: Dataset construction process

Prompt	F1	Accuracy
English	**0.72**	0.70
Persian	0.67	0.67
Translated (Fa to En)	0.70	0.69

Table 3: The results of GPT-3.5 in different task description languages.

First, we prompted GPT-3.5 with the 'Generic' task description in Persian, English, and GPT-3.5's translation of the Persian prompt to English. The inputs in these experiments are Persian bios. Table 3 shows the results for this set of analyses. We can see that the model performs better when prompted in the English language compared to the Persian prompt or the translated prompt. This last result explains why we continue the experiments with English prompts.

We then experimented with different level of details in the prompts. As shown in Table 2, the 'Detailed' prompt and 'Generic' prompt differ in the context provided to the model and the explanation of the first and second groups. Results are presented in Table 4. The two prompts display similar f-scores for the bio input, but the generic prompt outperforms the bio and tweets combination. We can see that the gap between the 'Detailed' prompt and the 'Extensive' prompt is more significant on the bio and tweets input. For the remainder of the analyses, unless stated otherwise, all experiments in this study use the 'Detailed' task description.

We subsequently experimented with different input design strategies the results of which are shown in Table 5. We observe that the best method for

	Bio		Bio and Tweets	
Prompt	F1	Accuracy	F1	Accuray
Generic	0.72	0.71	0.56	0.59
Detailed	**0.72**	0.70	**0.67**	0.68
Extensive	0.69	0.68	0.60	0.63

Table 4: The effect of the level of details provided in the prompt.

Bio	Tweets	Count	F1	Accuracy
✓	✗	-	**0.72**	0.70
✗	latest	5	0.44	0.51
		10	0.46	0.52
		20	0.45	0.52
✗	hashtags	5	0.52	0.57
		10	0.51	0.56
		20	0.52	0.57
✓	hashtags	5	0.62	0.63
		10	0.66	0.67
		20	**0.67**	0.68
✗	retweet	5	0.51	0.54
		10	0.52	0.54
		20	0.51	0.55
✓	retweet	5	0.65	0.66
		10	0.66	0.67
		20	0.64	0.64

Table 5: The effect of input in the GPT-3.5 response.

choosing the most informative tweets is associated with choosing the most used hashtags of the user. Considering that we started by scraping tweets with related hashtags for data collection, this strategy could capture the context more effectively than others, such as popular tweets or latest tweets. The best number of tweets included in the input would appear to be between 10 or 20, depending on the method of choosing tweets. However, using only the bio is still more effective.

Table 6 shows the results of our experiments on LLMs, Llama 2 Chat and WizardLM, which demonstrate that GPT-3.5 outperforms these models in our task. The performance of both LLMs is improved when provided with English translations of bios rather than the original Persian versions. WizardLM is outperforming Llama 2 on Persian, but GPT-3.5 still has a significant lead.

Model	Bio	F1	Acc
GPT-3.5	Original	**0.72**	0.70
	Translated	**0.77**	0.76
Llama 2	Original	0.42	0.45
	Translated	0.71	0.71
WizardLM	Original	0.63	0.62
	Translated	0.71	0.70

Table 6: Open source LLMs versus GPT-3.5.

bio		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.72	0.70
Fine-tuned ParsBERT	0.81	0.82
Fine-tuned RoBERTa	**0.86**	0.87
bio + tweets		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.67	0.68
Fine-tuned ParsBERT	**0.86**	086
Fine-tuned RoBERTa	0.85	0.85
tweets		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.52	0.57
Fine-tuned ParsBERT	0.80	0.81
Fine-tuned RoBERTa	**0.85**	0.85

Table 7: Comparison of the fine-tuned language models and GPT-3.5 across different inputs.

Table 7 include the comparison between GPT-3.5 and the fine-tuned classifiers. These results indicate that fine-tuned models outperform GPT-3.5 on the user classification task. RoBERTa shows the best performance with an f1 score of 0.86, while provided with the translation of the bio description. ParsBERT shows the same f1 score when provided with bios and 20 tweets that are chosen by the most-used hashtags of the user, which is the best setting of Table 5. All the results reported in Table 7 which include tweets have the same setting; that is, they correspond to the results of fine-tuned classifiers when trained and tested with only 20 tweets selected based on hashtags. Also, all of the reported f1 scores in the tables are weighted f1 scores.

From Table 7, we also see that including tweets in the input boosts ParsBERT's performance but adversely affects RoBERTa's performance. We observe that higher accuracy in tweet classification is achieved when tweets are translated to English, and a RoBERTa model is fine-tuned, rather than when using a fine-tuning ParsBERT with Persian tweets directly. This indicates that using RoBERTa on translated text for identifying the ideology of users results in the best performance, especially when the model is only based on their tweets.

Since RoBERTa demonstrates superior performance in Table 7, we conducted an additional qualitative analysis to better understand its performance compared to GPT-3.5. This involved reviewing instances where each model struggled to identify discrepancies in their predictions. Our analysis indicates that GPT-3.5 lacks context awareness, leading to incorrect predictions, even when familiar symbols or mottos of the Persian political context are involved. Conversely, in our analysis, RoBERTa appears to struggle with detecting sarcasm, a common element in Twitter communications. Finally, we also note that GPT-3.5's refusal to translate offensive language could imply that RoBERTa is working with less information due to translation losses.

Finally, we ran some additional experiments with the `gpt-4-1106-preview` model as well. Its performance on bios with the detailed task description is considerably better than GPT-3.5 with weighted f1 score and an accuracy level of 0.83. The performance is also improved with the bios that were translated to English using GPT-3.5. In this experiment, the f1 score and accuracy of GPT-4 is 0.82. Finally, the performance also improves when we translate the bios to English with GPT-4. This time we find an f1 score and accuracy of 0.84. While these results are better than GPT-3.5, we did not perform the experiments of this study on GPT-4 because of the expensive price of this model which makes it less practical.

6. Conclusions and Future Work

This study explored the application of LLMs for political ideology detection in the context of Persian Twitter users. Our results confirmed that the best approach to classifying ideology on Persian Twitter is to fine-tune a ParseBERT model with a combination of user biographies and tweets with the most popular hashtags. A RoBERTa model fine-tuned with translated biographies results in the same f-score, but the added cost of translation makes this approach less practical. However, there are several important limits to our analysis.

We acknowledge that this task is much more complex than similar works conducted in English. First, unlike in democratic countries, the broad spectrum of political views beyond the hyper-partisan ideologies is not well-defined in Iran. Second, Persian is a low-resource language, and LLMs are expected to perform worse in this language than in English. For these reasons, we limited our study to a computationally simplified task of hyper-partisan ideology detection by defining our ideological groups according to two extreme views: one that supports the Islamic government; and another that calls for a return of the overthrown monarchy. All other remaining opposition ideological groups were categorized in an "Others" for this analysis.

We evaluated `GPT-3.5-turbo` in different settings and showed that even in this simple computational task, while GPT-3.5 offered convincing results, it significantly performed worse than specialized models, such as fine-tuned RoBERTa and ParsBERT. These results highlight the importance of language-specific models for computational tasks that involve contextual nuances in a non-English space. Our results also confirm that investing in benchmark datasets to evaluate LLMs in non-English languages and non-standard tasks is extremely important. These datasets are crucially important for understanding the capabilities of LLMs; they are also necessary to develop specialized models that address the diverse needs of non-English speakers.

In future work, we intend to explore the landscape of ideology groups within the "Others" category. Given the diverse range of perspectives and the presence of numerous subgroups with complex boundaries within this category, we anticipate the need for a combination of unsupervised and supervised methodologies to effectively map and understand these varied ideological views.

Ethics Statement

We used the Twitter Research API to collect tweets for this study. In order to comply with Twitter's policies and to respect the users' privacy, we will not make the labelled dataset publicly available. However, our data collection methodology can be used by other researchers to explore similar tasks and scenarios.

There is a risk of mislabelling when users are labelled for political ideology based on their social media activities. Here, we mitigate this risk by labeling users that belong to two hyper-partisan groups ("Pro-Government" and "Pro-Monarchy") based on the explicit ideology identifiers found in the users' bio descriptions. Users without these identifiers were categorized as "Others". This implies that users labeled as "Pro-Government" and "Pro-Monarchy" self-identify with these classes publicly and actively engage in political discussions. We do not release any personally identifiable information for any of the users we studied.

Political ideology detection can be potentially misused by malicious actors to influence users, interfere in other countries' elections, or spread misinformation on social media. We emphasize that this task should never be employed to enable targeting of specific users. But there is no security by obscurity here. To counter such malicious uses, it is critical to develop strategies that reduce the spread of misinformation and extreme polarization, minimize the impact of bots, and promote safe and healthy online environments—tasks for which understanding ideology is essential (Pelrine et al., 2023; Tucker et al., 2017). Therefore, by minimizing risks through measures discussed above, such as not releasing identifiable data, political ideology detection research is beneficial to society.

Finally, it is important to note that GPT-3.5 is a closed system with unknown training data and strategies and frequent updates. Because of these factors, it is difficult to fully analyze and contextualize our results. Furthermore, these results may not remain valid for future versions of the model. However, this study is an important initial effort to explore the capabilities and limitations of general-purpose generative systems compared to fine-tuned supervised models for low-resource non-English languages, specifically the Persian language.

7. Bibliographical References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Pooya Azadi and Mohsen B. Mesgaran. 2021. The clash of ideologies on persian twitter. Working Paper 10, Stanford Iran 2040 Project, Stanford University.

Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. 2023. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *International Joint Conference on Artificial Intelligence*.

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

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

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. “so what if chatgpt wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

ERFI. 2023. The political stance in Iran. `https://erf.institute/stance`.

Mehrdad Farahani. 2020. Albert-persian: A lite bert for self-supervised learning of language representations for the persian language. `https://github.com/m3hrdadfi/albert-persian`.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53:3831–3847.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Rouzbeh Ghasemi, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. 2022. Deep persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, 48(4):449–462.

Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. 2016. Ideology Detection for Twitter Users with Heterogeneous Types of Links. *arXiv e-prints*, page arXiv:1612.08207.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ali Honari and Donya Alinejad. 2022. Online Performance of Civic Participation: What Bot-like Activity in the Persian Language Twittersphere Reveals About Political Manipulation Mechanisms. *Television & New Media*, 23(8):917–938. Publisher: SAGE Publications.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2022. Retweet-bert: Political leaning detection using language features and information diffusion on social networks. *ArXiv*, abs/2207.08349.

Hossein Kermani. 2023. #MahsaAmini: Iranian Twitter Activism in Times of Computational Propaganda. *Social Movement Studies*, 0(0):1–11. Publisher: Routledge _eprint: https://doi.org/10.1080/14742837.2023.2180354.

Hossein Kermani and Niloofar Hooman. 2022. Hashtag feminism in a blocked context: The mechanisms of unfolding and disrupting #rape on persian twitter. *New Media & Society*, 0(0):14614448221128827.

Hossein Kermani and Amirali Tafreshi. 2023. Walking with bourdieu into twitter communities: an analysis of networked publics struggling on power in iranian twittersphere. *Information, Communication & Society*, 26(8):1653–1674.

Afshin Khashei. 2021. A not-so-dangerous ai in the persian language.

Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2023. For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles. ArXiv:2307.03764 [cs].

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. Https://huggingface.co/blog/rlhf.

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

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

Ammar Maleki. 2022. IRANIANS' ATTITUDES TOWARD POLITICAL SYSTEMS: A 2022 SURVEY REPORT.

Kellin Pelrine, Anne Imouza, Zachary Yang, Jacob-Junqi Tian, Sacha Lévy, Gabrielle Desrosiers-Brisebois, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, and Reihaneh Rabbany. 2023. Party prediction for twitter.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):281–288. Number: 1.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Ilya Sutskever, Rewon Child, Gretchen Krueger, and Jong Wook Kim. 2021. Chat with gpt: Improving language generation and task-oriented dialogue. `https://openai.com/blog/chatgpt-plus`.

Miguel Ángel Rodríguez-García, Soto Montalvo Herranz, and Raquel Martínez Unanue. 2022. Urjc-team at politices 2022: Political ideology prediction using linear classifiers.

Nick Rogers and Jason Jones. 2021. Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time. *Journal of Social Computing*, 2.

Alireza Salemi, Emad Kebriaei, Ghazal Neisi Minaei, and Azadeh Shakery. 2021. Arman: Pre-training with semantically selecting and reordering of sentences for persian abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9391–9407.

Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International Conference Language Technologies for All (LT4All)*.

Heydar Soudani, Mohammad Hassan Mojab, and Hamid Beigy. 2022a. Persian natural language inference: A meta-learning approach. *arXiv preprint arXiv:2205.08755*.

Heydar Soudani, Mohammad Hassan Mojab, and Hamid Beigy. 2022b. Persian natural language inference: A meta-learning approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4306–4319, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Petter Tornberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Verity Trott. 2021. Networked feminism: counterpublics and the intersectional issues of #metoo. *Feminist Media Studies*, 21(7):1125–1142.

Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *Journal of democracy*, 28(4):46–59.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. ArXiv:2304.12244 [cs].

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification?

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Author Index