Exploring Word Alignment Towards an Efficient Sentence Aligner for Filipino and Cebuano Languages

Jenn Leana Fernandez and Kristine Mae Adlaon
College of Computer Studies
University of the Immaculate Conception
Father Selga Street, Davao City
{jfernandez_190000001847,kadlaon}@uic.edu.ph

Abstract

Building a robust machine translation (MT) system requires a large amount of parallel corpus which is an expensive resource for low-resourced languages. The two major languages being spoken in the Philippines which are Filipino and Cebuano have an abundance in monolingual data that this study took advantage of attempting to find the best way to automatically generate parallel corpus out from monolingual corpora through the use of bitext alignment. Byte-pair encoding was applied in an attempt to optimize the alignment of the source and target texts. Results have shown that alignment was best achieved without segmenting the tokens. Itermax alignment score is best for short-length sentences and match or argmax alignment score are best for long-length sentences.

1 Introduction

Word alignment is the task of discovering the corresponding words or terms in a bilingual sentence pair (Steingrímsson et al., 2021). Word-aligned corpora are a great source of translation-related knowledge. The estimation of translation model parameters usually relies heavily on word-aligned corpora (Liu et al., 2010). Therefore, the alignment of words is a crucial stage in the process of building a machine translation system (McCoy and Frank, 2017).

Sentence alignment is the task of aligning sentences in a document pair (Luo et al., 2021) or in a parallel corpus. In most cases, these sentence pairs share the same meaning or are contextually translated. Abundance of these parallel corpora is very evident for highly resourced languages while the collection and even building of such parallel corpus for low-resourced language is very difficult and is a very tedious task (Callison-Burch et al., 2004). The problem of aligning words in massively parallel texts containing hundreds or thousands of languages remains mostly unexplored (Östling, 2014) and that includes the Filipino and Cebuano languages .

The Philippines has a scarcity of language resources, particularly parallel corpora. Several studies have been conducted in an attempt to build a parallel corpus involving Philippine languages and most of them are paired with the English language (Michel et al., 2020; Ponay and Cheng, 2015; Lazaro et al., 2017). The study of Adlaon and Marcos (2019) had focused on the collection and building of both monolingual and parallel corpus for Filipino and Cebuano to build an NMT System (Adlaon and Marcos, 2018) for the said languages. The abundance of the collected monolingual corpora for the said language pair presents an opportunity for it to be transformed into a parallel corpus using an aligner.

Cebuano and Filipino are the two most spoken languages in the Philippines where the structure of these languages are morphologically-complex. Filipino language is flexible when it comes to word order. In fact, some Filipino sentence can be rearranged up to 6 different ways since sentence structures like SVO, VSO, VOS are accepted. While the Cebuano language, it is said to be predicated which explains why it follows the VSO format (Tariman, 2010). These languages can contribute and get benefits from our existing technology in different aspects, especially for machine translation. Translation studies and contrastive linguistics rely heavily on parallel corpora which are crucial for developing high-quality machine translation systems (Bañón et al., 2020; CLARIN, 2022). The transformation of the available monolingual corpus would be an addition to the existing Filipino-Cebuano parallel corpus. To date, the checking of the translation and the generation of parallel corpus is done manually which is a very laborious and tedious task especially when you have hundreds of thousands of sentences.

To the best of our knowledge, there is still no

Proceedings of the 29th International Conference on Computational Linguistic, pages 99–106
October 12–17, 2022.

word and sentence aligner effective for Cebuano and Filipino. In this paper, the researchers aim to conduct a preliminary investigation on the use of a word aligner for Cebuano and Filipino languages towards the development of an efficient sentence aligner and evaluate its performance in accordance to some ground truth.

2 Related Works

There have been different word alignment approaches that are widely used especially for machine translation. This section discusses the related studies of word and sentence alignment for machine translation.

The study of Kumar et al. (2007) describes a method for improving Statistical Machine Translation (SMT) performance in multiple bridge languages by leveraging multilingual, parallel, sentence-aligned corpora. Their solution includes a simple way for creating a word alignment system using a bridge language and a mechanism for integrating word alignment systems from various bridge languages. The researchers provide studies that show how this framework can be used to improve translation performance on an Arabic-to-English problem by using multilingual, parallel material in Spanish, French, Russian, and Chinese.

The paper goes over the many ways and challenges that come up when it comes to word alignment. Considering Hindi is based on subject object verb "SOV" and English is based on subject verb object "SVO," this study focuses on the major problem that occurs in word alignment. The report gives a survey on word alignment in the application of machine translation for foreign and Indian languages (Mall and Jaiswal, 2019).

Pourdamghani et al. (2018) described a strategy for enhancing word alignments by comparing words. This strategy is based on encouraging semantically comparable words to align in the same way. To estimate similarity, they employ word vectors trained on monolingual data. Additionally, by increasing the alignments of infrequent tokens, the researchers increase word alignments and machine translation in low-resource settings.

To improve the quality of Chinese-Vietnamese word alignment, Tran et al. (2017) incorporate linguistic relationship factors into the word alignment model. These are Sino-Vietnamese and content word linguistic relationships. The results of the experiments demonstrated that their strategy enhanced word alignment as well as machine translation quality.

Beloucif et al. (2016) presents a new statistical machine translation strategy that uses monolingual English semantic parsing to bias Inversion Transduction Grammar (ITG) induction and is specifically oriented to learning translation from low resource languages. The study shows that, in contrast to traditional statistical machine translation (SMT) training methods, which rely heavily on phrase memorization, the approach proposed focuses on learning bilingual correlations that aid in translating low-resource languages, with the output language semantic structure being used to further narrow ITG constraints.

Xiang et al. (2010) presented a novel approach for constructing and merging complementary word alignments for low-resource languages in order to increase word alignment quality and translation performance. In the study, they construct numerous sets of diverse alignments based on different incentives, such as linguistic knowledge, morphology, and heuristics, rather than focused on improving a single set of word alignments. By integrating the alignments acquired from syntactic reordering, stemming, and partial words, they demonstrate their strategy on an English-to-Pashto translation task. With much higher F scores and higher translation performance, the combined alignment surpasses the baseline alignment.

The researchers demonstrate that attention weights do accurately capture word alignments and propose two new word alignment induction methods, SHIFT-ATT and SHIFT-AET. The fundamental idea is to induce alignments when the to-be-aligned target token is the decoder input, rather than the decoder output, as in prior work (Chen et al., 2020).

In the study of Mao et al. (2022), they propose a word-level contrastive objective for many-to-many NMT that takes advantage of word alignments. For various language combinations, empirical studies demonstrate that this results in 0.8 BLEU gains. Analyses show that the encoder's sentence retrieval efficiency in many-to-many NMT is substantially correlated with translation quality, which explains why the suggested method has an impact on translation.

A study where the researchers used HMM-based models that were designed for bitext word and phrase alignment. The models are written in such a way that parameter estimation and alignment can be done quickly. Even with massive training bitexts, it has been founnd that Chinese English word alignment performance is comparable to IBM Model-4 (Deng and Byrne, 2005).

3 Methodology

3.1 Dataset

The parallel corpus that were used in this study come from the curated work of Adlaon and Marcos (2019). Their study aims to build a parallel corpus for Cebuano and Filipino where they used two different sources which is the biblical texts and the web. 500 sentence pairs in total of four domains were used for the experiments where it includes the bible texts, wikipedia, open domain, and news articles.

3.2 Data Cleaning and Transformation

In the dataset, the researchers performed data cleaning. This procedure was necessary in order to convert the data into a format that can be analyzed and be useful for the necessary experiments that will be applied to the corpus. Also, both Cebuano and Filipino texts were converted to lowercase. This is to avoid producing misleading results. Punctuations (i.e..,!?"';:-), numbers (i.e. 123...), and special characters (i.e. &*) were removed from the dataset which the researchers deemed to consider only alpha characters for this experimentation.

3.3 Preprocessing of the Corpus

Data preprocessing is a crucial step in doing an NLP task. This simply means transforming the data into a format that is predictable and easy to analyze (Menzli, 2021). In this experiment, the researchers performed subword tokenization specifically the Byte Pair Encoding to evaluate how tokenization contributes to distinguishing alignment of sentences of two different languages.

Byte Pair Encoding (BPE) or also known as diagram coding is a simple form of data compression in which the most common pair of successive bytes of data is replaced with a byte that does not present within that data (Mao, 2019). The BPE algorithm used in this study was from the work of Sennrich et al. (2016) where we set an average value of 35k merge operations per domain. Table 1 shows the comparison of a sentence without BPE, with BPE, and BPE with Lexicon trained on the corpus mentioned in section 3.1. The combined vocabulary of the four domains used in this study before BPE contains roughly 167k and 171k for Filipino and Cebuano respectively. After BPE, the vocabulary decreased its size to roughly 84k and 83k for Filipino and Cebuano respectively. The disparity of the size of the vocabulary from the set number of merge operations is attributed to the presence of scientific terms in the Wikipedia domain which the researchers supposed to exclude during the preprocessing phase.

In the study of Kudo (2018), they presented that BPE segmentation has the advantage of efficiently balancing vocabulary size and step size (the number of tokens required to encode the sentence). BPE uses a character frequency to train the merged processes. Early joining of frequent substrings will result in common words remaining as a single symbol. Rare character combinations will be broken down into smaller components, such as substrings or characters. As a result, even with a small fixed vocabulary (often 16k to 32k), the amount of symbols necessary to encode a sentence does not grow much, which is a crucial aspect for efficient decoding.

3.4 SimAlign Algorithm

There are different text aligners that are available and perform well on aligning two different languages. However, it requires a parallel data in order to generate great results. Also, the researchers aim to explore an embedding-based language model as several studies have shown that it could better capture both syntactic and semantic alignment(Jalili Sabet et al., 2020; Shen et al., 2017; Thompson and Koehn, 2019). In this paper, SimAlign algorithm which was proposed by Sabet et al. (2020) was utilized. The key concept of SimAlign is to use multilingual word embeddings for word alignment, both static and contextualized. In this study, we have used the pre-trained word embeddings available in the said study. For static embedding, for each language on Wikipedia, they used fastText (Bojanowski et al., 2016) to train monolingual embeddings. The embeddings are then mapped onto a shared multilingual space using VecMap (Artetxe et al., 2018). It must be noted that this algorithm operates without any cross-lingual supervision (e.g., multilingual dictionaries). On the other hand, multilingual BERT model (mBERT) was utilize in the contextualized embedding. It has been pre-trained on the 104 most popu-

Without BPE	**With BPE**
katulong umano ni velasco ang kanyang mga solid supporter sa kamara sa paggapang para maagapan ang inilulutong coup	*katulong umano ni velasco ang kanyang mga solid supporter sa kamara sa pagga@@ pang para maagapan ang inilu@@ lu@@ tong co@@ up*

Table 1: Comparison of a sentence without BPE and with BPE.

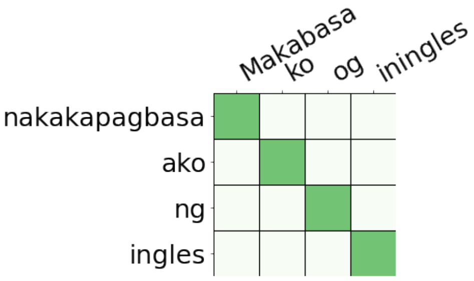

Figure 1: The alignment of Cebuano: Makabasa ko og iningles and Filipino: nakakapagbasa ako ng ingles with 4 words on each sentence. Which translates to *I can read English* in English.

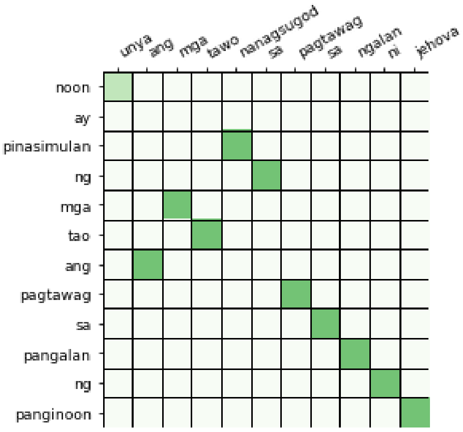

Figure 2: The alignment of Cebuano: unya ang mga tawo nanagsugod sa pagtawag sa ngalan ni jehova and Filipino: noon ay pinasimulan ng mga tao ang pagtawag sa pangalan ng panginoon with 11 and 12 words on each sentence for Cebuano and Filipino respectively. The input translates to *then the people started calling his name lord* in English.

lar Wikipedia languages. Also, only subword-level embeddings are offered by this model. Getting the average vectors of its subwords has been done to obtain a word embedding. Both the concatenation of all levels and word representations from each of the 12 layers are taken into account. It also has to be noted that the model has not been improved or finetuned. The study also proposed three different approaches namely, Itermax, Match, and Argmax to obtain alignments from similarity matrices. Itermax is a cutting-edge iterative approach, Match is a graph-theoretical technique focused on finding matches in a bipartite graph, while Argmax is a straightforward baseline. Figure 1, 2, and 3 shows how the alignment works of Cebuano and Filipino language of different word counts. The darker green shades are the sure links or equivalent translation of words for the both languages while the lighter green shade are the possible links or the translation that might have relation or if its not the exact translation of the word pair.

A gold standard must be created to measure the correctness of the different approaches in automatically aligning words using the SimAlign. The annotated gold standard used in this experiment was manually produced by the researchers where their mother-tongue language was Cebuano and Filipino language as their second language. The automatically generated alignment of Match, Inter, and Itermax will be evaluated using the 4 evaluation measures used for this experiment namely Precision, Recall, F1, and AER. AER requires a carefully annotated gold standard set of "Sure" and "Possible" links (referred to as S and P). Recall is measured using "sure" links, whereas Precision is measured using "possible" links. According to Och and Ney (2003), AER is derived from F-Measure. However, AER lacks one of F-most Measure's crucial features: the penalty for unbalanced precision and recall. The four measures are defined as:

$$Precision = \frac{|A \cap P|}{|A|}$$

$$Recall = \frac{|A \cap S|}{|S|}$$

$$F1 = \frac{2 Precision Recall}{Precision + Recall}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

4 Results and Discussions

In this section, we discussed the results for evaluating the aligned texts of the sentences with BPE and without BPE using the 4 evaluation measures namely Precision, Recall, F1, and AER. To determine the best alignment score, table 2 shows the three basis in choosing the best similarity matrix for each domain.

	Precision	Recall	F1	AER
Open Domain				
Match	0.778	**0.918**	0.842	0.16
Argmax	**0.873**	0.82	0.846	0.154
Itermax	0.813	0.908	**0.858**	**0.144**
Bible				
Match	0.634	**0.86**	0.73	0.273
Argmax	**0.798**	0.677	0.733	0.267
Itermax	0.726	0.817	**0.85**	**0.149**
Wikipedia				
Match	0.7	**0.9**	0.797	0.215
Argmax	**0.879**	0.758	**0.814**	**0.185**
Itermax	0.798	0.831	**0.814**	0.186
News Article				
Match	0.633	**0.858**	0.729	0.274
Argmax	**0.823**	0.688	0.749	0.249
Itermax	0.738	0.786	**0.761**	**0.239**
Applied with Byte-Pair Encoding				
Open Domain				
Match	0.746	0.895	0.814	0.188
Argmax	**0.867**	0.819	0.842	0.157
Itermax	0.816	**0.914**	**0.862**	**0.139**
Bible				
Match	0.515	**0.712**	0.598	0.405
Argmax	**0.649**	0.561	0.602	0.397
Itermax	0.589	0.646	**0.616**	**0.384**
Wikipedia				
Match	0.611	**0.832**	0.705	0.298
Argmax	**0.768**	0.702	0.734	0.266
Itermax	0.704	0.777	**0.739**	**0.262**
News Article				
Match	0.616	**0.836**	0.709	0.294
Argmax	**0.8**	0.669	0.729	0.27
Itermax	0.689	0.82	**0.749**	**0.254**

Table 2: Evaluation results of the aligned sentences with and without embedding. The best results per column on different domains are printed bold.

4.1 Without BPE

The alignments for the source and target texts are by tokens which was separated by white space. The result shows that without implementing BPE, the Open domain gets the highest score for *recall and F1*, with scores **0.918, 0.858** respectively which means the aligner was able to get the most number of matches compared to the other domains. Moreover it also gets the lowest score for AER with 0.144 which indicates that it has the lowest error rate among other domains. This could be attributed to its length that is shortest compared to the other domains.

It can also be observed that News Article domain gets the lowest score for *precision and F1*, with scores **0.633, 0.729** respectively. Additionally, it has the highest *AER* with the score **0.274** which tells us that this domain has the highest error rate. Upon the creation of the gold standard, we observed that the News Article corpus contains a lot of numbers, dates, and figures. However, since the dataset was preprocessed before the aligning of words, these numbers were removed and some necessary punctuations like hyphens which caused segmentation that makes the words incomprehensible and confusing that affects the alignment.

Based on the four domains used in this experiment, the Bible corpus has the most tokens per sentence which contains 1104 and 1507 sentences with number of tokens greater than 50 for Filipino and Cebuano respectively while there were no sentences greater than 50 tokens in the Open Domain. In line with this, we have observed that in short-length domains we acquire best results for Itermax while Match or Argmax are best for long-length domains. Figures 4, 5, 6, and 7 shows the examples of word alignments of Bible and Open Domain with and without BPE.

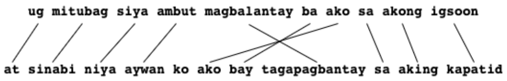
ug mitubag siya ambut magbalantay ba ako sa akong igsoon
at sinabi niya aywan ko ako bay tagapagbantay sa aking kapatid

Figure 3: Example word alignment of Bible Text without BPE

4.2 With BPE

We implemented the Byte Pair Encoding on the four domains to evaluate the difference when the tokens are segmented or not. The result shows that with BPE, the Open Domain gets the fore-

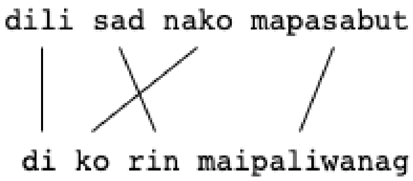

Figure 4: Example word alignment of Open Domain without BPE

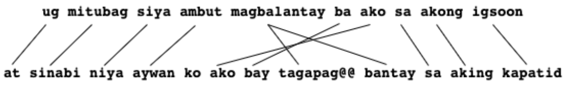
ug mitubag siya ambut magbalantay ba ako sa akong igsoon
at sinabi niya aywan ko ako bay tagapag@@ bantay sa aking kapatid

Figure 5: Example word alignment of Bible text with BPE

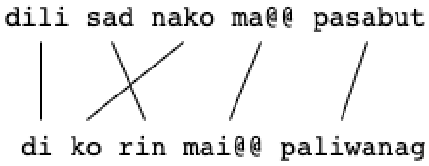

Figure 6: Example word alignment of Open Domain with BPE

most score for *precision, recall, F1, and AER*, with scores **0.867, 0.914, 0.862, 0.139** respectively which means the aligner was able to get the most number of matches compared to the other domains when applied with BPE.

It can be noticed that Bible domain gets the most unsatisfactory results for *precision, recall, F1, and AER*, with scores **0.515, 0.561, 0.602, 0.405** respectively.

Overall, if we compare the results of the dataset without BPE and with BPE, without BPE shows significantly higher scores than the dataset implemented with BPE. As what you have noticed in Table 1, on the 2nd column, the tokens are separated in a way that it is not understandable which also explains why the scores are low.

5 Conclusion

Sentence aligned parallel corpora are crucial in Machine Translation and choosing the most efficient aligner in different languages will be of great help in doing NLP tasks. In this study, we have observed that when aligning words, results are favorable when tokens are not segmented with BPE. Also, in the alignment from similarity matrices Match or Argmax are preferred for long-length sentences and Itermax for short-length sentences.

For future studies, it is recommended to increase the number of sentence pairs in the experimentation of the SimAlign to maximize the performance of algorithm. It is also recommended to explore a different embedding model that is specific to this kind of language to evaluate how embedding models affect the results of the alignment.

References

Kristine Mae M. Adlaon and Nelson Marcos. 2018. Neural machine translation for cebuano to tagalog with subword unit translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 328–333.

Kristine Mae M. Adlaon and Nelson Marcos. 2019. Building the language resource for a cebuano-filipino neural machine translation system. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2019, page 127–132, New York, NY, USA. Association for Computing Machinery.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Meriem Beloucif, Markus Saers, and Dekai Wu. 2016. Improving word alignment for low resource languages using English monolingual SRL. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 51–60, Osaka, Japan. The COLING 2016 Organizing Committee.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 175–es, USA. Association for Computational Linguistics.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

CLARIN. 2022. Parallel corpora.

Yonggang Deng and William Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology*

Conference and Conference on Empirical Methods in Natural Language Processing, pages 169–176, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic. Association for Computational Linguistics.

A.N. Lazaro, Nathaniel Oco, and Rachel Edita Roxas. 2017. Developing a bidirectional ilocano-english translator for the travel domain: Using domain adaptation techniques on religious parallel corpora. In *11th International Conference of the Asian Association for Lexicography*, Guangzhou, China.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Shengxuan Luo, Huaiyuan Ying, and Sheng Yu. 2021. Sentence alignment with parallel documents helps biomedical machine translation. *CoRR*, abs/2104.08588.

Shachi Mall and Umesh Chandra Jaiswal. 2019. Issues in word alignment from hindi-english languages. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8.

Lei Mao. 2019. Byte pair encoding.

Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan, and Sadao Kurohashi. 2022. When do contrastive word alignments improve many-to-many neural machine translation? In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1766–1775, Seattle, United States. Association for Computational Linguistics.

Tom McCoy and Robert Frank. 2017. Pivot-based word alignment.

Amal Menzli. 2021. Tokenization in nlp: Types, challenges, examples, tools.

Leah Michel, Viktor Hangya, and Alexander Fraser. 2020. Exploring bilingual word embeddings for Hiligaynon, a low-resource language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2573–2580, Marseille, France. European Language Resources Association.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Robert Östling. 2014. Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 123–127, Gothenburg, Sweden. Association for Computational Linguistics.

Charmaine Ponay and Charibeth Cheng. 2015. 23. building an english-filipino tourism corpus and lexicon for an asean language translation system.

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528, New Orleans, Louisiana. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *CoRR*, abs/2004.08728.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. 2017. Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189, Copenhagen, Denmark. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2021. CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jun Tariman. 2010. Cebuano 101: The cebuano language sentence structure. pages 22–26.

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), pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Phuoc Tran, Dien Dinh, Tan Le, and Long H. B. Nguyen. 2017. Linguistic-relationships-based approach for improving word alignment. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1).

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*.