Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2573–2580
Marseille, 11–16 May 2020

Exploring Bilingual Word Embeddings for Hiligaynon, a Low-Resource Language

Leah Michel, Viktor Hangya, Alexander Fraser
Center for Information and Language Processing
LMU Munich, Germany
l.michel@campus.lmu.de, {hangyav, fraser}@cis.lmu.de

Abstract

This paper investigates the use of bilingual word embeddings for mining Hiligaynon translations of English words. There is very little research on Hiligaynon, an extremely low-resource language of Malayo-Polynesian origin with over 9 million speakers in the Philippines (we found just one paper). We use a publicly available Hiligaynon corpus with only 300K words, and match it with a comparable corpus in English. As there are no bilingual resources available, we manually develop a English-Hiligaynon lexicon and use this to train bilingual word embeddings. But we fail to mine accurate translations due to the small amount of data. To find out if the same holds true for a related language pair, we simulate the same low-resource setup on English to German and arrive at similar results. We then vary the size of the comparable English and German corpora to determine the minimum corpus size necessary to achieve competitive results. Further, we investigate the role of the seed lexicon. We show that with the same corpus size but with a smaller seed lexicon, performance can surpass results of previous studies. We release the lexicon of 1,200 English-Hiligaynon word pairs we created to encourage further investigation.

Keywords: bilingual word embeddings, bilingual lexicon induction, post-hoc mapping, low-resource languages, Hiligaynon

1. Introduction

Since the introduction of Skip-gram and Continuous Bag-of-Words (Mikolov et al., 2013a) followed by the release of GloVe (Pennington et al., 2014), the use of word embeddings, *the continuous vector representations of words*, has become the norm for many natural language processing (NLP) tasks. From sentence classification (Kim, 2014), part-of-speech tagging (Abka, 2016), named entity recognition (Melamud et al., 2016), to sentiment analysis (Ruder et al., 2016), it is hard to imagine modern NLP without word embeddings. How this success of monolingual word embeddings (MWEs) can extend to a bilingual setup – to represent meaning and transfer knowledge in bilingual tasks – sparks interests that paved the way for investigating different models of bilingual word embeddings (BWEs). When two MWEs share a single vector space in the form of BWEs, it is likely that, following the concept of MWEs correlating distance with semantic similarity, the target language word closest to the source language word is the translation equivalent.

To evaluate the quality of BWEs, this paper conducts experiments with the task of bilingual lexicon induction (BLI). This task aims to mine accurate word translations from a source language to a target language. The output is a list of source words with corresponding lists of top-n translations in the target language based on cosine similarity in the BWE space. Unlike translations that are learned from parallel corpora which are sentence-aligned bilingual texts, BLI utilizes data that are not necessarily parallel. This task is crucial for many reasons: 1) many of the world's languages have either limited or no parallel corpora, 2) reliable, freely accessible dictionaries for low-resource languages are either scarce, highly erroneous or non-existent, 3) modern statistical and neural machine translation systems often rely on dictionaries and phrase tables, a downstream task of BLI, and 4) the tedious process of building parallel corpora for low-resource languages could potentially be alleviated through BLI (Hangya et al., 2018).

There are more than 7,000 living languages spoken in the world today and over 2,500 of them are now considered endangered (Eberhard et al., 2020). Out of these languages, research has been extensive on only a few major languages, such as English, Spanish, French, German, Chinese and Arabic. Linguistic resources in these languages abound; understandably, statistical and neural machine translation models that have seen impressive results focus on these languages, exploiting massive amounts of parallel corpora. High quality parallel corpora take years to build, backed by millions of dollars per language. To develop systems that can automatically generate translations relying only on monolingual corpora is therefore a huge step in machine translation, particularly for low-resource languages (Artetxe et al., 2019; Conneau and Lample, 2019).

Research on low-resource languages, however, poses many challenges. Unlike well-researched languages, they are mainly minority languages that do not have established grammar, dictionaries and orthographic standards. They are usually spoken at home, in small communities or in regions where people primarily learn the language by speaking, leading to lack of written documents, much less machine-readable data. There are usually no active communities of researchers that build linguistic resources. With these challenges, exacerbated by the diminishing number of speakers, the evaluation of systems for low-resource languages is a herculean and expensive task.

The suitability of models is also an issue. Models are often evaluated on languages with similar linguistic characteristics, making performance more consistent. This does not necessarily trickle down to languages of other families where linguistic typology is different. There is a need

to come up with innovative ways to develop systems that could either perform well even with limited data, or could achieve some degree of generalization when applied to languages of other origins.

1.1. The Case of Hiligaynon

To test how the BLI systems from previous papers perform when applied to an extremely low-resource language, this paper experiments on Hiligaynon, a Malayo-Polynesian language spoken in the Philippines. Commonly referred to as *Ilonggo*, it is part of the Bisayan language group that is predominantly spoken in the provinces of Western Visayas and SOCCSKSARGEN (South Cotabato, Cotabato, Sultan Kudarat, Sarangani and General Santos). Hiligaynon, which comes in 3 dialects, is the fourth most spoken language among the estimated 186 languages in the archipelagic country, having an estimated total of over 9 million native speakers. It is written in Latin script that used to follow Spanish orthographic conventions but is currently generally written based on the orthographic standards of Filipino, the Philippine national language.

Similar to Filipino, the country's colonial past left traces in Hiligaynon. Many loan words are adjusted to the orthography and pronunciation of native Hiligaynon sounds. Spanish words like *abyerto (abierto), timprano (temprano), gwapo (guapo), munyeka (muñeca), merkado (mercado), kambyo (cambio), ubra (obrah)*, to name a few, find frequent use. Code-switching to English, the country's another official language, is common, a footprint of America's post-Spanish occupation.

As is mostly the case with Philippine languages, there is only little NLP work on Hiligaynon. Since 2008, only one research paper for statistical machine translation (Oco and Roxas, 2018) has been published. Experiments yielded a BLEU score of 21.74 for Hiligaynon to English, and 24.43 for English to Hiligaynon, using the parallel corpora from the New Testament (Macabante et al., 2017). The lower performance of the Hiligaynon to English translation is attributed by Macabante et al. (2017) to the differences in word order between English and Hiligaynon; unlike the typical subject-verb-object (SVO) order of English, Hiligaynon has a free-word order, i.e., sentences are typically expressed in VSO, and sometimes, in VOS or SOV form, depending on emphasis.

For our experiments, we reached out to Macabante et al. (2017) for a copy of the parallel corpus, but were informed that the corpus is no longer available due to technical issues. We make do with the available Hiligaynon monolingual corpus consisting of a bit over 300,000 words in literary and religious texts from the now-defunct Palito website (Dita et al., 2009), datasets of which are still accessible online [1]. To produce a comparable dataset, religious and literary texts in English were also collected. 1,200 most frequent words from the English corpus were then extracted and translated into Hiligaynon by a native speaker: the first 1,000 pairs serve as the training seed lexicon, and the remaining 200 pairs as test set.

[1] https://www.dropbox.com/sh/b1dp56htdm9qux0/AABsNv12EzzdJDpQNop3gb5ea?dl=0

Since results of our experiments show that the small dataset is not sufficient to accurately mine translation pairs, we simulate the low-resource scenario for English to German, varying the size of the dataset and the seed lexicon to find out how both of these resources impact performance.

2. Related Work

One pioneering approach in creating BWEs for BLI is the model this paper investigates. Observing a linear relationship between languages, Mikolov et al. (2013b) mapped the MWEs of a source language to the MWEs of a target language by linear transformation. The bilingual signal to learn the transformation was a small seed lexicon of 5,000 word pairs. The model is simple, inexpensive, and showed impressive results for Spanish to English, and English to Czech translations. To test whether the same approach works for unrelated languages, over a billion Vietnamese phrases were trained to mine Vietnamese translations for English words and vice versa, applying previous techniques in vector representations of phrases (Mikolov et al., 2013c). Even with the large corpora, the accuracy is only 10 percent for English to Vietnamese and 24 percent for Vietnamese to English. In this paper, we attempt a word-word translation of English to Hiligaynon using a small available Hiligaynon corpus of a bit over 300,000 tokens.

Braune et al. (2018) further explored the feasibility of this approach by evaluating the quality of BWEs to mine rare and domain-specific words, as well as frequent words, in English to German. For frequent words in the general domain, the dataset consisted of monolingual corpora of 4,400,309 English and German sentences from parliament proceedings, news commentaries and web crawls taken from the WMT 2016 shared task (Bojar et al., 2016), a bilingual signal of 4,955 frequent word pairs, and 2,000 frequent words as test and validation sets. The baseline model failed to deliver promising results, but by applying ensembling techniques combined with orthographic cues, state-of-the-art results were achieved.

We extend the experiments of Braune et al. (2018) in mining frequent words for English to German, using their seed lexicon and test set. We set up datasets with different amounts of monolingual corpora, as well as different lexicon size, to determine the amount of data needed to arrive at comparable results.

Similar to our work, Adams et al. (2017) investigated word embeddings in low-resource setups. It was shown that the monolingual quality, i.e., the correlation of the cosine similarity of the embeddings with human word similarity judgements, drops drastically when only a few thousand sentences are available. To improve the performance, a large number of sentences from other languages were leveraged. To bridge the gap between the languages, a joint BWE model (Duong et al., 2016) and large bilingual dictionaries were used. They showed significant improvements in monolingual word similarity. However, they do not report results on bilingual quality, such as BLI, and rely on a large dictionary which is often not available for low-resource languages. In contrast, we focus on evaluating the bilingual quality of the models assuming only a small seed lexicon.

En-Hil	
Language	Words
English	345,583
Hiligaynon	319,934
En-De	
Language	Words
English	300,120
German	300,099

Table 1: Small datasets. The En-Hil set is composed of 70 percent literary and 30 percent religious texts each. The En-De set is from the Opus website made up of EU proceedings, news commentaries and books in equal distribution.

3. Approach

Following the experiments of Braune et al. (2018) on frequent words in the general domain, we use the code they made public on Github[2]. To find out if the model also works for an actual, low-resource language in the same BLI task, we conduct experiments in English to Hiligaynon. To see how the results fare compared to mining frequent words in related languages, we experiment with English to German, simulating the same low-resource setup.

We show that the model fails in a low-resource scenario, both for unrelated and related languages. We then increase the number of words of monolingual corpora for English to German. The goal is to determine at which corpora size does the performance become comparable to that achieved by Braune et al. (2018). We also look at how the seed lexicon affects performance by varying its size from 1,000 to 4,955 pairs.

3.1. Training Data

We test the model on two sets of data: 1) small datasets of comparable English to Hiligaynon (En-Hil) and English to German (En-De) as shown in Table 1, and 2) large datasets of En-De in varied sizes of monolingual corpora as shown in Table 2. How the monolingual corpora are gathered and prepared are detailed below.

3.1.1. Small data

The Hiligaynon corpus consists of literary and religious texts from the Palito corpus (Dita et al., 2009). To make the data comparable, we collected literary texts in English from Planet eBook[3], and religious texts from Sermon Online[4]. The En-De set, on the other hand, is taken from the Opus website [5] consisting of books, EU proceedings and news commentaries.

3.1.2. Large Data

To assess the impact of the size of monolingual corpora, we gathered texts in English and German from the Opus website consisting of books, EU proceedings, news commentaries, news, UN proceedings, Ted Talks and Wikipedia articles. We start with an En-De dataset of 1.5 million tokens, doubling the size per dataset until we reach our largest dataset of over 105 million tokens. This makes 7 En-De datasets in total. When possible, so that both languages have the same distribution of data, the amount of tokens per type of text from either language is adjusted. Table 2 shows the largest and smallest datasets.

104M		
Texts	En	De
books	1,412,247	1,330,089
europarl	37,808,676	37,631,718
newscomm	6,522,888	6,521,181
globalvoices	1,657,466	1,656,038
multi un	6,897,765	6,856,049
ted talk	2,679,589	2,677,251
wiki	48,090,535	48,014,011
Total	105,069,161	104,686,337
1.5M		
Texts	En	De
books	214,330	214,384
europarl	214,334	214,423
newscomm	214,205	214,393
globalvoies	214,367	214,524
multi un	214,389	214,381
ted talk	214,423	214,516
wiki	214,438	214,394
Total	1,500,486	1,501,015

Table 2: Largest and smallest En-De datasets from the Opus website. Numbers indicate number of tokens from books, EU proceedings, news commentaries, news from Global Voices, UN proceedings, Ted Talks, and Wikipedia pages from the Opus website.

3.2. Seed Lexicons

To learn BWEs, a seed lexicon is used as bilingual signal. The En-Hil lexicon consists of translation pairs translated by a native speaker.

1) For the En-Hil dataset: 1,000 pairs composed of frequent words from the English small corpus paired with Hiligaynon word translation

2) For the En-De small dataset: first 1,000 pairs from the frequent general domain seed lexicon of Braune et al. (2018) which are German translations of the most frequent English words in the WMT 16[6] dataset

3) For the large En-De datasets: all 4,955 pairs from the frequent general domain seed lexicon of Braune et al. (2018). In experimenting with the seed lexicon size, the number of pairs was decreased.

3.3. Test Data

These are translation pairs that are not in the seed lexicon. As is done with the seed lexicons, the En-Hil test set was translated by a native speaker while the En-De test sets

[2] https://github.com/braunefe/BWEeval
[3] www.planetebook.com
[4] www.sermon-online.de
[5] http://opus.nlpl.eu/
[6] www.statmt.org/wmt16/

(small and large) are taken from the test set used by Braune et al. (2018) for mining frequent words in the general domain.

1) For the En-Hil dataset: 200 of the next most frequent words in the English text (after the 1,000th word) paired with Hiligaynon translation
2) For En-De small dataset: the first 200 pairs from Braune et al. (2018).
3) For the large En-De dataset: all 1,000 pairs from Braune et al. (2018).

The En-De seed lexicon and tests sets are taken from Braune et al. (2018) which used a standard phrase-based SMT system trained on WMT 2017 data.

3.4. Training MWEs

The monolingual datasets in Section 3.1. are first normalized with Moses tools for tokenizing and lower-casing[7]. Punctuation marks are not removed. As there are plenty of digits from the religious texts, i.e., verse numbers, they are deleted along with the noticeable series of empty lines observed in the Hiligaynon texts. These are simply done with Unix commands.

After normalization, Skip-gram and CBOW are trained on the monolingual corpora using the toolkits word2vec[8] (Mikolov et al., 2013a) and fastText[9] (Bojanowski et al., 2017). Training takes longer with fastText as it uses character n-grams to represent word vectors, but the subword information it contains has been shown to better represent morphologically-rich languages like German.

Except for setting the dimensions to 50 and 300, and the minimum word count to 3 in order to compensate the small corpus size, all other parameters are set with default values. It is interesting to note that although the small En-Hil dataset has more English word tokens than the En-De dataset has, its number of English word types is lower. This is due to the domains of texts in the datasets; En-Hil consists only of 2 (literary and religious), while the small En-De dataset consists of 3 domains (EU proceedings, news commentaries and books). Having a larger vocabulary size could have both advantages and disadvantages. It could lead to a smaller number of out-of-vocabulary words, but on the other hand, the embeddings could be noisier due to low frequency.

3.5. Training BWEs

To project the MWEs into a single vector space, we implement the post-hoc mapping model of Mikolov et al. (2013b). This allows projection of the vector space of a source language *s* to the vector space of the target language *t* by learning a transformation matrix **W**.

This approach uses a small seed lexicon consisting of words from the source language $w_1^s, ..., w_n^s$ and their translations $w_1^t, ..., w_n^t$. The transformation matrix **W** is then learned using stochastic gradient descent by minimizing the squared Euclidean distance (mean squared error or MSE) between the previously learned monolingual embeddings of the source seed word x_i^s with the use of **W** and its translation x_i^t in the seed lexicon:

$$\Omega_{MSE} = \sum_{i=1}^{n} ||\mathbf{W}x_i^s - x_i^t||^2 \quad (1)$$

3.6. Mining Translation Pairs

The process of mining translation pairs to form a bilingual lexicon is simply done by computing the cosine similarity between the source word (in this case, English) contained in the gold standard and a target word (in this case, Hiligaynon or German) in the aligned BWEs. The German or Hiligaynon word closest to the English word is the top 1 translation candidate (highest cosine similarity), while the top 5 translation denotes that the translation candidate is taken from one of the 5 closest neighbors of the source word. That means, if one of these 5 translation candidates matches the translation from the gold standard, it is induced into the resulting lexicon. Accuracy is then determined by computing the total number of matches divided by the total number of the gold standard (i.e., 50 induced translations that match the 100 gold standard is 50 percent accuracy).

3.7. Ensembling

As ensembling produced better results in previous work (Braune et al., 2018), this technique is also applied in our experiments. This is done by generating the *n*-best translation candidates (*n*=100) from BWEs taken from different MWEs (word2vec Skip, word2Vec CBOW, fastText Skip, and fastText CBOW). The ensemble weight is computed by:

$$\sum_{i=1}^{M} \gamma_i Sim_i(s,t) \quad (2)$$

where $Sim_i(s,t)$ is the cosine similarity of a translation pair. $Sim_i(s,t)$ is valued at 0 if the translation candidate is not in the *n*-best list. The weighted sum of these cosine similarities then becomes the ensemble similarity score. For this, a validation set is used to fine-tune the weights γ_i for each test with the use of a grid search. The 1,000 validation set for ensembling is taken from Braune et al. (2018).

3.8. Ensembling + Edit Distance

By integrating a measure of similarity between word strings, Braune et al. (2018) showed that an even higher BLI performance can be achieved. To do this, the ensemble equation in Section 3.7. is extended with the orthographic similarity (one minus the Levenshtein distance) between the surface-forms of words *s* (source word i.e. English word) and *t* (target word, i.e., Hiligaynon or German). The *n*-best lists of candidate translations from different word similarity models including all BWE and orthography ($OSim(s,t)$) based models, are then generated. All these lists are then ensembled together.

[7] https://github.com/moses-smt/mosesdecoder

[8] https://github.com/dav/word2vec

[9] https://github.com/facebookresearch/fastText

4. Results

We present our results in this section starting with the experiments on the small En-Hil and En-De datasets, followed by various analyses on the large En-De datasets.

4.1. Small Data

It turned out that the model fails when applied to a very small dataset, even for related languages (En-De) as shown in Table 3.

Dimensions	En-De	En-Hil
50	0.0 (0.0)	0.5 (0.5)
300	0.0 (0.0)	0.0 (0.0)

Table 3: Bilingual lexicon induction of small En-De and En-Hil datasets in Top 1 (Top 5) percent accuracy. Shown here are MWEs trained with word2vec Skip-gram in 50 and 300 dimensions. MWEs trained with fastText have similar results.

The only word accurately predicted in Hiligaynon is *light – suga*. Words that are even orthographically close to each other are not accurately predicted, e.g. for En-Hil the words *possible – posible, color – kolor, angel – anghel* and for En-De the words *zone – zone, minute – minute*. This is, as also mentioned by Braune et al. (2018), due to the fact that there is no cross-lingual learning between two monolingual corpora since they have been trained separately. Hence, although fastText's subword information help better represent words with similar substrings, it does not prove effective in the face of very limited data.

4.2. Large Data

In order to have a deeper understanding of the required resources for building useful BWEs, we test various setups on En-De. Figure 1 shows the impact of the size of monolingual corpora to BLI performance. Performance having 1.5M words yield 0.2 percent accuracy. Between 12M and 25M, a steep learning curve (8 percent increase) can be observed. This reflects the results of previous works (Irvine and Callison-Burch, 2017; Mikolov et al., 2013b). Accuracy improves as the amount of data increases. We show results derived from word2vec Skip-gram since it consistently outperformed other MWE models across all sizes of monolingual corpora. Detailed comparison with fastText Skip-gram can be seen in Figure 2.

4.2.1. Ensembling + Edit Distance

Validating the work of Braune et al. (2018), Figure 2 shows increase in accuracy when the techniques of ensembling and ensembling with edit distance are applied. The poor performance of fastText, probably brought by the noise in the small corpus making character n-grams worse, drags down the effect of the technique. If compared with word2vec's lone performance, the increase is not as significant (only around 3 percent).

Due to the very limited data for Hiligaynon, the effect of ensembling with orthographic distance cannot be established in this paper. English and German have many words with closer orthographic distance and as such, there is a noticeable positive effect in accuracy. For languages that are not at all related like English and Hiligaynon, the effect could be insignificant.

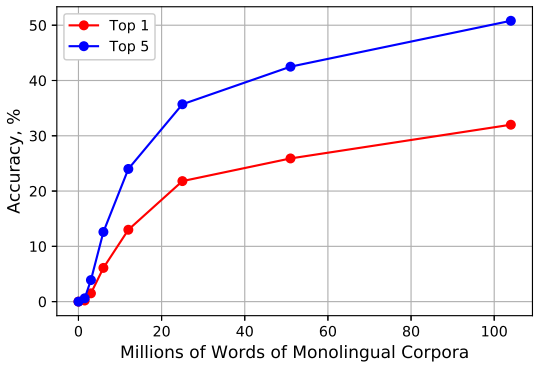

Figure 1: Learning curves with different sizes of English and German monolingual corpora using the Opus datasets. MWEs are trained with word2vec Skip-gram.

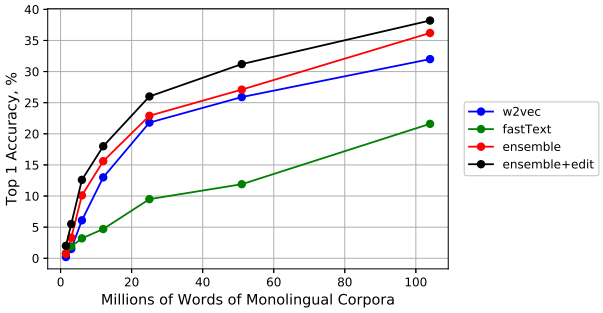

Figure 2: Learning curves trained with word2vec and fastText Skip-gram, including ensemble and ensemble + edit techniques over varying sizes of English and German monolingual corpora.

4.2.2. Impact of Lexicon Size

When considering resources, we must consider not only the monolingual corpora, but also the seed lexicon. It is therefore worth taking a look at how the size of the seed lexicon impacts the BLI learning curve. Figure 3 shows that unlike the impact of the size of monolingual corpora, the size of the seed lexicon is not proportionate to increase in accuracy. More seed lexicon pairs do not mean higher accuracy;

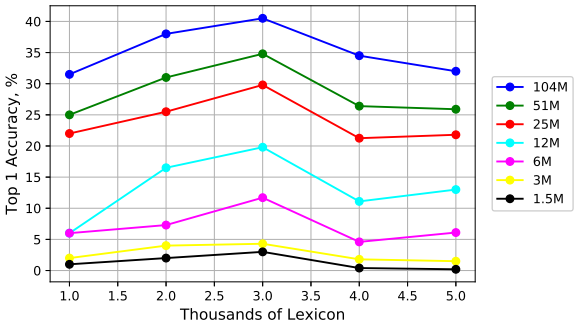

Figure 3: Learning curves over varying En-De lexicon sizes. MWEs are learned by word2vec Skip-gram.

they can even be detrimental.

With the different sizes of the seed lexicon shown in Figure 3, following the standard 80-20 ratio (80 train-20 test), the test data for the large En-De data in Section 3.3. is ignored from the 4,000 down to 1,000 lexicon size. For the 4,000 lexicon size, for example, the first 4,000 pairs from the lexicon are used as seed lexicon, and the next 800 are used as test set. The excess pairs are ignored.

As shown in Figure 3, learning peaked at 3K lexicon size and dropped thereafter, with at least 11 percent loss in accuracy in datasets 6M to 104M. Between 4K and 5K lexicon size, accuracy stays almost the same. The same phenomenon was observed by Vulić and Korhonen (2016). This, they reckon, is probably due to highly frequent words receiving more accurate representations (seed lexicon consists of 5,000 most frequent words with translation). The 2,000 additional less frequent words, therefore, could just be additional noise. From our dataset, it is hard to determine if and how much of the seed lexicon consists of frequent words since the seed lexicon is taken from a different dataset used by Braune et al. (2018). We refer the interested reader to (Lubin et al., 2019), who proposed a joint model for detecting noise in the seed lexicon while learning the mapping and showed improved BLI performance.

4.2.3. Translation Examples

Many induced top-1 German translations, as shown in Table 4, are semantically similar or synonymous to the words in the gold standard set. Since these are not counted as matches, the actual accuracy of the model can be higher than the calculated performance.

4.2.4. Impact of Dimensionality

Although not as significant as the impact of the seed lexicon size, it can also be noted that the effect of dimensionality in the performance of word2vec and fastText Skip-gram is the opposite (Figure 4). As has been shown by previous work (Mikolov et al., 2013b), word2vec's accuracy decreases as the dimensions get smaller. In this paper's experiment, fastText, though still at least 10 percent behind word2vec, performs better with the decrease in dimensionality. The improvement is only 2 percent (at best) though. From 200 to 100 dimensions, word2vec loses around 4 percent accuracy.

5. Error Analysis

5.1. Pre-processing

The experiments of Mikolov et al. (2013b) involved more pre-processing steps (e.g., removal of duplicate sentences, named entities and special characters, rewriting of numeric values, treating collocations like *ice cream* as one unit). Braune et al. (2018) and this paper simply normalized the monolingual corpora with the Moses tokenizer. As a result, for example, digits (like years) indeed induce numbers as translation candidates (which means semantic similarity is captured) – but they are mostly the wrong numbers, and therefore are not counted as matches, decreasing accuracy.

5.2. Lexicon Entries

Upon closer inspection of the seed lexicon, some of the translation pairs in the En-De lexicon are erroneous, e.g.

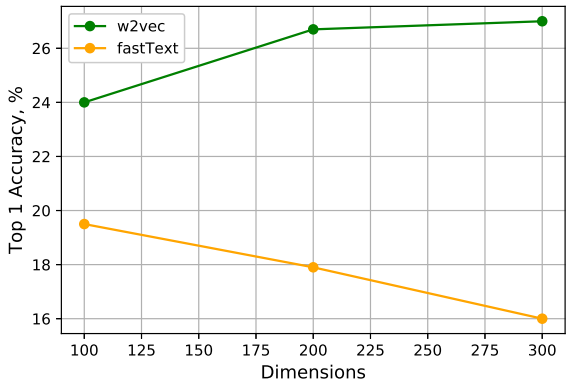

Figure 4: A comparison of learning curves in varying dimensons, trained with fastText Skip-gram and word2vec Skip-gram. Trained data consists of 51 million tokens of English and German monolingual corpora.

moon – fliegen (to fly) or *colonial – guatemala*. Spelling errors and tokenization errors (like *doesn* instead of *does n ' t*) were also retained. Stopwords were also observed in the En-De lexicon, contributing to the noise in the embeddings.

5.3. Hiligaynon Linguistic Attributes

In creating the English-Hiligaynon lexicon, one difficulty is the word-word translation. Plurality in Hiligaynon is formed through particles, e.g. from *banana – saging* to *bananas – **mga** saging*, from *Juan (si Juan)* to *Juan and those with him – **sanday** Juan*. Although some nouns still retain the Spanish convention of ending male or female nouns with 'o' or 'a', e.g. *male teacher – maestro* or *female teacher – maestra*, some nouns are of neuter gender. To indicate noun in masculine or feminine form, the word for male or female is added, linked with *nga*, e.g. *son – bata **nga lalaki*** (literal translation: *child that is male*), *daughter – bata **nga babayi*** (literal translation: *child that is female*). Comparatives are expressed in two words, preceded by the Spanish *mas*, e.g. *smaller – **mas** gamay*. Reduplication is applied to intensify, e.g. from *many – damo* to *too many – damo-damo*, or to diminish meaning, e.g. from *house – balay* to *playhouse – balay-balay*. Comparatives, plural and gender-specific nouns are replaced with their lemma in the lexicon, e.g. *big* instead of *bigger*, *child* instead of *children*, *child* instead of *son* or *daughter*.

Affixes come in different forms: prefix, infix (usually inserted after the first consonant of the stem), and suffix. A stem can be a root or a root with affixation, which means affixed forms can go through further affixation. Hence, some words consist only of a root while others are complex forms of a root with affixes (Wolfenden, 1971).

Variations in orthography also add to the noise. The 'i' is interchangeable with 'e' (*babaye* is the same as *babayi*), likewise in the case of 'o' to 'u' (*damo* is the same as *damu*), such that different authors observe different spelling conventions. Further, code-switching between Hiligaynon and English or between Hiligaynon and Filipino is also prevalent in the Hiligaynon corpus, adding more noise to the data.

English Word	Induced German Translations	Gold Standard Entry
abuses	menschenrechtsverletzungen *(human rights violations)*	missbrauch
recognition	akzeptanz *(acceptance)*	annerkennung
spiritual	religiosen *(religious)*	geistige
duration	laufzeit *(runtime)*	dauer
warning	meldung *(announcement)*	warnung
usd	euro	usd
amounts	mengen *(quantities)*	beträge
fears	befürchtungen *(apprehensions)*	ängste

Table 4: Example top-1 translations of English words to German, manually chosen from the results of MWEs trained with w2vec Skip-gram. Data consists of 25M tokens and seed lexicon of 3,000 En-De pairs.

6. Conclusion and Future Work

We encourage further research on Hiligaynon by releasing the English-Hiligaynon lexicon we created for this paper. Many other languages are still considered low-resource, and with the growing diversity of languages in digital devices and in the Internet, more research focusing on other language families with limited data should gain more attention.

With this work, we provide additional proof that data scarcity is still a hindrance to training quality MWEs and consequently, quality BWEs. Because of the data-driven nature of existing models where the learning curve is strongly influenced by the number of words of monolingual corpora, there is still a lot to be explored as to what models, both for monolingual and bilingual word embeddings, can overcome the challenge of limited data. There is of course no one-size-fits-all model for all languages, as every language or language family has its own unique syntax and word concepts.

This paper also reveals that the frequency of the seed lexicon does not play a significant role in mining accurate translations with BLI. Moreover, in searching for ways to reduce resources while keeping performance on par with previous work, this paper shows that the seed lexicon, considered *inexpensive* as it is in comparison with other BWE models, can further be minimized – with even better results. With 25 million words of monolingual corpora using only 3,000 seed lexicon, performance of the word2vec Skipgram (29.8 percent) even surpasses results released by previous study (Braune et al., 2018) which trained 100-million-word corpora using 5,000 seed lexicon (27.1 percent).

There is still a lot of room for improvement. For instance, the problem of polysemy should be addressed in MWE models so that the two or more polysemous senses of a single word type are not represented using the same vector.

Another future work is training BWEs with max margin ranking loss (Lazaridou et al., 2015). As also shown by Braune et al. (2018), this technique generates better results than the post-hoc mapping model applied in our experiments. Additionally, Hiligaynon could also benefit from cross-lingual transfer learning, exploiting high-resource related languages like Cebuano, Filipino or even Spanish.

Using a simple, inexpensive model, the experiments and analysis in this paper provide various insights into different factors affecting performance for mining translation pairs – from the size of the monolingual corpora, the frequency and size of the seed lexicon, down to the impact of dimensionality in the performance of word2vec and fastText.

7. Acknowledgements

We would like to thank the reviewers for their valuable input. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

8. Bibliographical References

Abka, A. F. (2016). Evaluating the use of word embeddings for part-of-speech tagging in bahasa indonesia. In *2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 209–214, October.

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 937–947.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An Effective Approach to Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August.

Braune, F., Hangya, V., Eder, T., and Fraser, A. (2018). Evaluating bilingual word embeddings on the long tail. 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 188–193, New Orleans, Louisiana, June.

Conneau, A. and Lample, G. (2019). Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Dita, S., Roxas, R., and Inventado, P. (2009). Building online corpora of philippine languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, volume 2, pages 646–653.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning Crosslingual Word Embeddings without Bilingual Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2020). *Ethnologue: Languages of the World*. SIL International.

Hangya, V., Braune, F., Kalasouskaya, Y., and Fraser, A. (2018). Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Workshop on Spoken Language Translation*.

Irvine, A. and Callison-Burch, C. (2017). A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310, June.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October.

Lazaridou, A., Dinu, G., and Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July.

Lubin, N. Y., Goldberger, J., and Goldberg, Y. (2019). Aligning Vector-spaces with Noisy Supervised Lexicons. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–465.

Macabante, D. G., Tambanillo, J. C., Cruz, A. D., Ellema, N., Octaviano, M., Rodriguez, R., and Roxas, R. E. (2017). Bi-directional english-hiligaynon statistical machine translation. In *Proceedings of TENCON 2017 – 2017 IEEE Region 10 Conference*, pages 2852–2853, Penang, Malaysia, November.

Melamud, O., McClosky, D., Patwardhan, S., and Bansal, M. (2016). The role of context types and dimensionality in learning word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1030–1040, San Diego, California, June.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, October.

Oco, N. and Roxas, R. (2018). A survey of machine translation work in the Philippines: From 1998 to 2018. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 30–36, Boston, MA, March.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas, November.

Vulić, I. and Korhonen, A. (2016). On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany, August.

Wolfenden, E. P. (1971). *Hiligaynon Reference Grammar*. University of Hawai'i Press.