CSC 372 Final Project
Integrating Toki Pona to SMaLL-100 for Enhanced Natural Language Understanding

Ilina Navani and **Katie Kowalyshyn**
{ilnavani,kakowalyshyn}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

In this paper, we extend SMaLL-100, a multilingual machine translation model built for low-resource languages, to Toki Pona. Toki Pona is an artificial language constructed by Sonja Lang, a linguist, to experiment with the number of words needed to express the same meaning. By fine-tuning SMaLL-100 using a parallel corpus, we were able to create a model capable of translating language from Toki Pona to English. We found that our model performed poorly according to the BLEU score and often mixed up gender as well as subjects in translations. However, the model was able to generally conserve meaning and produce semantically correct translations, thereby achieving reasonable performance in its task of machine translation.

1 Introduction

This project began with a simple goal to examine the effects of languages with data scarcity issues and language expansion errors, and turned into a model extension project. Languages with a lower number of speakers naturally have less data to support machine translation projects. As opposed to English, most other languages aren't often colloquially labeled as world languages and instead have far less literature and internet data available for machine translation research. Hence, as Natural Language Processing (NLP) research has progressed, it's become natural to use English as a base language for new models and techniques. In recent years, the trend has been towards a research field which demands more and more data for precision of machine translation, yet there remain languages where data scarcity prove to be a roadblock. Hence, our goal was to fine-tune a model to perform translation into English from a language with less data as well as complexity of grammar, at which point we came across Toki Pona.

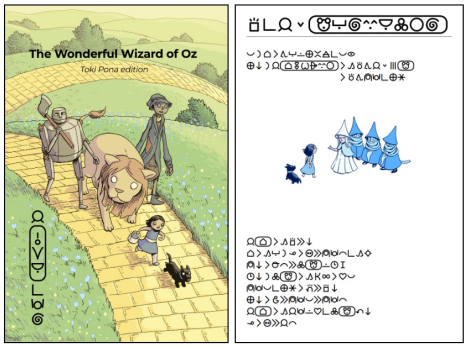

Figure 1: The Wonderful Wizard of Oz (Toki Pona edition) translated by Sonja Lang, illustrated by Evan Dahm (Lang 2014)

Toki Pona is a language created in 2001 by Canadian linguist Sonja Lang to "understand the meaning of life in 120 words" (Lang 2001). This language's grammar is explained in eighteen minutes via the website's linked YouTube video and even has its own Discord channel featuring 5000+ speakers. In 2022, ISO 639-3 adopted the code "tok" for Toki Pona, making it a world language. In Figure 2, readers can see the latin alphabet equivalents for Toki Pona letters.

Figure 2: Latin Alphabet Chart for Toki Pona (Lang 2014)

Toki Pona literally translates to the "language of good." Its origins, though artificial, have expanded drastically. Though one is able to write in Toki Pona using only fourteen letters from the latin alphabet, it is also a glyph-based language. The latin writing system remains the most common form of Toki Pona communication, while two other forms of logographic writing systems are also common.

The first of these systems was created by Sonja Lang and is called "Sitelen Pona." Its glyphs each represent a word of the language, and can be seen in Figure 3.

Figure 3: Sitelen Pona Glyphs
(Lang 2014)

Adjectives are written inside or above the noun's symbol itself. Hence, as seen in Figure 4, Toki Pona's language symbol is the symbol for "toki" with a line inside of the glyph to represent the adjective "pona."

Figure 4: Toki Pona Symbol
(Lang 2014)

The second logographic writing system is sitelen sitelen, created by Jonathan Gabel (Gabel 2021). This style is more complex than Sitelen Pona, and includes logograms representing words and an alphasyllabary for writing syllables. While this paper will not go into depth about this writing expression, readers are referred to Jonathan Gabel's website to learn more about its roots (Gabel 2021).

Toki Pona functions as a form of communication as any other natural language does, despite its artificial roots. Its lack of data in comparison to English and other commonly spoken languages isolated it as a perfect case study for our investigation.

Thus, our research questions were:

1. How does Toki Pona function as a source language when targeting English translation?
2. To what extent does amount of training data affect the accuracy of machine translation?

The remainder of our paper will provide further context to the datasets used, the expansion of the machine translation model, and results. We will conclude by identifying potential areas for further research.

2 Background

Corpora

To gather sufficient data, we sought out two parallel corpora. The first corpus came from Tatoeba, an online platform with thousands of example sentences and their translations into various languages. The Tatoeba dataset provides 2615 parallel sentences in English and Toki Pona. The second corpus, obtained from Hugging Face, followed a similar structure, but was much larger with 21,030 parallel sentences in each language. Hence, by combining both datasets we were able to generate a parallel corpus with 23,645 sentences in each language.

Toki Pona's grammatical structure is far more expansive than English's grammar. In an effort to reduce the complexity of the language via a lower word count, Sonja Lang's creation expresses more complex thoughts via a string of simpler words. For example, a sentence that may take only a few words in English will inevitably take more words in Toki Pona and simultaneously lose some meaning. A simple example of this is the translation of "They construct a building", which in Toki Pona is "ona li pali e tomo." As sentences get more complex, this language expansion becomes exponentially more difficult. Rather than expanding a sentence to be irrationally long, Toki Pona speakers subjectively choose adjectives of the thing they're describing in order to get their point across. While some people may identify a bunny's cuteness as most important in a story, another may identify a bunny by their large ears. This naturally complicates translation, as the number of equivalencies for just one word, "bunny" becomes much more difficult to identify.

Ultimately, our choice to use Toki Pona came down to an interest in artificially-created language as well as the research question of studying low-resource languages. Our choice to use SMaLL-100 as our machine translation model came from its creation for low-resource languages and the

ability to expand it to new languages like Toki Pona.

Machine Translation Model
SMaLL-100 covers more than 10k language pairs and 94 languages. It was designed to compete with M2M-100 while being smaller and faster. Through Hugging Face, this model functions as an extension of Transformers and uses the sentencepiece package. For an evaluation metric, SMaLL-100 uses spBLEU and the SentencePiece tokenizer. The model has been very successful in translation tasks, leading to two published papers discussing the creation of the model and its performance. (Mohammadshahi et al. 2022b) (Mohammadshahi et al. 2022a). SMaLL-100 is trained with uniform sampling across all language pairs, allowing it to focus on the performance of low-resource languages (Mohammadshahi et al. 2022a). Additionally, the papers expand on compression for Multilingual Neural Machine Translation models (MNMT) for various language groups, gender, and semantic biases, and ultimately find that compression amplifies gender and semantic biases (Mohammadshahi et al. 2022b). Interestingly, this examination of compressed MNMTs is not exclusive to low-resource languages, but also expands to high-resource languages. When deciding how to implement Toki Pona to this model, we used the same evaluation metric as the publsihed papers to maintain comparable results.

3 Experiments

Training
When fine-tuning SMaLL-100 for translation from Toki Pona to English, we first split our dataset such that 80% was used to train the model and 20% was held out for testing. As a result, our model was trained on a total of 18,916 parallel sentences in each language. We passed in source and target sentences in batch sizes of 16, given the computational resources available to us. We used the Adam optimizer with a learning rate of 0.0001, thereby allowing us to perform backpropagation and update model parameters to decrease loss throughout training.

Testing
Once training was complete, we tested performance by comparing translated sentences from our model to reference sentences from the corpus. We fed the fine-tuned model a test set of 4729 Toki Pona sentences, generating an equivalent number of predicted English translations. We were then able to calculate BLEU and METEOR scores to compare the similarity of the predictions to corresponding reference sentences in the test set.

Evaluation
The BLEU score is a common evaluation metric to determine the similarity of machine-translated text to a set of high quality reference translations. It does so by exact word matching, that is, looking at the number of common words between two sentences. The score is a number between 0 and 1, and the closer the number is to 1, the more accurate the translation. Python's NLTK library has a built-in module to calculate BLEU scores, which is what was used in this project.

Since BLEU only measures direct word-by-word similarity without considering semantic or syntactic similarity, we decided to implement a second evaluation metric, METEOR score. This score assess the quality of machine translation by evaluating overall alignment between the translated text and the reference text. It considers the order in which words appear as it penalizes the results having incorrect syntactical orders, however it pays attention to synonyms, stems, and paraphrases. Therefore it can recognize translations that use different words or phrases while still conveying the same meaning as the reference text. Our calculation of METEOR scores utilized a built-in Python module within the Hugging Face evaluate library.

4 Results

Our test set produced an average BLEU score of 0.086, indicating very poor performance. This was expected, given the fact that BLEU scores rely highly on the precision of a translation. As explained earlier, a high BLEU score is only achieved if there is an exact match between words in the translated text and reference text, regardless of whether the two sentences express the same meaning using different words. Upon a closer examination of our results, we realized that the model was indeed performing well in terms of conveying meaning. While specific words did not always align, overall meaning was generally conserved across both sentences, as seen through the examples in Figure 5.

The METEOR scores were significantly better, resulting in an average score of 0.5. Figure 5 shows how semantically similar sentences – such as 'Smith is alive and wants to see you' and 'A Mr. Smith has come to see you' – have higher METEOR scores than other sentence pairs that are less similar. Hence, although the predictions were not perfectly aligned with their targets, the model was able to learn translation from Toki Pona to English with reasonable performance.

```
Translated sentence:  ['He heated my picture.']
Reference sentence:  She burned my photo.
BLEU score:  1.2882297539194154e-231
METEOR Score:  0.20000000000000004

Translated sentence:  ["It's going to rain."]
Reference sentence:  The clothes got wet.
BLEU score:  0
METEOR Score:  0.09803921568627452

Translated sentence:  ['Smith is alive and wants to see you.']
Reference sentence:  A Mr Smith has come to see you.
BLEU score:  4.797597231912944e-78
METEOR Score:  0.5377777777777778

Translated sentence:  ['If the sun rose, the wind blossom is gone.']
Reference sentence:  The fog dissipated as the sun rose.
BLEU score:  6.08970970641905e-155
METEOR Score:  0.4481927710843373
```

Figure 5: Examples of Results

We hypothesize that one reason for imperfect performance is the batch size. Given the GPU space available

to us, we were able to train our data using a maximum batch size of 16. As a result, the model could be overfitting, since parameters are being updated based on a smaller subset of the training data. We were also limited by the amount of training data available to us, since Toki Pona does not have very large publicly available corpora. Additionally, our model was of course built for its ninety-four original languages, meaning that Toki Pona was not part of the original architecture. This naturally limits Toki Pona translations to any language. Extending a model has its limits, and while the model itself may be comparable to M2M-100, our adaptation of the model is not close to achieving the same results. Hence, our project was successful from a completion standpoint but not entirely from a research perspective.

5 Conclusions

Our goal to fine-tune the SMaLL-100 model was successful, though our results weren't necessarily emblematic of a very accurate translation model. Our research questions were answered to an extent, showing that more training data certainly helped increase the translation quality and that Toki Pona functions adequately as a source language, although its linguistic differences to English remain unsolved after this research. Moving forward, a research project could investigate specifically the language expansion distinctions between Toki Pona and English, or to expand this model further to compare Toki Pona translations with languages other than English. SMaLL-100 serves as a tool for future translation endeavors, using NLP tools to investigate linguistic differences via translation.

6 Contributions

We were both able to offer input on every aspect of the assignment and complete the code as well as the paper during our in-person meetings. We each proof-read the entire document as well.

References

Gabel, J. 2021. sitelen sitelen.

Lang, S. 2001. Toki pona. `https://tokipona.org/`. Accessed: April 28, 2024.

Lang, S. 2014. *Toki Pona : the Language of Good*. Marston Gate: Great Britain.

Mohammadshahi, A.; Nikoulina, V.; Berard, A.; Brun, C.; Henderson, J.; and Besacier, L. 2022a. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8348–8359. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Mohammadshahi, A.; Nikoulina, V.; Berard, A.; Brun, C.; Henderson, J.; and Besacier, L. 2022b. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4308–4329. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.