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

Raphaël Merx[ɕ] **Aso Mahmudi**[ɱ] **Katrina Langford**[ʇ]
Leo Alberto de Araujo[ʃ] **Ekaterina Vylomova**[ɱ]
[ɕ]Catalpa International [ɱ]The University of Melbourne [ʇ]Timorlink [ʃ]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/

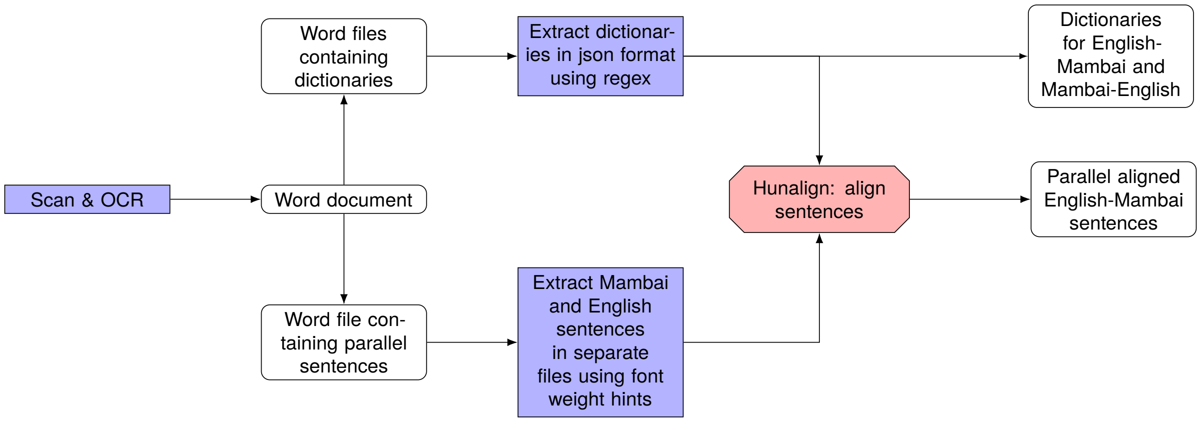

Figure 1: Overview of our process for extracting dictionaries and a parallel corpus from the Mambai Language Manual

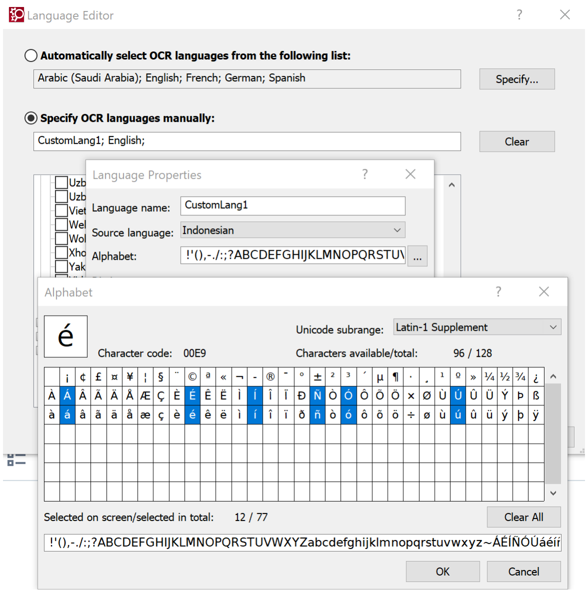

Figure 2: Mambai configuration in ABBYY FineReader 15.

(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:
```

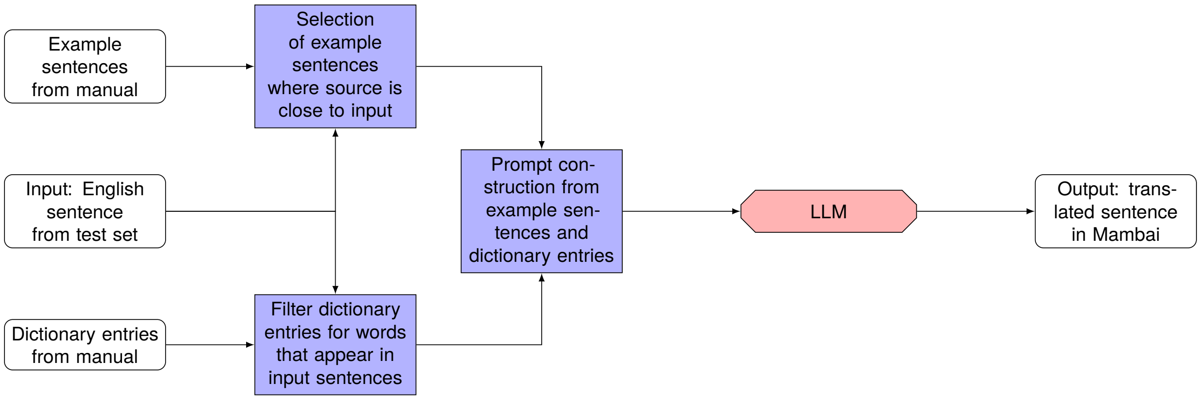

Figure 3: Overview of our process for translating English sentences to Mambai using both dictionary entries and sentence pairs in few-shot LLM prompting.

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 pre-training 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*.