Exploring Bilingual Word Embeddings for Hiligaynon with Geographical and Linguistic Context

Ria Leilani Baldevia

[1] Student, King's College London, UK
ria_leilani.baldevia@kcl.ac.uk

Abstract. Previous research by Michel et al (2020) explored bilingual word embeddings for Hiligaynon, a low-resource Philippine language, showing promise in improving downstream NLP tasks. This paper builds upon that work by creating and evaluating bilingual word embeddings between Hiligaynon and Cebuano, a major language in the Philippines. Cebuano, while a major language, has few linguistic resources available when compared to English and German languages utilized in Michel et al. Hiligaynon, being geographically proximate to Cebuano, shares lexical, syntactic, and semantic properties that may benefit bilingual embedding models. This research trains bilingual skip-gram word embeddings on Hiligaynon-Cebuano parallel data consisting of Bible, Opus data, Wikipedia, and other literary translations. The embeddings are evaluated both intrinsically through word similarity tasks and extrinsically through part-of-speech tagging and named entity recognition. The final submission will address findings and a discussion focused on the implications for low-resource language preservation. This research also highlights opportunities for future work in expanding resources for Philippine languages through cross-lingual embedding techniques. This work contributes to the promising discourse of expanding computational capabilities for morphologically complex, under-studied languages.
.

Keywords: NLP, word embeddings, Hiligaynon

1 Introduction

The rapid development of artificial intelligence (AI) and Large Language Models (LLM) has led to major advancements in natural language processing. In particular, the creation of word embedding models that capture semantic and syntactic information has been instrumental. Mikolov et al. [1] introduced an efficient method for learning distributed vector representations of words called word2vec. Meanwhile, Pennington et al. [2] proposed GloVe word vectors using matrix factorization. Going beyond words, Bojanowski et al. [3] incorporated sub-word information into fastText embeddings. More recently, contextualized representations like ELMo [4] and bidirectional transformer models in BERT [5] have achieved state-of-the-art results on various language tasks by incorporating context.

1.1 The Case of Hiligaynon

Hiligaynon is an Austronesian language spoken mainly in the Western Visayas region of the Philippines. It is the majority language of Panay island and Negros Occidental province; both in the central region of the Philippines. Hiligaynon has about 7 million native speakers and is also spoken in other parts of the Philippines by an additional 2 million second-language speakers. The Hiligaynon language is believed to have developed from the Kinaray-a language spoken on Antique island. As migration and trade spread, the use of Hiligaynon to neighboring islands like Panay and Negros over the past few centuries, it gradually grew. Today, it is the predominant language across Panay and Negros islands. Hiligaynon serves as the major language for the Western Visayan region for both everyday conversation and literature.[6]

2 Related work

This paper was inspired by the 2020 work of Michel et al. [7] where they released a English-Hiligaynon lexicon for further work on the role of under-resourced languages in the digital spaces. Their work shows data limitations still hinder high-quality multiword and bilingual embeddings. The data-hungry nature of current models means small monolingual corpora restrict learning, so techniques to overcome limited resources merit exploration. No one model suits all languages, as syntax and concepts differ across families. Moreover, seed lexicon frequency played an insignificant role in translation mining via BLI. In fact, reducing the seed lexicon, an inexpensive resource, improved results. With just 3,000 terms and 25 million corpus words, Skipgram (29.8%) outperformed a prior 100-million word, 5,000 term study (27.1%). Overall, using a simple model, their research provided insights into factors impacting translation pair mining, from corpus and seed lexicon size to dimensionality's effect on word2vec and fastText. Research into under-resourced languages, such as Hiligaynon, is very few; and the work Michel et al. is one of the few current published academic work on Hiligaynon.

3 Approach

The methodology consists of four key components: gathering parallel data from multiple sources; preparing the data for analysis; developing and training models; and testing and evaluating the models. The research replicates the approach of Michel et al.; however, the language of German is replaced with Cebuano, a major language in the Philippines that is more geographically and contextually appropriate compared to the Western languages used in the previous research. The research leverages the experiment by Braun et al. [8] This paper attempts to identify which corpora size does the performance become comparable to that achieved by Braune et al.

3.1 Training Data

The research leverages two sets of data from the OPUS project. [9] The corpus yielded the number of extracted instances for each language pair, as displayed in Table 1.

Table 1. Language corpora data from OPUS project

Language Pair	**Instances**
English-Cebuano	728,739
English-Hiligaynon	428, 302

3.2 Test Data

A sample of instances from the Opus language pair corpora in Table 1 was used for validation and testing.

4 Evaluation & Findings

The evaluation is currently underway and if this working paper is accepted for further consideration, the final paper will have the evaluation and findings documented.

5 References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
2. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
3. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.
4. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proc. of NAACL.
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
6. Hiligaynon [Hiligaynon]. (n.d.). In Ethnologue. Retrieved January 9, 2023, from https://www.ethnologue.com/language/hil
7. Michel, L., Hangya, V., Frase, A. (2020). Exploring Bilingual Word Embeddings for Hiligaynon, a Low-Resource Language. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) (pp. 2573-2580). Marseille, France: European Language Resources Association (ELRA)
8. Braune, F. (n.d.). BWEeval. Retrieved from https://github.com/braunefe/BWEeval

9. Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214– 2218