An Unsupervised Approach for Enrichment of Cross-Lingual Word Embeddings

Intelligent Systems MSc Project

Chris M Warren

(Candidate No: 144166)

January 10, 2018

Abstract

This is the abstract

Acknowledgements

I would like to thank my supervisor Dr Julie Weeds, who has provided me with excellent support and guidance during this project, from the initial concept and theoretical context, to the technical implementation, pacing and planning of the work. Also, a big thank you to my wife, Sarah, who helped me find the courage to revisit academia in later life and has been a constant support throughout this mind-expanding journey.

Contents

Chapter 1

Introduction

1.1 Cross-Lingual Embeddings

Embeddings are the outputs of language modelling algorithms that learn vocabulary representations in continuous vector space. Typically, each embedding vector represents a single word in the vocabulary space. The proximity of the word vectors may be modelled to represent various aspects of linguistic similarity. One such aspect, which many researchers have attempted to model, is the semantic equivalence of words in a vocabulary space of two or more languages. These models are termed *cross-lingual embeddings*, and are a popular focus area in current research.

Figure 1.1 shows an example of English and Chinese words in a cross-lingual embedding space [1]. The embedding space would typically be of the order of 300-500 dimensions. However, the figure has been visualized in 2-dimensions using a dimensional reduction technique, to make the relationships between the word representations more intuitive for humans to interpret graphically.

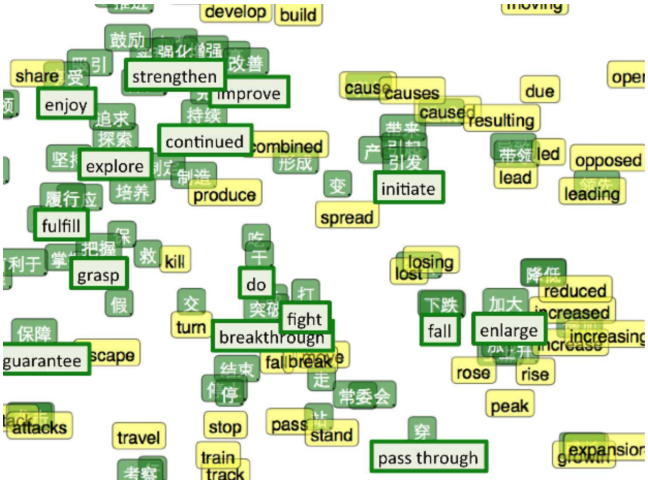


Figure 1.1: A graphical representation of a cross-lingual embedding space. Figure reproduced from (Ruder 2016).

1.1.1 Applications

Cross-lingual embedding models have been proposed as a means to improve the coverage and accuracy of monolingual word similarity, particularly for low-resource languages where there is a shortage of training data available to generate a monolingual distributional similarity model. The intuition is that word similarity relationships in the low-resource language may be learnt by mapping the words into an embedding space shared with a better resourced language. This could assist in transferring other Natural Language Processing (NLP) techniques into low-resource languages, for example, super-sense tagging, part-of-speech tagging, dependency parsing, and named entity recognition (Och and Ney 2003; Ruder 2016).

Cross-lingual embedding models can also be used to derive direct word translations, as is the goal in Machine Translation (MT). However, due to the morphological and idiomatic differences between languages (§1.1.5), most state-of-the-art MT approaches now focus on phrase-based translation. Nevertheless, word-based cross-lingual embeddings could help in

[1] `http://ruder.io/cross-lingual-embeddings/`

lexical induction tasks to locate a source language word within a semantic neighbourhood in the target language, which can give a fuzzy indication of meaning. For some tasks this kind of fuzzy translation will be good enough, and may be attained using more cheaply trained algorithms than state-of-the-art MT. For example, some applications could be information retrieval or document classification across a multi-lingual domain e.g. a set of international news articles, or a multi-lingual Twitter feed.

1.1.2 Training Data

Cross-lingual embedding models have been devised with a variety of training data inputs. The choice of training data requires a balance between quality of signal and the cost/effort required to obtain the data, or in the case of low-resource languages, the feasibility of obtaining it at all (Ruder 2016).

- Word-aligned data - e.g. manually curated bilingual dictionaries. This type of input gives a high quality signal, but is costly to maintain and may not exist for some language pairs.
- Sentence-aligned data - e.g. parallel translations of governmental documentation or religious texts. This type of data is already in existence, covers many low-resource languages and is often freely available from open sources. It has a high quality signal, but the data is only available in moderate quantities.
- Document-aligned data - e.g. Wikipedia pages, or news reports about the same event. These are documents on the same subject matter from different language sources, which are not precisely aligned in their meaning, but are likely to share common terms. This kind of data is freely available in vast quantities from the internet, but provides a very noisy signal.

In this project, we focus on sentence-aligned training data, as this seems to offer the best balance between signal quality and availability of data.

1.1.3 Algorithmic Approaches

Many approaches for learning cross-lingual embeddings have been attempted. The various approaches can be broken out into four categories, based on the way in which they use monolingual and parallel corpus inputs to learn the embeddings (Ruder 2016). Many of the algorithms involve neural network language modelling, which typically requires complex parameter tuning, whereas some methods avoid this by taking a classical matrix factorization approach and are attractive for their relative simplicity.

- Monolingual Mapping - A monolingual embedding space is learnt independently for each language as a pre-processing step. A geometric matrix transformation is then learnt, which enables mapping between the two monolingual representations e.g. (Mikolov, Le, and Sutskever 2013).
- Pseudo-Cross-Lingual - Corpora from each language are merged or shuffled together in some way, to produce a single corpus that mixes words from both languages. A monolingual embedding algorithm is then applied to this pseudo-language corpus e.g. (Vulić and Moens 2016).
- Cross-Lingual Training - The model focuses exclusively on optimizing the cross-lingual objective e.g. (Chandar et al. 2014; Søgaard et al. 2015; Levy, Søgaard, and Goldberg 2017).
- Join Optimization - The model jointly attempts to optimize both the monolingual and cross-lingual objectives at the same time e.g. (Gouws, Bengio, and Corrado 2014).

We will explore the cited algorithms in each category in more detail in the Background chapter (§2). The experimental work in this project will focus on (Levy, Søgaard, and Goldberg 2017)'s method, which is in the Cross-Lingual Training category. However, their argument, based on their systematic comparison of cross-lingual embedding algorithms, is that the feature set used to train the model is a more significant factor than the algorithm used.

1.1.4 Evaluation Approaches

Cross-lingual representation models have been evaluated on a wide range of tasks in the literature. These can be divided into *intrinsic* and *extrinsic* tasks.

Intrinsic tasks are aimed at evaluating the accuracy of the model in isolation, eliminating as far as possible other factors that are not related to the model itself. Typical tasks in this category are *bilingual dictionary induction* i.e. predicting

dictionary translations of words, and *word similarity* i.e. representing similar word vectors in close proximity to one another within the embedding space.

Extrinsic tasks apply the model to some realistic application scenario and assess its usefulness. In this category almost any NLP application could be relevant. However, popular evaluations in previous work have been cross-lingual document classification (CLDC), word alignment, machine translation, sense tagging, part-of-speech tagging, dependency parsing and named entity recognition.

A significant contribution of the (Levy, Søgaard, and Goldberg 2017) paper on which this project is based, is to recreate several previous approaches and make a systematic comparison using the same input feature set and a consistent set of evaluation benchmarks.

1.1.5 Challenges

One of the major challenges of deriving cross-lingual unigram similarity relationships between embedding representations is the morphological and syntactic variation between languages.

Languages will often use a different number of words to convey the same meaning. For example, the the German noun *Schullehrer* would be translated as a bigram in English *school teacher*, the French verb *quitter* would often be translated using a satellite-framed verb phrase in English *to go out*, and the subject of the verb *tenemos* in Spanish is indicated with an inflectional suffix whereas in English we need a separate pronoun to indicate the subject of the verb *we have*.

There is also variation in word order between languages, which can be a challenge for some embedding algorithms that rely on proximity of words within parallel sentence sequences. Words that are close neighbours in one language may be widely separated in another language.

Studies of these *word alignment* challenges in MT pre-date the current research focus on embedding methods, but are relevant foundational concepts that will be explored in the Background (§2.1).

1.2 Problem Context

This study builds on recent work by Levy, Søgaard, and Goldberg 2017, in which a parallel corpus of sentence-aligned translations in 57 languages from the Bible is used to train a multi-lingual embedding model. Other approaches have used document-aligned parallel training corpora, which tend to be more readily and cheaply available in large volumes e.g. crowd-source internet content such as Wikipedia pages, parallel newspaper articles, etc. Whereas Levy et al demonstrated that their approach, using a relatively small corpus (approx 25,000 sentences) yields competitive accuracy and efficiency. Levy et al suggest that this is due to the improved signal strength provided by precise sentence-aligned data, coupled with the signal boosting effect of training simultaneously on a multi-lingual corpus as opposed to separately on individual language pairs.

Using the Bible as a training corpus lends itself particularly well to this approach, as it is freely available in a wide selection of languages and its human translators aim to translate the text precisely 'chapter and verse', which provides high quality sentence alignment. Levy et al also demonstrated the same technique on a corpus of proceedings from the European parliament. This is another multi-lingual domain in which human translators pay close attention to high quality sentence alignment, since precise semantic equivalence is critical in legal proceedings. By contrast, for artistic reasons when translating creative writing, human translators may be less precise in their sentence-alignment, so that their prose reads more idiomatically in the target language. The Europarl corpus is larger than the Bible corpus (approx 180,000 sentences), but contains fewer language translations (21 languages).

One significant disadvantage, however, of training on a Bible corpus for most modern-day applications, is the limited and antiquated vocabulary available in the biblical domain. The Europarl corpus contains a larger vocabulary that is less antiquated, but it still uses vocabulary and predominant words senses typical of the parliamentary domain.

1.3 Aim of Research

The aim of this study is to explore how Levy et al's cross-lingual embedding approach and the Bible multi-lingual parallel corpus could be used as an initial bootstrap step in an enrichment pipeline, resulting in a cross-lingual embedding model that has broader vocabulary coverage and better modern-day relevance. The motivation is to derive a cross-lingual embedding space that is accurate enough to be useful in novel task domains where such high quality sentence-aligned parallel corpora as the Bible and Europarl corpora are not available. The model is evaluated on a bilingual dictionary induction task, using the crowd-sourced multi-lingual dictionary 'Wiktionary'.

A secondary aim is to investigate the efficacy of this approach as a lower cost alternative to prior work, both in terms of the infrastructure required to execute the pipeline, and the use of freely available data sources as training and evaluation corpora. This was partly dictated due to the modest resources available to the masters-level researcher. Nevertheless it is

a worthwhile research aim to explore less resource-intensive approaches in MT. This would be more accessible on a low budget than the current state-of-the-art neural MT approaches, which require large-scale parallel computing capabilities. All the experiments detailed in this study were executed on a single instance commodity laptop [2].

1.4 Experimental Approach

We begin by reproducing Levy et al's experiments using their source code and produce a baseline set of results on the same training and evaluation corpora used by Levy et al. We then explore techniques for enriching cross-lingual embeddings derived from the Bible corpus, using large quantities of monolingual text from social media, a freely available source of high volume, modern text, spanning many varied subject domains. The monolingual enrichment corpora are obtained by searching Twitter and Wikipedia sites, using the respective language vocabularies of the evaluation set as search terms. We propose and evaluate a novel approach for ranking the sentences in the enrichment corpora according to their applicability to the evaluation vocabulary, with the aim of improving the signal to noise ratio and thereby improving the efficiency of the training process.

1.5 Chapter Summary

TBD - revisit after writing each chapter

Chapter 2 reviews the literature, detailing the background context of the problem, the broad families of approaches that have been tried previously and approaches used for benchmarking cross-lingual embedding methods.

In chapter 3 we introduce the data sets used, and document some preliminary exploratory analyses that were carried out.

A detailed description of the cross-lingual embedding model and the proposed new enrichment pipeline is given in chapter 4, while chapter 5 details the evaluation methods used.

The experimental results are presented and evaluated in chapter 6.

A critical appraisal of the work and its implications is presented in chapter 7, including suggestions for future work. Concluding remarks are made in chapter 8.

Appendix A contains a self-assessment checklist covering the ethical concerns for the project, which highlighted no items requiring further ethical review.

Appendix B contains the raw result tables.

A full listing of the code developed in the study will be submitted electronically.

[2] Machine specification: MacBook Pro (Retina, 15-inch, Mid-2015), 2.2GHz quad-core Intel Core i7, 16GB 1600 MHz DDR3 RAM, 60GB SSD, plus 2TB external SSD via USB3.

Chapter 2

Background

In this chapter, we explore the background context of cross-lingual embeddings in the literature. We begin with a section on foundational techniques applied to the problem of word alignment in machine translation, which will be revisited when we explore recent work on cross-lingual embeddings. We then describe state-of-the-art neural network modelling techniques for deriving monolingual embeddings. Lastly, we review in detail the cross-lingual embedding approaches investigated by (Levy, Søgaard, and Goldberg 2017) and the algorithm proposed by them, which is the baseline for this project.

2.1 Word Alignment Algorithms in Machine Translation

An *alignment* is a mapping of a word in the source language to word(s) in the target language, which takes into account the word sequence within the source and target sentences. In the examples in figure (2.1), sentence (a) shows a straightforward 1:1 alignment where the word order is the same in the German source and its English translation, whereas in (b) the word mapping is still 1:1 but the word order differs in the English translation. In sentence (c), one of the German words (*klitzeklein*) is mapped to two words in the English translation (*very small*), whereas in sentence (d) there is a German word (*ja*) which has no equivalent in the translated English sentence. Sentence (e) introduces a NULL token, which is used to indicate a word that has been introduced to make sense of the English translation but has no mapping within the German source sentence. This last example demonstrates several of these complexities occurring within the same sentence, and is more typical of real-world translation tasks.

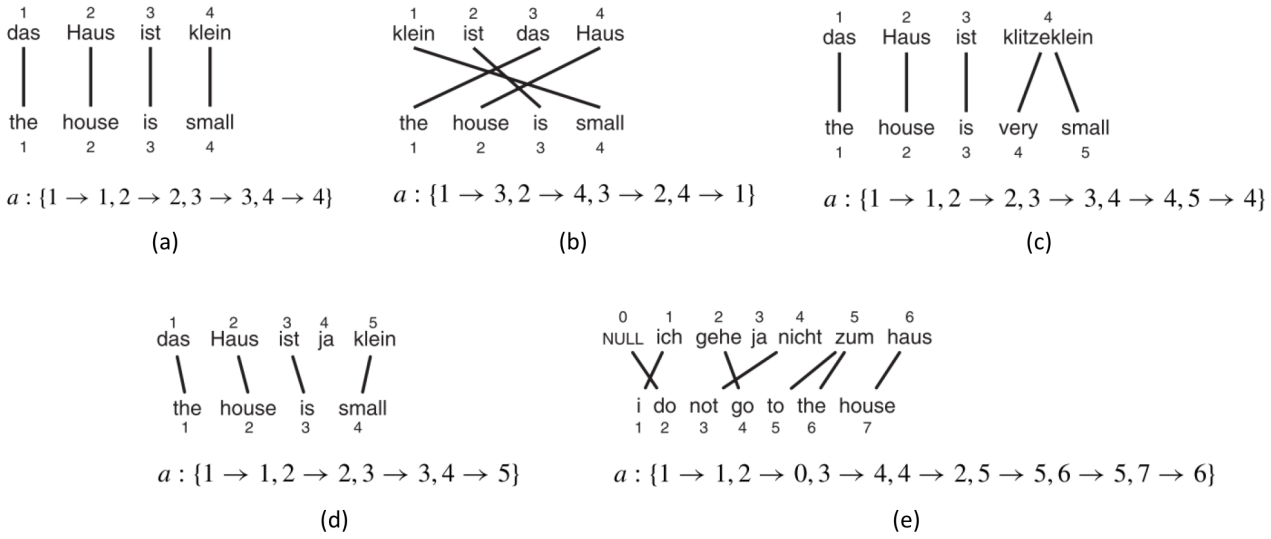


Figure 2.1: Examples of word alignment between English and German sentences. Figure reproduced from (Koehn 2010)

An alignment function a maps each English output word at position i to a foreign input word at position j:

$$a : i \mapsto j \tag{2.1}$$

2.1.1 Statistical Models

2.1.1.1 Lexical Translation Probability Distribution

The task of lexical translation between two natural languages is complicated by the concept of lexical divergence i.e. homonymy and polysemy in both languages. There may be several possible words in the target language that could be given as a correct translation for a word in the source language, which we need to disambiguate. The respective likelihood of each of these translation words can by described by a *lexical translation probability distribution* (Koehn 2010).

More formally, the function:

$$p_f : e \mapsto p_f(e) \tag{2.2}$$

which, given a foreign word f, maps each potential English translation word e to a probability $p_f(e)$, the likelihood of that translation. This *translation* probability will be denoted as the conditional probability function $t(e|f)$, to distinguish it clearly from other probability functions in the equations that follow.

2.1.1.2 IBM Model 1

IBM Model 1 (Brown et al. 1993) is a simple generative model for word-based machine translation based solely on the lexical translation probability distribution and the notion of word alignment. The translation probability of a foreign sentence $\mathbf{f} = (f_1 ... f_{l_f})$ of length l_f to an English sentence $\mathbf{e} = (e_1 ... e_{l_e})$ of length l_e with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \mapsto i$ is computed as follows:

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) \tag{2.3}$$

where ϵ is a normalization parameter, such that the probabilities of all possible English translations e and alignments a sums to one. The denominator of the normalization fraction takes into account the NULL token, so there are $l_f + 1$ words which could be mapped by $(l_f + 1)^{l_e}$ different alignments.

The core term in equation (2.3) is a product of conditional translation probabilities over all generated output words. The Expectation Maximization (EM) algorithm (Koehn 2010) allows the conditional translation probabilities to be *learnt* from a sentence-aligned parallel corpus.

2.1.1.3 HMM Alignment & Higher IBM Models

IBM Model 1 is a simplistic model, which makes several assumptions and has many flaws. For instance, it assumes that each alignment is equally likely, so the best alignment for each word is independent of the decision about best alignments of the surrounding words; whereas in reality, alignments do tend to preserve locality between languages i.e. neighbouring words in the source language are often aligned with neighbouring words in the target language. The model is also weak in terms of adding and dropping words.

A more sophisticated Hidden Markov Model (HMM) alignment approach is often used (Och and Ney 2003). This has a transition probability term in which the current alignment is conditional on the previous one, more effectively utilizing information about the sequence of words in the source sentence.

Further to this, five models of increasing complexity were proposed in the original work by IBM (Brown et al. 1993):

- IBM Model 1: lexical translation
- IBM Model 2: adds absolute alignment model
- IBM Model 3: adds the concept of 'fertility' - the number of target words which will be generated from the source words
- IBM Model 4: adds relative alignment model - probability of a connection depends on the words connected
- IBM Model 5: fixes deficiency - does not waste probability mass on impossible strings in the target language

2.1.2 Heuristic Models

In (Och and Ney 2003)'s comparison study of statistical alignment models, they benchmarked the statistical models described in the previous section against heuristic models. At the time, although the statistical models were already well-developed, the heuristic models were widely favoured in many NLP applications, due to their relative simplicity.

2.1.2.1 Dice Aligner

The Dice Aligner is a simple heuristic model in which the Dice coefficient is used as the similarity function. As with the statistical alignment models, the training input is a sentence-aligned parallel corpus. For each parallel sentence pair, a matrix of similarity scores between every target word e at position i and source word f at position j is obtained:

$$dice(i,j) = \frac{2.C(e_i, f_j)}{C(e_i).C(f_j)} \tag{2.4}$$

where $C(e_i, f_j)$ denotes the count of sentences in which e and f co-occur in the parallel training corpus, and $C(e)$ and $C(f)$ denote the counts of occurrences of e and f in the target and source sentences, respectively.

A simple heuristic is then applied, to select the word alignments that maximize the Dice similarity scores:

$$a_j = \arg\max_i \{dice(i,j)\} \tag{2.5}$$

Take the following sentence-aligned parallel corpus as a very small-scale example:

$$\begin{aligned} s_{e_0} &= \text{[the cat sat on the mat]} \\ s_{e_1} &= \text{[the cat chased the mouse]} \\ s_{e_2} &= \text{[the mouse hid under the mat]} \\ \\ s_{f_0} &= \text{[el gato se sentó en la alfombra]} \\ s_{f_1} &= \text{[el gato persiguió al ratón]} \\ s_{f_2} &= \text{[el ratón se escondió debajo de la alfombra]} \end{aligned} \tag{2.6}$$

The correct alignment result for sentence s_0 is shown in figure (2.2). We will now demonstrate the Dice aligner algorithm and attempt to learn the alignments.

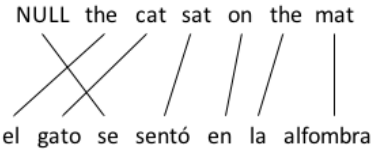


Figure 2.2: The correct English/Spanish alignments for the example sentence *The cat sat on the mat*

For sentence s_0, we place the sentence co-occurrence counts over the whole corpus into the matrix $C(e_i, f_j)$ as follows:

$$C(e_i, f_j) = \begin{array}{c} \\ el \\ gato \\ se \\ sentó \\ en \\ la \\ alfombra \end{array} \begin{array}{c} \begin{array}{ccccc} the & cat & sat & on & mat \end{array} \\ \begin{pmatrix} 3 & 2 & 1 & 1 & 2 \\ 2 & 2 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 1 & 2 \end{pmatrix} \end{array} \tag{2.7}$$

The Dice similarity score is evaluated using equation (2.4, to derive the $dice(i,j)$ matrix:

$$dice(i,j) = \begin{array}{c} \\ el \\ gato \\ se \\ sentó \\ en \\ la \\ alfombra \end{array} \begin{array}{c} \begin{array}{ccccc} the & cat & sat & on & mat \end{array} \\ \begin{pmatrix} \frac{2\times3}{6\times3} & \frac{2\times2}{2\times3} & \frac{2\times1}{1\times3} & \frac{2\times1}{1\times3} & \frac{2\times2}{2\times3} \\ \frac{2\times2}{6\times2} & \frac{2\times2}{2\times2} & \frac{2\times1}{1\times2} & \frac{2\times1}{1\times2} & \frac{2\times2}{2\times2} \\ \frac{2\times2}{6\times2} & \frac{2\times1}{2\times2} & \frac{2\times1}{1\times2} & \frac{2\times1}{1\times2} & \frac{2\times2}{2\times2} \\ \frac{2\times1}{6\times1} & \frac{2\times1}{2\times1} & \frac{2\times1}{1\times1} & \frac{2\times1}{1\times1} & \frac{2\times1}{2\times1} \\ \frac{2\times1}{6\times1} & \frac{2\times1}{2\times1} & \frac{2\times1}{1\times1} & \frac{2\times1}{1\times1} & \frac{2\times1}{2\times1} \\ \frac{2\times2}{6\times2} & \frac{2\times1}{2\times2} & \frac{2\times1}{1\times2} & \frac{2\times1}{1\times2} & \frac{2\times2}{2\times2} \\ \frac{2\times2}{6\times2} & \frac{2\times1}{2\times2} & \frac{2\times1}{1\times2} & \frac{2\times1}{1\times2} & \frac{2\times2}{2\times2} \end{pmatrix} \end{array} = \begin{array}{c} \\ el \\ gato \\ se \\ sentó \\ en \\ la \\ alfombra \end{array} \begin{array}{c} \begin{array}{ccccc} the & cat & sat & on & mat \end{array} \\ \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{1}{3} & \boxed{1} & \boxed{1} & \boxed{1} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \boxed{1} & \boxed{1} & \boxed{1} \\ \frac{1}{3} & 1 & \boxed{2} & \boxed{2} & 1 \\ \frac{1}{3} & 1 & \boxed{2} & \boxed{2} & 1 \\ \frac{1}{3} & \frac{1}{2} & \boxed{1} & \boxed{1} & \boxed{1} \\ \frac{1}{3} & \frac{1}{2} & \boxed{1} & \boxed{1} & \boxed{1} \end{pmatrix} \end{array} \tag{2.8}$$

The boxed values in matrix (2.8) show which English words give the maximum Dice scores for each Spanish word i.e. the result of $\arg\max_i \{dice(i,j)\}$. There are several ties for the maximum association score. This is probably due in part to the artificially small size of the example corpus, as it is unlikely there is sufficient signal to accurately learn the alignments. However, we can see even with such little input data, that the result does correctly indicate a high degree of correlation between some of the words, for example between *sat on* and *sentó en*. The algorithm is less successful on common words that carry little meaning, such as the definite article *the*, which translates into several forms in the Spanish (masculine and feminine:*el* and *la*; and contracted with the preposition *a* in *al*).

2.1.3 Statistical vs Heuristic Comparison

Och and Ney suggested that statistical methods are more mathematically coherent, whereas the choice of similarity function used in heuristic models was often completely arbitrary. They compared two variations on the heuristic Dice coefficient model with IBM Models 1-5, Hidden Markov Models (HMMs) and their own novel proposal, IBM Model 6. They demonstrated that the statistical models consistently outperformed the Dice heuristic model.

Subsequently, these statistical alignment models have become the foundation of modern phrase-based statistical machine translation models, of which the current state-of-the-art neural translation models are a further extension.

2.1.4 Word Alignment & Cross-Lingual Embeddings

In the Introduction, we described some of the word alignment challenges that are inherent to the cross-lingual embedding problem. These two problems are closely related.

There is a category of cross-lingual embedding algorithms (Klementiev, Titov, and Bhattarai 2012; Faruqui and Dyer 2014) which overcomes these challenges by depending upon the word alignment algorithms described here as a pre-processing step. The automatically induced word alignments are the inputs to the cross-lingual embedding algorithm.

In section (2.3.3), we will also see how (Levy, Søgaard, and Goldberg 2017) have drawn parallels between state-of-the-art cross-lingual embeddings algorithms and the Dice aligner to derive their SID-SGNS method that forms the baseline for this project.

Furthermore, one of the frequently used benchmark evaluations for cross-lingual embedding algorithms is the task of discovering word alignments in parallel text.

2.2 Monolingual Neural Network Language Models

The monolingual Recurrent Neural Network Language Model (RNNLM) architectures proposed by (Mikolov et al. 2013b) have become extremely popular in the distributional similarity problem space. Mikolov et al built upon previous neural network language models to maximize accuracy while minimizing computational complexity, enabling the models to be trained over very large datasets. They provide two distinct model implementations in the word2vec framework, CBOW and SkipGram, described in section (2.2.2).

2.2.1 Learning Embeddings Using RNNLMs

A useful side-effect of training an RNNLM is that word embeddings are implicitly generated. Figure 2.3 shows a simple RNNLM, which comprises an input layer, a hidden layer with recurrent connections, an output layer and the corresponding weight matrices. The input vector $w(t)$ represents the current word in a sequence of T training words $w_1, w_2, w_3, ...w_T$ belonging to a vocabulary V of size $|V|$. The output vector $y(t)$ generates a probability distribution over the set of words in the vocabulary. Due to the recurrent connections, the hidden layer $s(t)$ has a historic input $s(t-1)$, creating a feedback loop which enables it to maintain a cumulative representation of the sentence history.

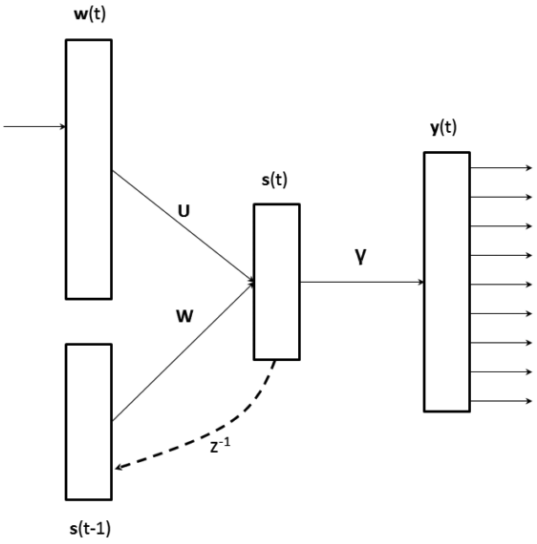


Figure 2.3: Recurrent Neural Network Language Model. Figure reproduced from (Mikolov, Yih, and Zweig 2013)

The hidden and output layer values are computed as:

$$s(t) = f(Uw(t) + Ws(t-1)) \tag{2.9}$$

$$y(t) = g(Ys(t)) \tag{2.10}$$

where the respective activation functions at each layer are:

$$f(z) = \frac{1}{1+e^{-z}}, g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{2.11}$$

The RNN model is trained with back-propagation, with the explicit training objective to maximize the log-likelihood J_θ of the data with respect to the model parameters θ using a generative approach:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log y(t) \tag{2.12}$$

However, an implicit outcome of this process is that the column values of the learnt weight matrix U may be used as vectorized word representations in continuous space (embeddings). Since similar words are likely to be found in similar contexts, the intuition is that these embeddings can be used in a wide variety of distributional similarity tasks that are unrelated to the original training objective.

2.2.2 word2vec Log-Linear Models

Prior to (Mikolov et al. 2013b), the term *embeddings* was first coined by (Bengio et al. 2003) and the idea was further developed by (Collobert and Weston 2008). The two main benefits of Mikolov et al's *word2vec* models were that they were less expensive to train, due to the removal of the costly hidden layer, and that they allowed the language model to take additional context into account.

Within a context window, the Continuous Bag-of-words (CBOW) model is used to predict the current word w_t given the context $[w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}]$, whereas the Skip-gram model is used to predict the context $[w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}]$ given the current word w_t.

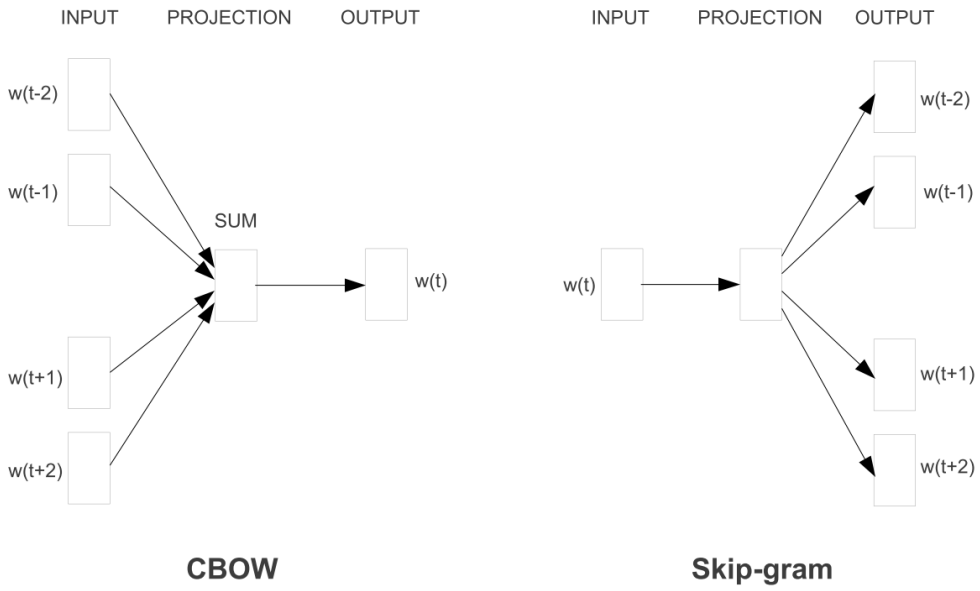


Figure 2.4: The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word. Figure reproduced from (Mikolov et al. 2013b)

2.2.3 Continuous Bag-of-Words (CBOW)

The RNNLM used a recurrent self-connection as an input to capture a cumulative representation of the sentence history as it parses through the sentence, with the aim of predicting the next word in the sentence. Whereas in the CBOW model, since the goal is to obtain word embeddings over the whole vocabulary, the model can take into account both past and future text to make its predictions. A feed-forward NNLM is used with an input window centred on the current word and including both the n preceding and subsequent words in the sequence. It is known as Continuous Bag-of-Words, since it uses a continuous representation in which the order of words in the projection is of no importance.

The objective function is similar to the RNNLM log-likelihood function, but takes the n words before and after the current word as the context window.

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-n}, ..., w_{t-1}, w_{t+1}, ..., w_{t+n}) \tag{2.13}$$

2.2.4 Continuous Skip-gram with Negative Sampling (SGNS)

The Skip-Gram model (Mikolov et al. 2013a) has the reverse architecture of the CBOW model; that is, instead of predicting the current word based on the context, it uses a log-linear classifier with continuous projection layer to predict words within a certain range of the current word.

The conditional probability term in the objective function is therefore reversed:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \sum_{-n\leq j\leq n, \neq 0} \log p(w_{t+j}|w_t) \tag{2.14}$$

There is an additional summation term to account for some modifications made to the typical n-gram word context definition. Firstly, the context window has dynamic size, such that the parameter k denotes the *maximal* window size. Context words are sampled uniformly for window sizes $k' = \{1, ..., k\}$. Secondly, words appearing less than a minimum threshold number of times in the corpus are not considered as words or contexts; and words above a certain frequency threshold are down-sampled, which has the effect of increasing the effective size of the window. This enables context words which carry useful topical content and which are relatively remote from the current word to be considered, making the learnt similarities more topical (Goldberg and Levy 2014).

2.2.5 Linguistic Properties of Embeddings

Although there is no knowledge of syntax or semantics built into the RNNLM, it has been demonstrated that the generated word representations can have interesting syntactic and semantic properties (Mikolov, Yih, and Zweig 2013). In high-dimensional space, the vector offsets between words can demonstrate consistent geometric relations that model various aspects of syntactic and semantic similarity, as shown in Figure 2.5.

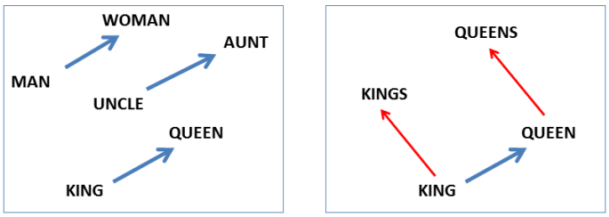


Figure 2.5: Vector offsets for gender relations (left) and singular/plural relations (right). For illustration, the high-dimensional vector space has been reduced to 2D using Principal Component Analysis (PCA), with different projections to show the different relationships. Figure reproduced from (Mikolov, Yih, and Zweig 2013)

2.2.6 Cross-Lingual Similarity in Embeddings Space

Mikolov et al applied this concept of linguistic regularity in word representations to the problem of machine translation (Mikolov, Le, and Sutskever 2013), and observed that these geometric relations can hold between languages. Figure 2.6 shows an example of relations between numbers and animals in English and Spanish, demonstrating similar geometric configuration of corresponding source/target words in each respective monolingual embedding space. They experimented with an approach that learns a linear projection matrix using stochastic gradient descent, to translate word vectors from the source language embedding space to the target language embedding space, using a small (5,000 word) bilingual dictionary as a training corpus.

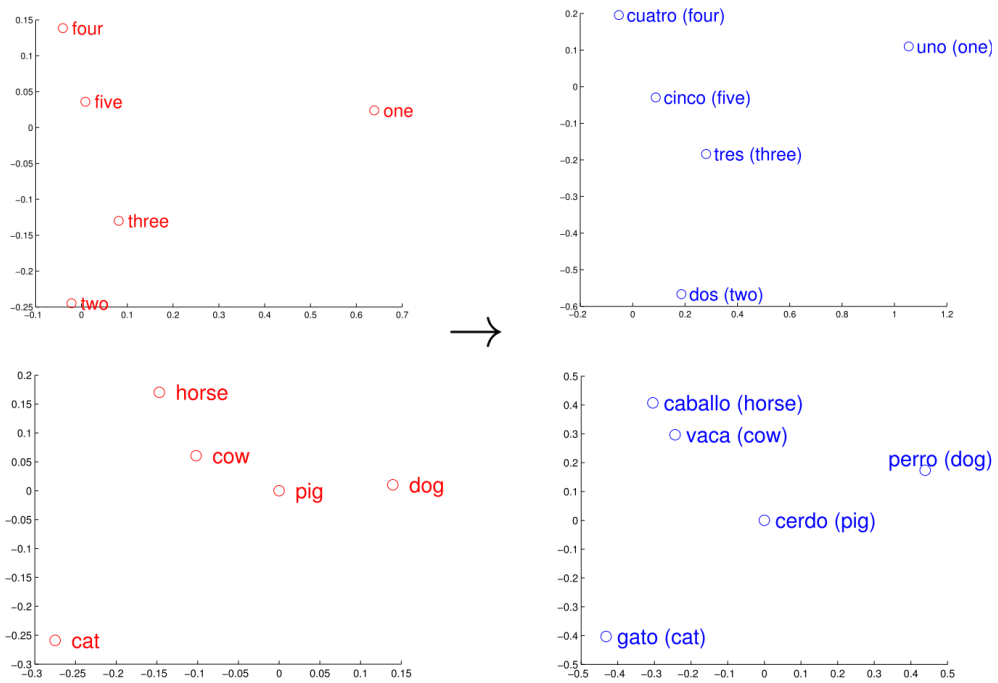


Figure 2.6: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). Figure reproduced from (Mikolov, Le, and Sutskever 2013)

2.3 Learning Cross-Lingual Embeddings from Sentence Alignments

Having laid out some of the foundational concepts of word alignment and word embedding algorithms, we now move on to discuss the ideas and models presented in (Levy, Søgaard, and Goldberg 2017), which is the baseline starting point for this project. Levy et al re-evaluated several previously proposed cross-lingual embedding algorithms, which are outlined in (§2.3.2). They focused on the granularity of alignment of the training inputs, which can be divided into three categories: word-level alignments, document-level alignments and sentence-level alignments (§2.3.1).

They established the importance of sentence-aligned training inputs, and then contrasted the sentence-aligned cross-lingual embedding algorithms with the sentence ID feature space used in word alignment algorithms such as IBM Model 1 (Brown et al. 1993). They demonstrated that the sentence-aligned embedding algorithms and the alignment algorithms both perform similarly on word alignment and bilingual dictionary induction tasks.

They proposed a new cross-lingual embedding algorithm SID_SGNS (skip-gram with negative sampling using sentence IDs), described in (§2.3.3), which is a generalization of the Dice aligner (Och and Ney 2003). They used this to simultaneously train a multi-lingual embedding space using all 57 languages in the Bible corpus, taking advantage of the combined signal this provides to achieve better performance on the dictionary induction task than the best-performing prior art.

2.3.1 Alignment Granularity of Training Inputs

Approaches based on *word-aligned* training inputs have used either bilingual dictionaries (Mikolov, Le, and Sutskever 2013; Xiao and Guo 2014), or automatically generated word alignments (Klementiev, Titov, and Bhattarai 2012; Faruqui and Dyer 2014). The disadvantage of depending on bilingual dictionaries is that they may not exist for many low-resourced language pairs, while the accuracy achieved in approaches using automatically generated word alignments depends predominantly on the accuracy of the word alignment algorithm, rather than the word-embedding algorithm itself.

Document-aligned training corpora are usually obtained from comparable texts from different language sources, which are not necessarily translations, such as Wikipedia articles on the same topic in different languages, or new reports about the same event. These approaches (e.g. Søgaard et al. 2015; Vulić and Moens 2016) aim to make use of training data that has a much weaker signal strength, but is readily available in a wide variety of domains and in huge quantities resulting in computationally expensive training algorithms.

The third category of *sentence-aligned* parallel corpora provide a clean and accurate signal over a relatively small corpus, and training data from sources such as the Bible is freely available in a wide range of languages.

2.3.2 Baseline Cross-Lingual Embedding Algorithms

Levy et al reproduce four algorithms from prior work as the baseline for their study of sentence-aligned training approaches for generating cross-lingual embeddings: BilBOWA (Gouws, Bengio, and Corrado 2014), BWE-SkipGram (Vulić and Moens 2016), Autoencoders (Chandar et al. 2014) and Inverted Index (Søgaard et al. 2015).

2.3.2.1 BilBOWA

Bilingual Bag-of-Words without Alignments (Gouws, Bengio, and Corrado 2014) uses a *joint-optimization* approach; that is, monolingual models are trained on large corpora in the two respective languages using an objective function $\mathcal{L}(.)$, while a much smaller parallel corpus of sentence-aligned text is used to apply a cross-lingual regularization function $\Omega(.)$, which constrains the two monolingual models such that translation word pairs are assigned similar embeddings in the two languages. The motivations for this approach are to make use of corpora that are cheaply available i.e. two very large monolingual corpora combined with a much smaller sentence-aligned parallel corpus, and to improve on the computational efficiency of prior approaches. The joint-optimization approach was originally proposed by (Klementiev, Titov, and Bhattarai 2012) and extended by (Zou et al. 2013). Gouws et al used a skip-gram with negative sampling (SGNS) model for the monolingual objective function $\mathcal{L}(.)$, similarly to Zou et al, but they proposed a novel computationally-efficient cross-lingual regularization function $\Omega(.)$.

To optimize an *exact cross-lingual objective* between the two monolingual embedding matrices entails two key challenges: deriving or learning the alignments between the words in the respective vocabularies, which is both costly and noisy; and efficiently evaluating $\Omega(.)$ during training, since a naïve evaluation would scale as the product of the two vocabulary sizes $O(|V_e|.|V_f|)$ at each training step.

To achieve improved computational efficiency, Gouws et al assume a uniform word-alignment model, such that each source language word can potentially be aligned with each target language word for each observed sentence pair in the sentence-aligned corpus. They apply this assumption iteratively, sampling one sentence pair at a time, minimizing the L_2

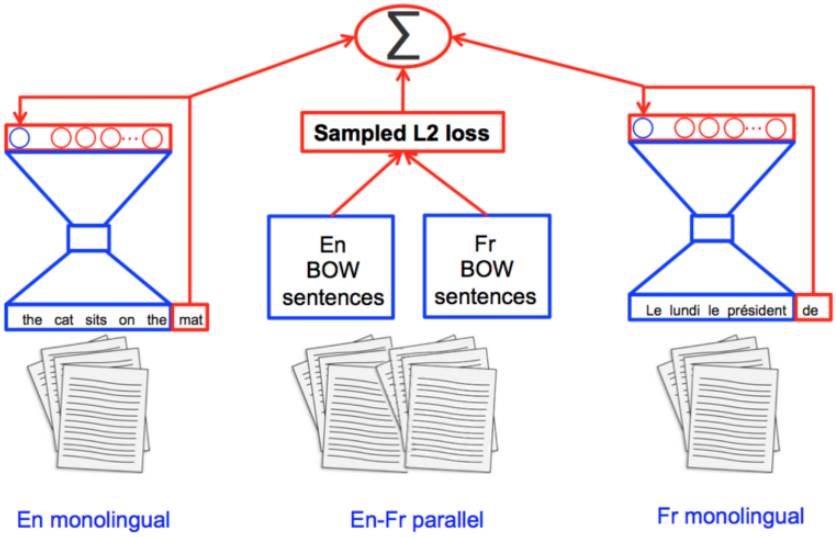


Figure 2.7: Joint-optimization: source and target monolingual models are constrained by a cross-lingual regularization function learnt from a smaller sentence-aligned corpus. Figure reproduced from (Gouws, Bengio, and Corrado 2014)

loss between the bag-of-words sentence vectors of the parallel sentences. The efficiency gain is achieved since the length of each sentence computed at each iteration is much less than the vocabulary sizes.

The BilBOWA-loss, a *sampled approximate cross-lingual objective* is computed as:

$$\Omega_{\mathbf{A}}^{(t)}(\mathbf{R}^e, \mathbf{R}^f) = ||\frac{1}{m}\sum_{w_i \in s^e}^{m} \mathbf{r}_i^e - \frac{1}{n}\sum_{w_j \in s^f}^{n} \mathbf{r}_j^f||^2 \tag{2.15}$$

where Ω is calculated over training iteration steps t and the assumed uniform word-alignment matrix of the sentence pair at each iteration step $\mathbf{A}$, and $\mathbf{R}$ denotes the word-embedding matrices in English e and French f respectively, $\mathbf{r}$ denotes an individual word-embedding row vector, m and n are the lengths of the English and French sentences respectively, and w denotes a word position i in the English version and position j in the French version of the aligned sentence pair.

2.3.2.2 BWE-SkipGram

The BWE-SkipGram approach (Vulić and Moens 2016) uses document-aligned corpora and uses a merging algorithm to combine words from an aligned document pair into a single *pseudo-bilingual document*. A monolingual embedding model such as SGNS (Mikolov et al. 2013a) is then trained on these pseudo-bilingual documents. (Note: BWE stands for 'bilingual word embeddings').

Again, their motivation is to demonstrate that bilingual word-embeddings may be derived from cheaply available data. In this case, document-aligned data, such as Wikipedia articles on the same topic, or news articles reporting on the same event. They also recommend their approach for its simplicity.

They experiment with two approaches for merging the documents, a random shuffle and a deterministic length-ratio shuffle, which is more likely to give consistent results. The deterministic approach is shown in Figure (2.8). Document lengths are denoted as m_S and m_T for an aligned tokenized document pair (d_j^S, d_j^T). Assuming $m_S \geq m^T$, the algorithm proceeds as follows (if, in fact, $m_T > m_S$, then the roles of d_j^S and d_j^T are simply reversed):

1. Pseudo-bilingual document d'_j is empty: $d'_j =$
2. Compute ratio: $R = \lfloor \frac{m_S}{m_T} \rfloor$
3. Scan through aligned documents d_S and d_T simultaneously and (3.1) append R word tokens from d_j^S into d'_j; then (3.2) append 1 word token from d_j^T. Repeat steps 3.1 and 3.2 until all word tokens from d_j^T have been inserted into d'_j.
4. Insert remaining $m_S \mod m_T$ word tokens from d_j^S into d'_j.

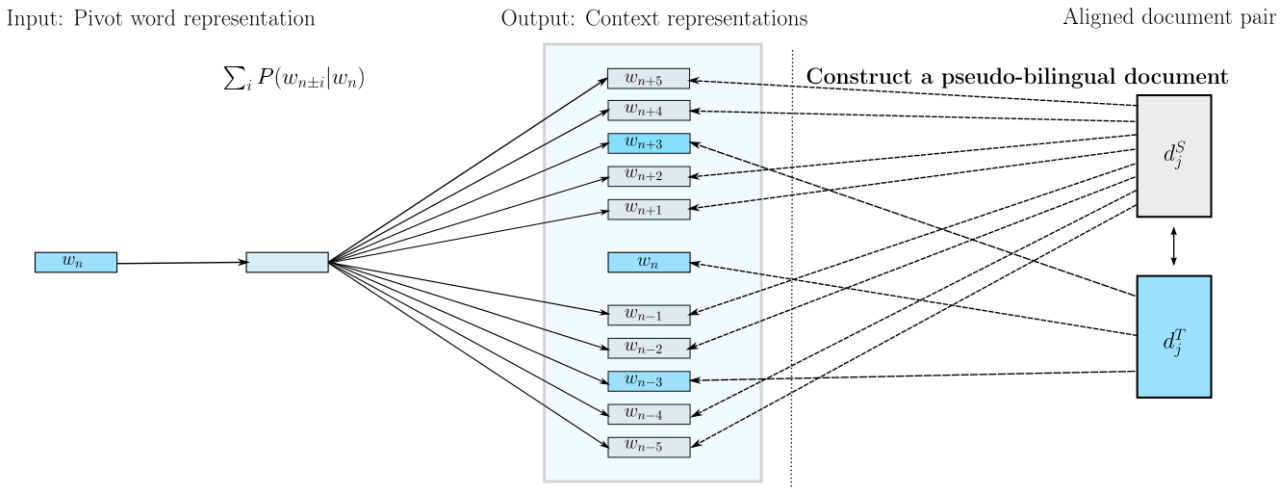


Figure 2.8: Pseudo-bilingual document generation: Length-ratio shuffle. Figure reproduced from (Vulić and Moens 2016)

The right-hand side of Figure (2.8) shows construction of a pseudo-bilingual document in the case where $R = 2$. A typical Skip-gram architecture shown on the left-hand side, similar to Mikolov et al's model in Figure (2.4), is used to train the model on the generated pseudo-bilingual document corpus.

An example follows, using a paragraph of parallel text from the original English and the Spanish translation of *Harry Potter and the Philosopher's Stone* (Rowling 1997, Spanish translation by Alicia Dellepiane):

$d_j^S = \{$ Nothing like this man had ever been seen in Privet Drive . He was tall , thin and very old , judging by the silver of his hair and beard , which were both long enough to tuck into his belt . He was wearing long robes , a purple cloak which swept the ground and high -heeled , buckled boots . His blue eyes were light , bright and sparkling behind half -moon spectacles and his nose was very long and crooked , as though it had been broken at least twice . This man 's name was Albus Dumbledore . $\}$ $m_S = 102$

$d_j^T = \{$ En Privet Drive nunca se había visto a un hombre así . Era alto , delgado y muy anciano , a juzgar por su pelo y barba plateados , tan largos que habría podido sujetarlos con el cinturón . Llevaba una túnica larga , una capa color púrpura que barría el suelo y botas con tacón alto y hebillas . Sus ojos azules eran claros , brillantes y centelleaban detrás de unas gafas de cristales de media luna , y su nariz era muy larga y torcida , como si se la hubiera fracturado alguna vez . El nombre de aquel hombre era Albus Dumbledore . $\}$ $m_T = 105$

$m_T > m_S$, therefore: $R = \lfloor \frac{m_T}{m_S} \rfloor = \lfloor \frac{105}{102} \rfloor = \lfloor 1.0294 \rfloor = 1$

$d'_j = \{$ En Nothing Privet like Drive this nunca man se ever había been visto in a Privet un Drive hombre . así He . was Era tall alto , , thin delgado and y very muy old anciano , , judging a by juzgar the por **silver** su of pelo his y hair barba and **plateados** beard , , tan which largos were que both habría long podido enough sujetarlos to con tuck el into cinturón his . belt Llevaba . una He túnica was larga wearing , long una robes capa , color a púrpura purple que cloak barría which el swept suelo the y ground botas and con high tacón -heeled alto , y **buckled hebillas** boots . . Sus His ojos **blue azules** eyes eran were claros light , , brillantes bright y and centelleaban sparkling detrás behind de half unas -moon gafas spectacles de and cristales his de nose media was luna very , long y and su **crooked** nariz , era as muy though larga it y had **torcida** been , broken como at si least se twice la . hubiera This fracturado **man** alguna 's vez name . was El Albus nombre Dumbledore de . aquel **hombre** era Albus Dumbledore . $\}$

In a closely sentence-aligned parallel document such as this example, the shuffling process can sometimes result in word translations being in close proximity (e.g *blue* and *azules*, or *buckled* and *hebillas* above), but due to variations in word order in the two languages and variations in word-level alignments within phrases, it can often result in word translations being more distant from one another (e.g. *silver* and *plateados*, *crooked* and *torcida*, or *man* and *hombre*). Recall that the Skip-gram rules of the (Mikolov et al. 2013a)'s SGNS algorithm will down-sample frequent tokens, such as punctuation and stop-words, which will tend to have the effect of bringing meaning-laden words from the English and Spanish versions together into the same context window in cases where they have become separated by the shuffling algorithm.

Since the focus of (Levy, Søgaard, and Goldberg 2017) is on the importance of the sentence-alignment signal in the training input, they re-apply this pseudo-bilingual merging technique to sentence-aligned training data to produce their baseline.

2.3.2.3 Autoencoders

Whereas (Mikolov et al. 2013b) based their word2vec models on RNNLM architecture, the bilingual autoencoder approach proposed in (Chandar et al. 2014) is based upon a different neural network model, which had previously been used in the monolingual form shown in Figure (2.9).

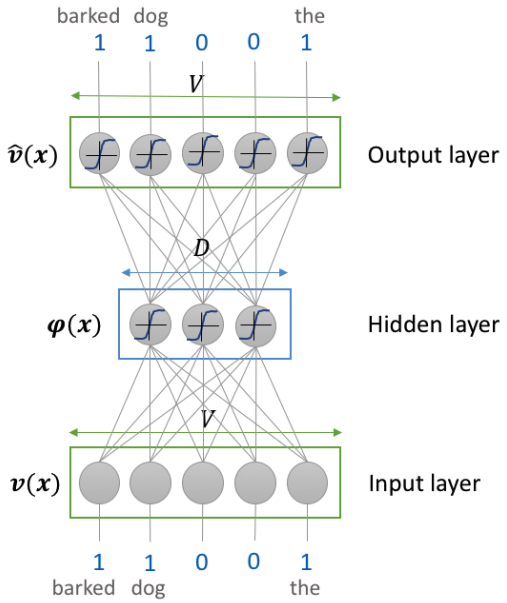


Figure 2.9: Monolingual bag-of-words autoencoder. The sigmoid symbols shown in the nodes of the hidden and output layers indicate where non-linear activation functions are applied.

The monolingual autoencoder takes as its input a sentence $\mathbf{x}$ in the form of a bag-of-words $\{x^{(t)}\}_{t=1}^{T}$, with x_i a word index in a fixed vocabulary of V words. The sentence is expressed as a sparse binary vector $\mathbf{v}(\mathbf{x})$ of dimension V, such that $\mathbf{v}(\mathbf{x})_{x_i}$ is 1 if the word x_i is present in the sentence and 0 otherwise.

The objective is to learn generalized encoder and decoder functions, such that the model encodes any $\mathbf{v}(\mathbf{x})$ to a D-dimensional representation $\phi(\mathbf{x})$ at the hidden layer and then decodes that back to an approximation of the starting state $\hat{\mathbf{v}}(\mathbf{x})$.

The encoder function is of the form:

$$\phi(\mathbf{x}) = \mathbf{h}(\mathbf{c} + \mathbf{W}\mathbf{v}(\mathbf{x})) \tag{2.16}$$

where $\mathbf{h}(.)$ is a non-linear activation function such as sigmoid or hyperbolic tangent, and $\mathbf{c}$ is a D-dimensional bias term.

The decoder function is of the form:

$$\hat{\mathbf{v}}(\mathbf{x}) = \sigma(\mathbf{V}\phi(\mathbf{x}) + \mathbf{b}) \tag{2.17}$$

where $\mathbf{V} = \mathbf{W}^T$, $\mathbf{b}$ is V-dimensional bias term and $\sigma(a) = \frac{1}{1+e^{-a}}$ the sigmoid function. The model is trained by minimizing a loss function $\ell(\mathbf{v}(\mathbf{x}))$ using stochastic or mini-batch gradient descent (Chandar et al. 2014).

This technique might typically be used in a neural network architecture to reduce data from a large sparse representation down to a lower-dimensional representation at the hidden layer, whilst minimizing loss of information. However, a useful by-product is that word representations (embeddings) are found in the columns of the matrix $\mathbf{W}$.

Chandar et al extended the monolingual autoencoder to propose a new technique for learning bilingual word representations (embeddings). Assuming an aligned sentence pair $(\mathbf{x}, \mathbf{y})$, in source language $\mathcal{X}$ and target language $\mathcal{Y}$ respectively, we would like to learn to encode/decode bags-of-words bi-bidirectionally from source to target language and vice versa. The two languages may have different vocabulary sizes $V^{\mathcal{X}}$ and $V^{\mathcal{Y}}$, but the representation at the hidden layer will be of the same size D. We define language-specific word representation matrices $\mathbf{W}^{\mathcal{X}}$ and $\mathbf{W}^{\mathcal{Y}}$, which are

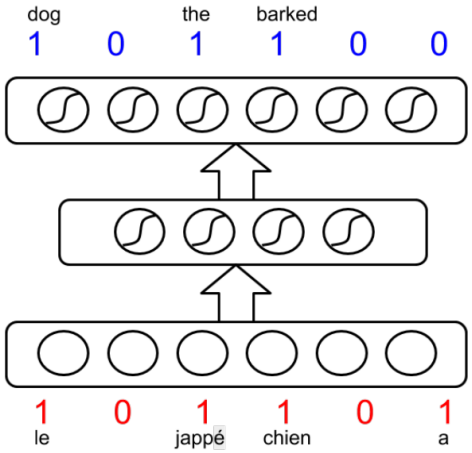


Figure 2.10: Bilingual bag-of-words autoencoder. The English sentence *the dog barked* is reconstructed from the French sentence *le chien a jappé*. Figure reproduced from (Chandar et al. 2014)

used in the respective encoder functions as shown:

$$\phi(\mathbf{x}) = \mathbf{h}(\mathbf{c} + \mathbf{W}^{\mathcal{X}}\mathbf{v}(\mathbf{x})), \quad \phi(\mathbf{y}) = \mathbf{h}(\mathbf{c} + \mathbf{W}^{\mathcal{Y}}\mathbf{v}(\mathbf{y})) \tag{2.18}$$

Each language also has its own respective decoder parameters $(\mathbf{b}^{\mathcal{X}}, \mathbf{V}^{\mathcal{X}})$ and $(\mathbf{b}^{\mathcal{Y}}, \mathbf{V}^{\mathcal{Y}})$. Chandar et al also added a cross-lingual regularization term to the objective function, to encourage high correlation.

2.3.2.4 Inverted Index

The *inverted index* method of (Søgaard et al. 2015) is a strikingly simple count-based approach, which aims to make use of the large-scale, freely available, cross-lingual information in Wikipedia to create inter-lingual representations. Wikipedia is a crowd-sourced encyclopedia which contains articles in numerous languages linked cross-lingually into an ontology of *concepts*.

Across a set of languages, they identify the common subset of Wikipedia concepts, and create a matrix of terms used to describe that concept across the different languages (e.g. eq. 2.19). They then invert the concept-to-term matrix, to represent a word by the Wikipedia concepts it is used to describe. They perform dimensionality reduction using Singular Value Decomposition (SVD).

$$M_{(word,concept)} = \begin{array}{c} \\ w_{en_1} \\ w_{en_2} \\ \vdots \\ \hline w_{fr_1} \\ w_{fr_2} \\ \vdots \\ \hline w_{de_1} \\ w_{de_2} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} C_1 & C_2 & C_3 & C_4 & C_5 & \cdots \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & \cdots \\ 0 & 1 & 1 & 0 & 0 & \cdots \\ & & \vdots & & & \\ \hline 1 & 0 & 0 & 1 & 1 & \cdots \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ & & \vdots & & & \\ \hline 1 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & \cdots \\ & & \vdots & & & \end{pmatrix} \end{array} \tag{2.19}$$

The main benefits of this approach are its simplicity (far fewer parameters to optimize than in the neural network models), it does not rely on the availability of sentence-aligned parallel corpora or word-aligned lexicons, and it allows for simultaneous multi-lingual training leading to truly inter-lingual word representations.

In the baseline tests carried out by (Levy, Søgaard, and Goldberg 2017), they re-apply this method using sentence-aligned training data, so that they can benchmark it alongside other methods using the same training corpus.

2.3.3 SGNS with Sentence IDs

Finally, we discuss the SGNS with Sentence ID method, proposed in (Levy, Søgaard, and Goldberg 2017), which is the starting point for the investigations carried out in this project. They reflect on the use of word-feature matrices and

vector similarity metrics (e.g. cosine similarity) in the word similarity literature. They re-frame several of the baseline methods discussed in the previous sections in terms of word-feature matrices, association metrics and matrix factorization approaches.

2.3.3.1 Word-Feature Matrices

Given a vocabulary space V_W and a feature space V_F, a word-feature matrix M of dimensions $|V_W| \times |V_F|$ may be defined, where each entry in M represents some statistic of the word and feature combination. This may be a large sparse matrix of raw binary co-occurrence metrics, or smarter association metrics may also be used, such as:

L_1 Row Normalization

$$M^{L_1}_{w,f} = \frac{I(w,f)}{I(w,*)} \tag{2.20}$$

Inverse Document Frequency (IDF)

$$M^{IDF}_{w,f} = \log \frac{|V_F|}{I(w,*)} \tag{2.21}$$

Pointwise Mutual Information (PMI)

$$M^{PMI}_{w,f} = \log \frac{\#(w,f) \cdot \#(*,*)}{\#(w,*) \cdot \#(*,f)} \tag{2.22}$$

where $I(w,f)$ is the co-occurrence binary indicator function, and $\#(w,f)$ is the co-occurrence count function. The $*$ represents a wildcard value, such that any function with a wildcard argument should be interpreted as the sum of all possible instantiations, e.g. $I(w,*) = \sum_x I(w,x)$.

Since V_W and V_F may be extremely large and the word/feature matrix is often very sparse, the matrix is typically decomposed to lower dimensions before extracting the word vector embeddings. This can be achieved using Singular Value Decomposition (SVD), as in Inverted Index (§2.3.2.4). Alternatively, the SGNS algorithm (Mikolov et al. 2013a) has been shown to be equivalent to factorization of M^{PMI} using a weighted non-linear objective (Goldberg and Levy 2014).

2.3.3.2 Generalized Dice

Levy et al demonstrate that the Dice coefficient (Och and Ney 2003) which we discussed with an example in section (§2.1.2.1; eq. 2.4) is mathematically equivalent to the dot-product of two L_1-normalized sentence-ID word-vectors, multiplied by 2. i.e.

$$w_s \cdot w_t = \frac{dice(w_s, w_t)}{2} \tag{2.23}$$

They demonstrate their claim, firstly by expressing the dot-product in terms of the co-occurrence binary indicator function $I(w,f)$ introduced earlier:

$$w_s \cdot w_t = \sum_i \left(\frac{I(w_s,i)}{I(w_s,*)} \cdot \frac{I(w_t,i)}{I(w_t,*)} \right) \tag{2.24}$$

where i is the index of the aligned sentence.

TBD - Try to paraphrase somehow to indicate more of my own understanding?
Since $I(w_s,*) = S(w_s,*)$ and $I(w_t,*) = S(*,w_t)$, and both are independent of i, we can rewrite the equation as follows:

$$w_s \cdot w_t = \frac{\sum_i I(w_s,i) \cdot I(w_t,i)}{S(w_s,*) \cdot S(*,w_t)} \tag{2.25}$$

Since $I(w,i)$ is an indicator function of whether the word w appeared in sentence i, it stands to reason that the product $I(w_s,i) \cdot I(w_t,i)$ is an indicator of whether both w_s and w_t appeared in i. Ergo, the numerator of Equation (2.25) is exactly the number of aligned sentences in which both w_s and w_t occurred: $S(w_s,w_t)$. Therefore:

$$w_s \cdot w_t = \frac{S(w_s,w_t)}{S(w_s,*) \cdot S(*,w_t)} = \frac{dice(w_s,w_t)}{2} \tag{2.26}$$

(Levy, Søgaard, and Goldberg 2017)

2.3.3.3 Multi-lingual SID-SGNS

Given their intuition on the importance of Sentence-ID and their observations about equivalences between the classic Dice aligner and more recent embedding approaches, they propose a new 'Generalized Dice' algorithm called 'SGNS with Sentence IDs (SID_SGNS)'. They construct a word/sentence-ID matrix in a similar fashion to Inverted Index, using a co-occurrence count metric $n_{w,s} \geq 0 \in \mathbb{Z}$ (eq. 2.27), instead of the binary co-occurrence metric previously shown in (eq. 2.19).

$$M_{(word,sentence\_id)} = \begin{array}{c} \\ w_{en_1} \\ w_{en_2} \\ \vdots \\ \hline w_{fr_1} \\ w_{fr_2} \\ \vdots \\ \hline w_{de_1} \\ w_{de_2} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & \cdots \end{array} \\ \left(\begin{array}{cccc} n_{en_1,1} & n_{en_1,2} & n_{en_1,3} & \cdots \\ n_{en_2,1} & n_{en_2,2} & n_{en_2,3} & \cdots \\ & \vdots & & \\ \hline n_{fr_1,1} & n_{fr_1,2} & n_{fr_1,3} & \cdots \\ n_{fr_2,1} & n_{fr_2,2} & n_{fr_2,3} & \cdots \\ & \vdots & & \\ \hline n_{de_1,1} & n_{de_1,2} & n_{de_1,3} & \cdots \\ n_{de_2,1} & n_{de_2,2} & n_{de_2,3} & \cdots \\ & \vdots & & \end{array} \right) \end{array} \tag{2.27}$$

They then apply an SGNS factorization to this matrix (Mikolov et al. 2013a). They are not entirely clear in their paper on how this was done, so this will be revisited later in the Pipeline Design chapter (4), where the findings from reverse engineering Levy et al's code will be described.

They achieved comparable performance to other sentence ID based methods when this approach was used on bilingual data. However, a significant benefit that this approach shares with Inverted Index, is that it is possible to train multi-lingually. Their best results were achieved by training simultaneously on 57 languages of the sentence-aligned Bible corpus (Christodouloupoulos and Steedman 2015), due to the enhanced signal derived from multi-lingual training data.

2.3.4 Systematic Baseline Comparison

The cross-lingual embedding literature presents numerous different approaches which have often been evaluated against diverse tasks without much comparability of benchmarks. There is also a wide variation in training signals used, which makes it difficult to detect whether variations in results are due to the data quality or the algorithms.

In their experiments, (Levy, Søgaard, and Goldberg 2017) have performed a systematic comparison of four previous baseline methods alongside their own new proposals.

2.3.4.1 Training Data

The parallel sentence-aligned training data was sourced from the Bible corpus (Christodouloupoulos and Steedman 2015) and Europarl (Koehn 2005), although their reported results are for the Bible corpus experiments only. Both corpora were decapitalized and tokenized using white spaces after splitting at punctuation

One of the baseline models (BilBOWA) also required monolingual corpora in each language as an input. It is not clear from Levy et al's paper what source they used for this, or indeed whether they added any monolingual inputs at all, since additional input data might bias their systematic comparison with the other algorithms. However, the original BilBOWA work (Gouws, Bengio, and Corrado 2014) used the freely available, pre-tokenized Wikipedia datasets from (Al-Rfou, Perozzi, and Skiena 2013).

2.3.4.2 Evaluation Tasks & Benchmarking Methods

Levy et al made their systematic comparison using word alignment and lexical induction evaluation tasks. They used 16 manually annotated word alignment benchmarks (Graca et al. 2008; Lambert et al. 2005; Mihalcea and Pedersen 2003; Holmqvist and Ahrenberg 2011; Çakmak, Acar, and Eryiğit 2012). They created a new lexical induction benchmark based on 16 bilingual dictionaries obtained from Wiktionary, an online crowd-sourced cross-lingual dictionary. They have made the Wiktionary benchmarks available for download and re-use, along with their codebase.

Note: A concurrent paper by (Upadhyay et al. 2016) also aimed to address this inconsistency of training and evaluation approaches. They proposed standardized evaluation using the intrinsic tasks of monolingual word similarity in English and cross-lingual dictionary induction, and extrinsic tasks of cross-lingual document classification and cross-lingual syntactic dependency parsing. Rather than creating their own evaluation set, they based their dictionary induction evaluation task on a gold standard from the Open Multilingual WordNet data released by (Bond and Foster 2013). This

may be a better-curated gold standard than Wiktionary, although it is not as cheap or self-maintaining as a crowd-sourced gold standard.

Chapter 3

Dataset Analysis

In this chapter, we present the findings of preliminary analyses on the data sets used in the following experiments.

3.1 Parallel Corpora

The Bible and Europarl parallel training corpora used in (Levy, Søgaard, and Goldberg 2017) are available for download with their source code [1]. This was the format in which they were used for these experiments, although there may be other formats available from their respective originators (Bible corpus: Christodouloupoulos and Steedman 2015; Europarl corpus: Koehn 2005).

3.1.1 Bible Corpus

The Bible corpus contains 24,785 parallel tokenized sentences in 57 languages (table 3.1), uniquely identified and aligned using sentence IDs.

af: Afrikaans	et: Estonian	kn: Kannada	pa: Paite (Chin)	te: Telugu
ar: Arabic	fa: Farsi	ko: Korean	pl: Polish	th: Thai
bg: Bulgarian	fi: Finnish	la: Latin	pt: Portuguese	tl: Tagalog
cb: Cebuano	fr: French	lt: Lithuanian	qe: Q'eqchi'	tr: Turkish
cf: Haitian Creole	he: Hebrew	mg: Malagasy	ro: Romanian	vi: Vietnamese
cs: Czech	hi: Hindi	mi: Maori	ru: Russian	ww: English
da: Danish	hr: Croatian	ml: Malayalam	sk: Slovak	xh: Xhosa
de: German	hu: Hungarian	mr: Marathi	sl: Slovene	zh: Chinese
el: Greek	id: Indonesian	my: Myanmar	so: Somali	zm: Zarma
en: English	is: Icelandic	ne: Nepali	sq: Albanian	
eo: Esperanto	it: Italian	nl: Dutch	sr: Serbian	
es: Spanish	ja: Japanese	no: Norwegian	sv: Swedish	

Table 3.1: Bible corpus languages (57 in total). Note: Two different English translations have been included (codes: en and ww)

The vocabulary and linguistic style used in the Bible is likely to be very characteristic of the domain across all languages. We can only quantitatively analyse the languages that are familiar to us, but to take some examples from the English, we see antiquated words (e.g. maketh, shewed, saith, unto, thy, brethren), domain-specific words (e.g. holy, prophets, almighty, apostles, alleluia), an absence of everyday words in the modern world (e.g. car, television, shop, university, big), and some polysemous words seem to be predominantly used in the biblical sense (e.g. **lamb** of God, God the **father**, **Lord** God). However, many words are unambiguous and used similarly in normal modern-day English (e.g. sea, sailor, city, fear, labour, blue, large).

In Figure (3.1), we analyse the number of words and range of vocabulary required to express the same meaning conveyed by the set of Bible sentences in a sample of the languages available in the corpus. This gives a 'big picture' quantitative insight into the challenges of the word alignment task due to morphological differences between languages. At the two extremes, we have Myanmar which has the smallest vocabulary but uses the most words to convey meaning,

[1] `http://www.bitbucket.org/omerlevy/xling_embeddings`

and Korean which has the largest vocabulary but needs relatively few words to convey the same meaning. This is likely to be caused by the isolating vs polysynthetic nature of the languages, which characterizes the number of morphemes per word, and is an indicator of how much information can be conveyed per word (i.e. these results suggest that Myanmar is isolating, whilst Korean is polysynthetic). Given this variation in the amount of meaning conveyed by words in different languages, it looks unlikely that we would be able to find clear 1:1 alignments between sentences for some language pairs, which will also make the task of lexical induction more challenging.

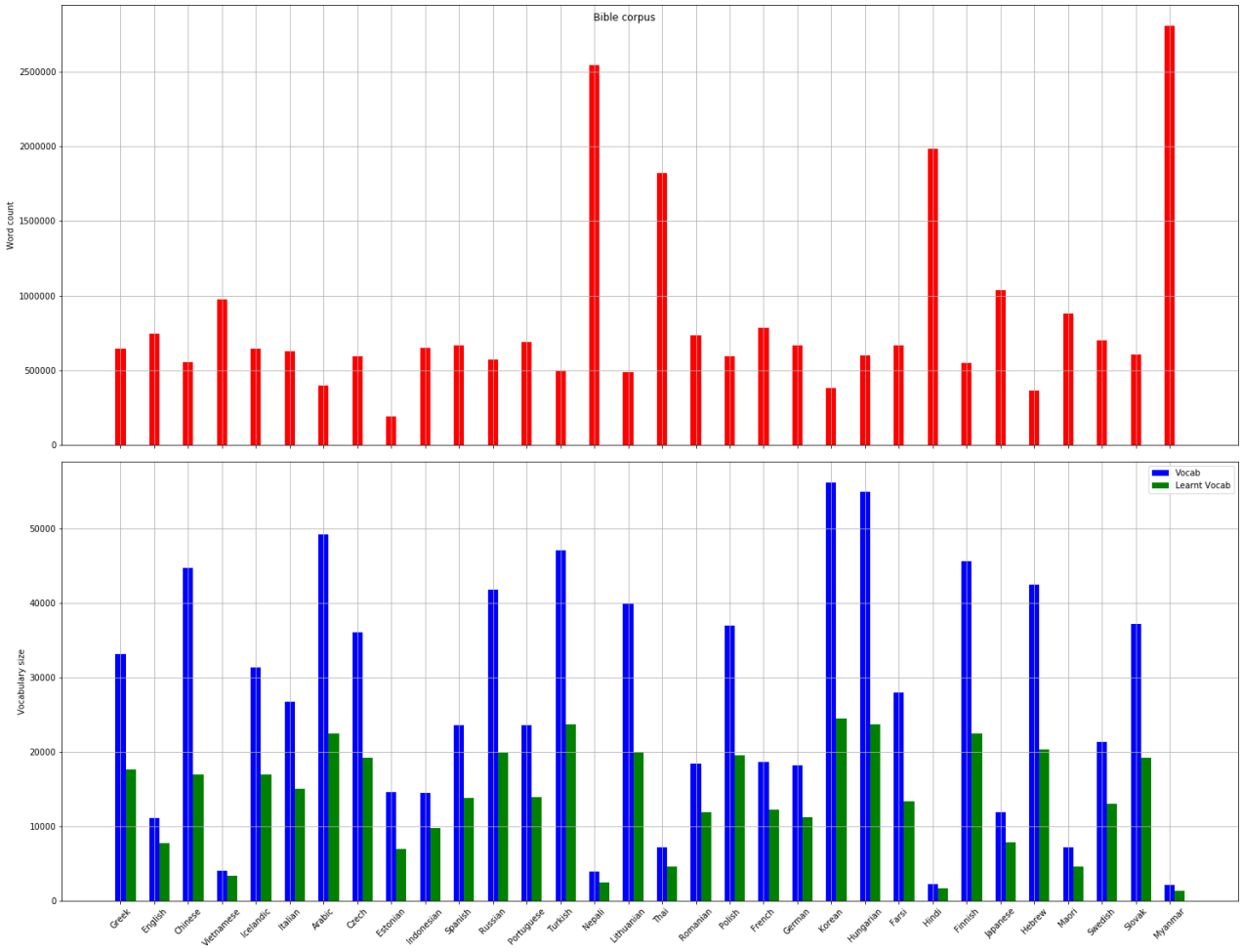


Figure 3.1: Analysis of the word counts correlated with the vocabulary sizes for a subset of languages in the Bible corpus.

Due to the Zipf long-tailed frequency distribution typical of language vocabularies, the baseline SID-SGNS algorithm thresholds word tokens at a minimum frequency of 2 during the training process, so that rare words that provide minimal training signal are discarded. In the following sections, this accounts for the difference between the 'vocabulary size' and 'learnt vocabulary size' statistics in figure (3.1).

3.1.2 Europarl Corpus

The Europarl corpus contains 178,375 parallel tokenized sentences in 21 languages (table 3.2), uniquely identified and aligned using sentence IDs.

bg: Bulgarian	en: English	hu: Hungarian	pl: Polish	sv: Swedish
cs: Czech	es: Spanish	it: Italian	pt: Portuguese	
da: Danish	et: Estonian	lt: Lithuanian	ro: Romanian	
de: German	fi: Finnish	lv: Latvian	sk: Slovak	
el: Greek	fr: French	nl: Dutch	sl: Slovene	

Table 3.2: Europarl corpus languages (21 in total).

Levy et al reported that they had trained their model on the Europarl corpus during their experiments, although they did not report their results, so we have no baseline for comparison. For that reason, this corpus has not been used further in these experiments. Nevertheless, it is interesting to contrast it with the Bible corpus, to gain some intuition on their relative strengths and weaknesses when applied to different problems.

Firstly, we note that the Europarl corpus is about 7x larger in terms of the number of sentences, but covers less than half the number of languages. The languages are also restricted to a set of fairly well-resourced European languages, whereas the Bible corpus contains some languages (e.g. Maori, Somali, Q'eqchi'), which we might suppose to be considered low-resource (Cieri et al. 2016).

The linguistic style is characteristic of the parliamentary/legal domain. In the English, we can see administrative documentation (e.g. "action taken on parliament 's resolutions : see minutes"), oral debate (e.g. "mr president , i rise on the issue of palestine .") and domain-specific words (e.g. commissioner, revenue, expenditure, minutes). However, there are a wide variety of current topics being debated (e.g. data protection, banking conduct, multi-cultural issues, climate change, agricultural policy), which we could expect to yield a broad coverage of vocabulary across these varied domains.

The objective of this project is to experiment with enriching the multi-lingual SID-SGNS model with the aim of bootstrapping embeddings using the strong multi-lingual signal from a small quantity of freely available parallel data, and expanding the vocabulary coverage with data from large monolingual corpora. For that objective, the Bible corpus is better suited than Europarl, since it has broader coverage of languages but a smaller vocabulary.

3.2 Evaluation Datasets

The Wiktionary evaluation dataset used for the lexical induction task in (Levy, Søgaard, and Goldberg 2017) was also downloaded with their source code. There are 8 bilingual dictionaries, which appear to be a complete dump from the Wiktionary site [2] at some time in the recent past. Due to the crowd-sourced nature of the dataset, the vocabulary of translation definitions in each dictionary is different.

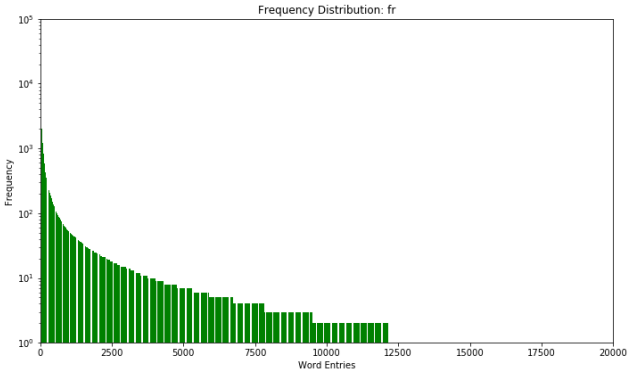
Source	Target	Volume	INV (%)		Transliterations (%)	
			Source	Target	Total	Capitalised
English	Arabic	7319	19.2	14.2	0.1	0.1
English	Spanish	35162	11.6	13.5	7.0	2.4
English	Finnish	48736	10.3	6.3	2.5	1.2
English	French	40106	10.6	11.7	14.2	2.9
English	Hebrew	6118	22.9	15.6	0.0	0.0
English	Hungarian	20504	14.8	10.4	1.7	0.8
English	Portuguese	34980	12.8	15.3	5.4	1.9
English	Turkish	8786	18.9	19.2	6.7	4.0

Table 3.3: Lexical induction evaluation set: Wiktionary bilingual dictionaries. The in-vocabulary (INV) percentages indicate the proportion of evaluation words that are included in the learnt vocabulary of Levy et al's cross-lingual embeddings trained on the Bible corpus.

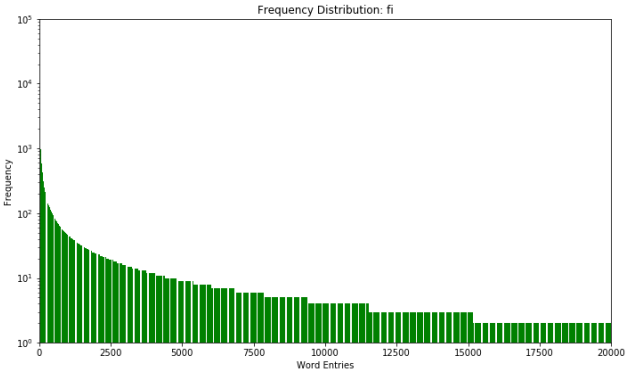
There is broad domain coverage, and a large proportion of modern-day words represented, which would be absent from the Bible corpus. There is considerable duplication of word pairs within the dictionaries, which probably arises from navigating inter-wiki links in the Wiktionary ontology between homonyms of words in the different languages. There are also different senses of polysemous words represented (e.g. in en-fr: 'bank'='banque' [financial institution]; 'bank'='rive' [river bank]).

The precision-at-1 (P@1) results presented in (Levy, Søgaard, and Goldberg 2017) ignore any words that are out-of-vocabulary (OOV) in the training corpus of either source or target language, i.e. the denominator of the P@1 percentage is the total of word pairs that are in-vocabulary (INV) for both languages. Figure (3.2) shows an analysis of the vocabulary frequency distribution in the French and Finnish translation of the Bible corpus. Only a small proportion of the Bible corpus words are in the evaluation set, and their frequencies vary widely from the highest (approx 10^3) down to 1. Equivalent plots for English and Spanish are shown in appendix section (§B.1.1 & B.1.2).

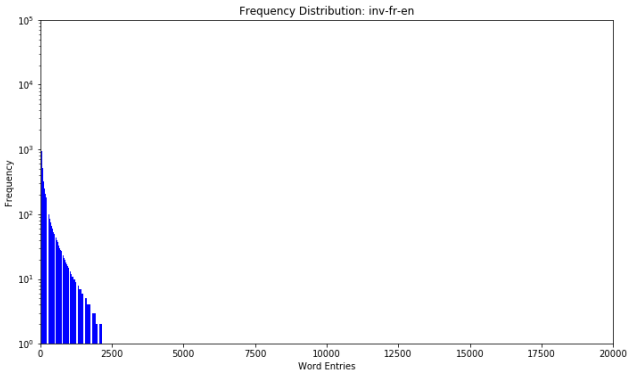
[2] https://en.wiktionary.org/wiki/Wiktionary:Main_Page



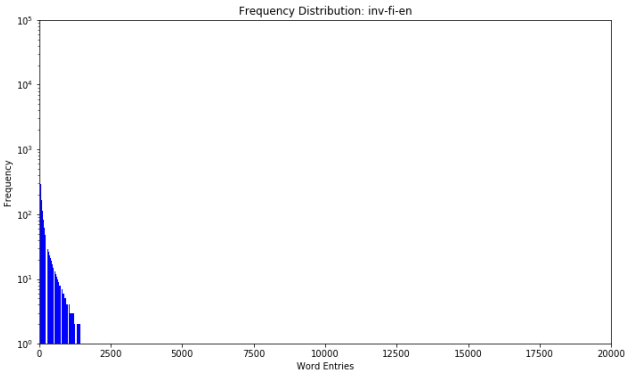
(a) French: $freq(w) : w \in C_{B_{fr}}$



(b) Finnish: $freq(w) : w \in C_{B_{fi}}$



(c) French/English: $freq(w) : w \in \{C_{B_{fr}} \cap E_{W_{fr-en}}\}$



(d) Finnish/English: $freq(w) : w \in \{C_{B_{fi}} \cap E_{W_{fi-en}}\}$

Figure 3.2: Plots showing (3.2a & 3.2b) the learnt vocabulary frequency distribution in the French and Finnish translations of the Bible corpus $C_{B_{l_s}}$. Plots (3.2c & 3.2d) show the subset of French and Finnish Bible corpus words that are considered INV i.e. Bible words for which the corresponding word pair in the Wiktionary evaluation set $E_{W_{l_s-l_t}}$ is included in **both** the source and target language translations of the Bible corpus.

In some language pairs there is a higher proportion of transliterations (see table 3.3); that is, translation definitions in which the source and target language words are the same. In some cases, these are proper nouns (usually detectable due to their capitalization) or acronyms (e.g. 'HTML'), whereas some are borrowed words (e.g. in French the borrowed word 'weekend' is used with the same meaning as in English). Although foreign language transliterations are heuristically easy for humans to learn, they are unlikely to be any easier for the SID-SGNS algorithm, since the source and target versions of the transliterated words will be prefixed with a language code making them effectively different words with different vectors in the embedding space. The algorithm does not look at the similarity of the characters within the word, only the similarity of sentence contexts in which they are found.

Chapter 4

Processing Pipeline Design

4.1 Recreating the Multilingual SID-SGNS Baseline

Prior to designing the enrichment pipeline, Levy et al's source code [1] was executed to recreate the baseline results of their Multilingual SID-SGNS experiment, trained on the Bible corpus. The training and evaluation pipeline code was adapted from the original shell scripts to iPython notebooks executed in an Anaconda 2 environment. A framework was created to execute 10 independent training runs, tabulate the results and perform significance testing as described in section (5.4), to establish whether the results of the recreated codebase were consistent with those reported in (Levy, Søgaard, and Goldberg 2017).

Code inspection revealed that the core algorithm `word2vecf` (Goldberg and Levy 2014) is closely derived from the `word2vec` implementation of (Mikolov et al. 2013a), which is written in C. It is not entirely clear from the explanation in (Levy, Søgaard, and Goldberg 2017) how the SGNS algorithm (figure 2.4) is applied to the word/Sentence-ID feature matrix (eq. 2.27). However, reverse engineering the `word2vecf` code in comparison with `word2vec` , it appears that the neural network architecture for the SID-SGNS embedding algorithm is approximately as shown in figure (4.1). The main differences are the construction of the training vector at the output layer from word/sentence-ID co-occurrence counts, and that the word input does not relate to a sequential list of words so there is no equivalent concept of a context window around the current word. Rather, context is defined by the Sentence-IDs in which the words appear. The Skip-gram and negative sampling features of the algorithm appear to be largely unaltered from the `word2vec` implementation.

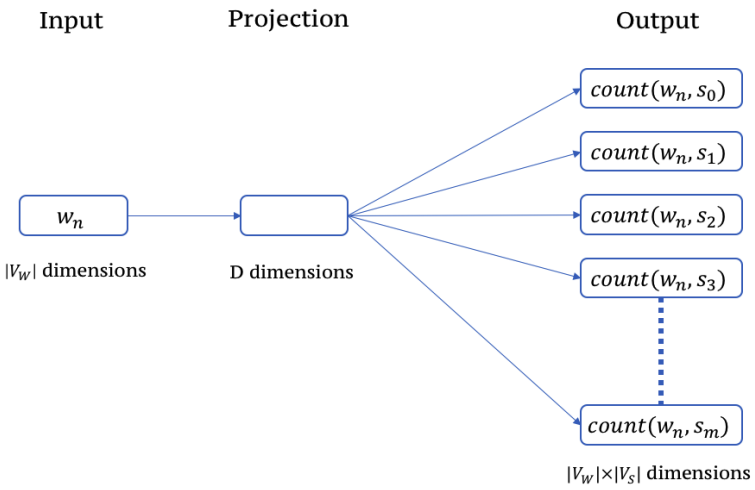


Figure 4.1: An approximation of the SID-SGNS Neural Network Architecture, as reverse engineered from the `word2vecf` code of (Goldberg and Levy 2014)

The training inputs to the model implementation are a `pairs` file, a `counts.words.vocab` file and a `counts.contexts.vocab` file. An intermediate file `counts` is used to aggregate the `counts.*.vocab` files:

- `pairs` - Maps words to the Sentence ID in which they appeared in the training corpus (Format: language_prefix,

[1] bitbucket.org/omerlevy/xling_embeddings

word, sentence_id)

- `counts` - How many times each word appears in each sentence (Format: count, language_prefix, word, sentence_id)
- `counts.words.vocab` - aggregates counts by word (Format: language_prefix, word, count in descending order)
- `counts.contexts.vocab` - aggregates counts by Sentence ID (Format: language_prefix, sentence_id, count in descending order)

The training outputs are:

- `words` - contains the word embedding vectors for all languages
- `contexts` - contains dimensionally reduced vectors describing the word features that apply to each Sentence ID (not used in the subsequent pipeline)

The `words` file is then split out into separate files for each language, and also reformatted as a serialized `numpy` array. The learnt vocabulary listing for each language is also extracted into a separate file.

The embedding dimension is set to 500, the number of iterations is 100, and the number of negative samples is 1. These baseline parameter settings have been re-used unmodified throughout the subsequent experiments.

4.2 Monolingual Corpus Extraction

The aim of this project is to extend the work of Levy et al, by enriching the baseline cross-lingual embedding with monolingual data inputs in each of the target languages. The goal of this enrichment is to add additional vocabulary into the embedding space and build cross-lingual relationships between words in a broader range of domains, especially more modern-day words than are found in the Bible.

To achieve this, we need a source of large quantities of freely available text spanning these vocabularies and domains. Ideally, the enrichment data would contain high-quality grammatically correct sentences, consistent with the quality of the parallel corpora used by Levy et al. Two social media sources were experimented with for this purpose: Wikipedia and Twitter.

4.2.1 Wikipedia Corpus

Data extraction code was developed in Python to invoke the Wikipedia API, download HTML page content and use scraping techniques to extract sentences from the body of the article. An alternative option to use monthly Wikipedia dump files was discarded, because the page content was in Wikimedia mark-up, which is less well-supported than HTML with code libraries for scraping content.

In order to extract good quality grammatical sentences, the content was filtered to include only the HTML paragraph (<p>) tags. The paragraphs were segmented into sentences using the Punkt tokenizer (not available for Hebrew and Arabic). The sentences were then split into word tokens, decapitalized and split at punctuation using the NLTK tokenizer, in a similar fashion to the Bible corpus provided by Levy et al. Any remaining wiki mark-up was removed using regular expression pattern matching. This process required thorough testing on example pages, to eliminate various anomalies in the content formatting, but nevertheless it is likely that some poor quality sentences remained in the sample. Each sentence was allocated a unique sentence-ID composed of the language code and a sequential number.

For the initial round of experiments, a set of 1,000 pages was randomly selected from each language's Wikipedia site. This resulted in a sample of approximately 15,000 sentences in each language. Due to the random selection approach, the subject matter and content of these pages would have varied substantially in each language sample.

4.2.2 Twitter Corpus

Data extraction code was developed in Python to invoke the Twitter API and download tweets. The API offers two possible approaches for this: a streaming approach where tweets are captured in real-time based on a filter criterion, and a query option where tweets are retrieved from recent history based on a keyword search.

Some additional extract management code was needed to operate within Twitter's usage constraints. The query API has a rate limit mechanism, which allows 180 API calls every 15 minutes. The streaming API allows the connection to be left open indefinitely, but in order to change the filter criteria the connection must be dropped and re-connected, and excessively frequent re-connections are penalized.

The streaming approach was used initially simply to retrieve tweets to use in initial trials. This was subsequently replaced with the query approach, in order to target tweets containing a particular subset of vocabulary (we will expand on this in section 4.4).

The tweet content was extracted and tokenized using the NLTK tokenizer, without any additional data cleansing. As is typical of Twitter data, the tweets contained very noisy sentence examples, with many non-words, ungrammatical constructions, hyperlinks, etc. Each sentence was allocated a unique sentence-ID composed of the language code and a sequential number.

4.3 Multilingual SID-SGNS Model with Monolingual Enrichment

Having prepared monolingual corpora in each language, pre-processed in a similar way to the multilingual Bible corpus, the baseline cross-lingual embedding model was retrained with the added enrichment data. This was achieved by adding the co-occurrence counts for the words in the enrichment corpora into the word/sentence-ID matrix, as shown in eq. (4.1).

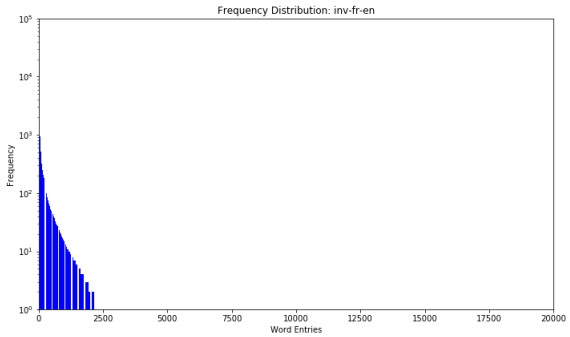
Given a multi-lingual corpus C_B (such as the Bible corpora), which contains a set of aligned sentences $s_j \in S_B \in C_B$ in languages $l \in L_B$, comprised of words w_i that are each a member of only one language vocabulary V_{B_L}; and further given monolingual corpora C_l (such as the Wikipedia corpora) in languages $l \in L_W \ni [en, fr, fi, es]$, which contain a set of non-aligned documents comprised of sentences $s_k \in S_{W_l} \in C_{W_l}$ and words w_h belonging to the corpus vocabulary V_{W_l}; we define a word/sentence-ID matrix $M_{(W,S)}$ with elements equal to the co-occurrence count metric $n_{w,s} \geq 0 \in \mathbb{Z}$, as shown:

$$
M_{(W,S)} = \begin{array}{c|cccccccccc}
 & s_j \in C_B & \cdots & s_k \in C_{W_{en}} & \cdots & s_k \in C_{W_{fr}} & \cdots & s_k \in C_{W_{fi}} & \cdots & s_k \in C_{W_{es}} & \cdots \\
w_i \in V_{B_{en}} & n_{i,j} & \cdots & n_{i,k} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_i \in V_{B_{fr}} & n_{i,j} & \cdots & 0 & \cdots & n_{i,k} & \cdots & 0 & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_i \in V_{B_{fi}} & n_{i,j} & \cdots & 0 & \cdots & 0 & \cdots & n_{i,k} & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_i \in V_{B_{es}} & n_{i,j} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & n_{i,k} & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_h \in V_{W_{en}} \notin V_{B_{en}} & 0 & \cdots & n_{h,k} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_h \in V_{W_{fr}} \notin V_{B_{fr}} & 0 & \cdots & 0 & \cdots & n_{h,k} & \cdots & 0 & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_h \in V_{W_{fi}} \notin V_{B_{fi}} & 0 & \cdots & 0 & \cdots & 0 & \cdots & n_{h,k} & \cdots & 0 & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\
\hline
w_h \in V_{W_{es}} \notin V_{B_{es}} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & n_{h,k} & \cdots \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots &
\end{array}
\tag{4.1}
$$

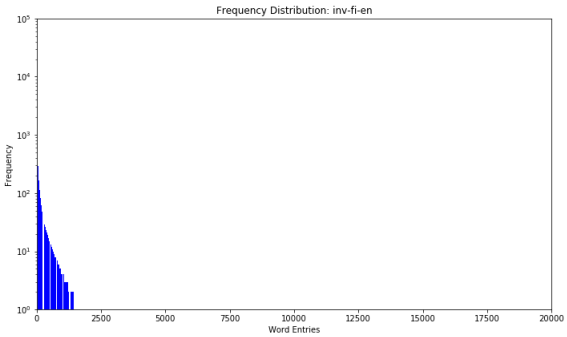
The SID-SGNS multilingual embedding model was otherwise unaltered, maintaining the same SGNS parameters as the baseline.

Effectively, this would boost the occurrence of some of the INV words from the Bible corpus, although by unequal amounts for each language since the enrichment corpora do not contain comparable content. It would also add new vocabulary, although the additional vocabulary would be different for each language.

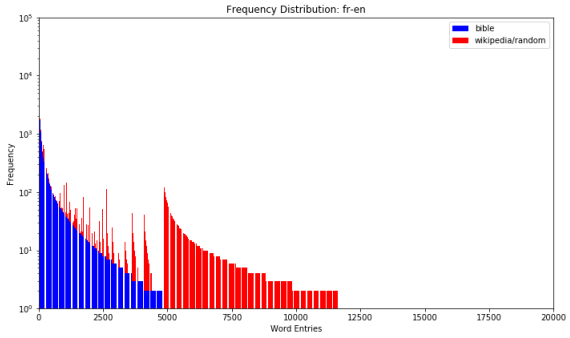
An initial attempt of the enrichment approach was tested using a randomly sampled set of Wikipedia sentences. A second trial used a set of Twitter sentences extracted by using the query API to search for tweets containing the OOV words (i.e. not in the Bible corpus) from the evaluation set. The effects of the enrichment on the vocabulary frequency distributions can be seen in figure (4.2 a-f). The proportion of evaluation words found in the Bible corpus (blue) increases relative to the baseline, because some words that were previously INV in the source language but OOV in the target language are now in scope. The Twitter distribution appears to have the best coverage of the target evaluation vocabulary, but we remained sceptical whether it would perform well, due to the high level of noise in Twitter data. Equivalent plots for Spanish, and the reverse directions of translation, are shown in appendix section (§B.1.3 & B.1.4).



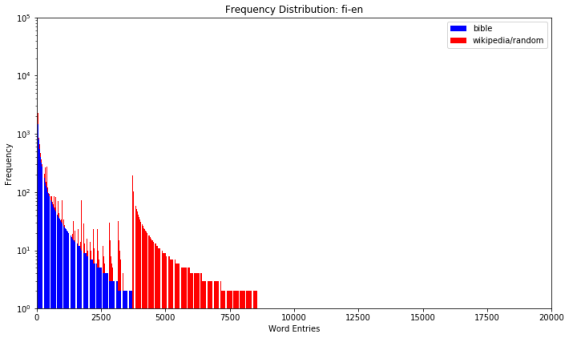
(a) Bible Baseline
$freq(w) : w \in \{C_{B_{fr}} \cap E_{W_{fr-en}}\}$



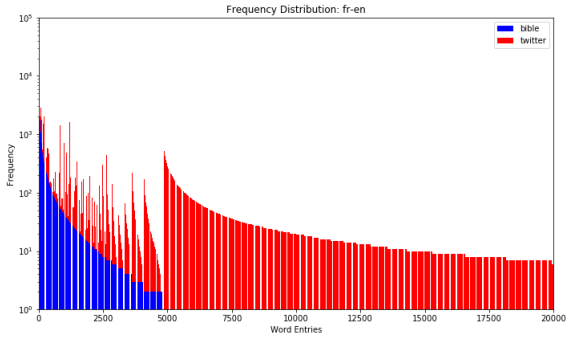
(b) Bible Baseline
$freq(w) : w \in \{C_{B_{fi}} \cap E_{W_{fi-en}}\}$



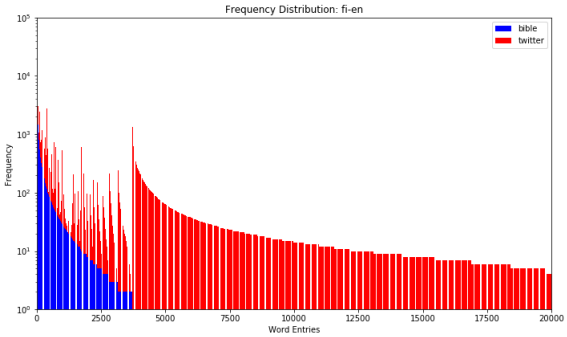
(c) Randomized Wikipedia Enrichment
$freq(w) : w \in \{\{C_{B_{fr}} \cup Randomized(C_{W_{fr}})\} \cap E_{W_{fr-en}}\}$



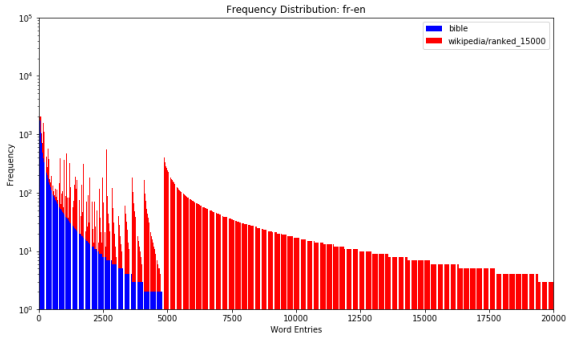
(d) Randomized Wikipedia Enrichment
$freq(w) : w \in \{\{C_{B_{fi}} \cup Randomized(C_{W_{fi}})\} \cap E_{W_{fi-en}}\}$



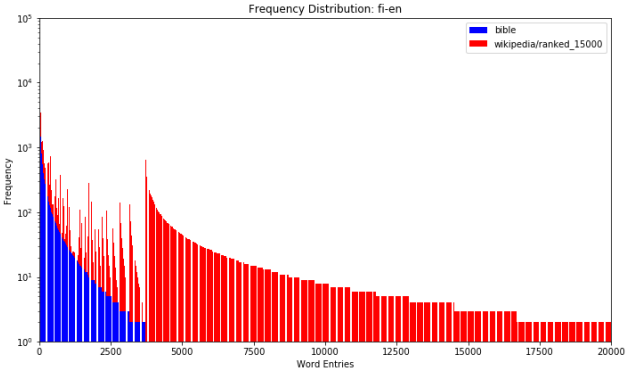
(e) Twitter Enrichment
$freq(w) : w \in \{\{C_{B_{fr}} \cup C_{T_{fr}}\} \cap E_{W_{fr-en}}\}$



(f) Twitter Enrichment
$freq(w) : w \in \{\{C_{B_{fi}} \cup C_{T_{fi}}\} \cap E_{W_{fi-en}}\}$



(g) Ranked Wikipedia Enrichment
$freq(w) : w \in \{\{C_{B_{fr}} \cup Ranked(C_{W_{fr}})\} \cap E_{W_{fr-en}}\}$



(h) Ranked Wikipedia Enrichment
$freq(w) : w \in \{\{C_{B_{fi}} \cup Ranked(C_{W_{fi}})\} \cap E_{W_{fi-en}}\}$

Figure 4.2: Plots showing how enrichment of the baseline Bible corpus $C_{B_{l_1}}$ (blue) with sentences from Twitter $C_{T_{l_1}}$ and Wikipedia $C_{W_{l_1}}$ sources added new vocabulary (red) to the frequency distributions. The plots are filtered to show only the vocabulary that intersects with the target words in the Wiktionary evaluation set $E_{W_{l_1-l_2}}$. The left column shows distributions for French/English, while the right column shows Finnish/English.

4.4 Enrichment Sentence Ranking Method

A third trial was attempted, using a set of Wikipedia sentences extracted by using the query API to search for pages containing the OOV words (i.e. not in the Bible corpus) from the evaluation set. However, this resulted in a corpus of approx 19 million sentences (9GB of uncompressed text files). Since the SID-SGNS code trains the model by loading the word/sentence-ID feature matrix into main memory, it would not have been possible on the hardware available to train using such a large corpus combined with the baseline Bible corpus.

Since the Wikipedia sentences appeared to be generally of better quality than the tweets, a ranking method was devised to attempt to filter the size of the Wikipedia corpus down to a more manageable size, by discarding sentences that would provide the least amount of signal from which to learn the target vocabulary in the evaluation set. The goal was to select a subset of whole sentences, with an overall effect on the vocabulary distribution of down-sampling the most and least frequent words, and any words that were not present in the evaluation set. Stop words occur extremely frequently, and are not discriminative in terms of their co-occurrence statistics, so they are not helpful for learning semantic relationships. Whereas words with very low frequencies do not provide enough co-occurrence signal.

The sum of the TF-IDF (Term Frequency - Inverse Document Frequency) scores for the set of words in each sentence was calculated, as follows.

First calculate:

- The Term Frequency $\text{TF}_{w,s}$, the number of occurrences of word w in sentence s.
- The Document Frequency DF_w, the number of sentences in the corpus containing the word w.

Now compute the Inverse Document Frequency IDF_w:

$$\text{IDF}_w = \log \frac{N}{\text{DF}_w}, \text{ where } N \text{ is the total number of sentences in the corpus.} \tag{4.2}$$

Then multiply by the Term Frequency $\text{TF}_{w,s}$ to compute the TF-IDF score:

$$\text{TF-IDF}_{w,s} = \text{TF}_{w,s} \times \text{IDF}_w \tag{4.3}$$

To perform this calculation on such a large corpus requires persistent memory in a database. MongoDB was selected due to its simple key/value document-based model, which works very naturally with Python dictionaries (hashmaps). The collections shown in figure (4.3) were created to model the data. The Wikipedia corpus was loaded from raw files into the `sentences` collection. To ensure no duplication of data, unique indexes were applied to the `sentence_id` and `sentence_text` values (the latter using an MD5 hashing algorithm to compress the full sentence text into a unique string). Derived collections analogous to the `pairs` and `counts` files were created, and a `words` collection analogous to the `counts.words.vocab` file. The evaluation dataset was read into the `eval_set` collection, to enable look-ups so that TF-IDF scores were only allocated to words that were present in the evaluation set.

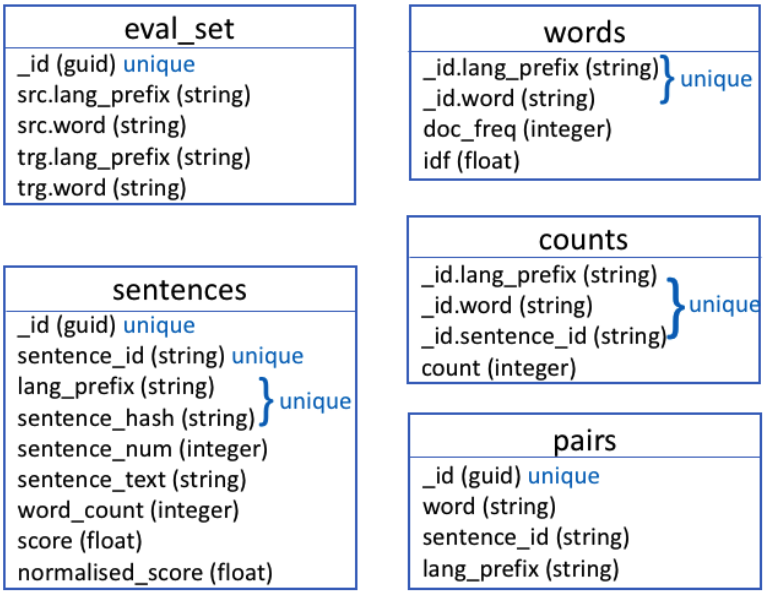


Figure 4.3: MongoDB database collections used to rank sentences in enrichment corpora

The computation was then performed as follows:

- $\text{TF}_{w,s}$ was calculated by aggregating the `pairs` collection on `lang_prefix`, `word` and `sentence_id` and inserting the aggregated count into the `counts` collection.
- DF_w was calculated by aggregating the `pairs` collection on `lang_prefix` and `word` and updating the sentence count into the `word` item.
- N was calculated by performing a count on the `sentences` collection.
- A word lookup was performed to the `eval_set` collection, and if there was a match, then the IDF_w was calculated using equation (4.2) and updated onto the `word` item.
- The $\text{TF-IDF}_{w,s}$ was calculated for all items in the `counts` collection for which the corresponding `word` item was populated with the IDF_w.
- For each item in the `sentences` collection, the $\text{TF-IDF}_{w,s}$ was summed over corresponding entries in the `counts` collection, and updated into the `score` value.
- An alternative `normalized_score` was calculated by dividing the `score` by the word count in the sentence.
- The sentences were ranked by the chosen scoring metric, and the top 15,000 sentences for each language were selected as inputs to the enrichment process. This created a corpus of roughly equivalent size to the randomized approach, to make the results comparable.

The un-normalized `score` metric was used in all experiments reported here. Due to lack of time, it was not possible to execute trials with both scoring methods. The resulting vocabulary frequency distributions are shown in figure (4.2 g-h).

TBD - Overall pipeline diagram

Chapter 5

Evaluation Methods

5.1 Evaluation Languages

To reduce the scope of the experiments to a more manageable size, given the time and resources available, a subset of languages was selected for experimentation. The rationale for the selection was to use languages with which the author has some familiarity (English, French, Spanish), since this makes qualitative evaluation of results more intuitive. However, given the analysis of isolating vs polysynthetic languages carried out in section (3.1), a more polysynthetic language was also selected for trial, to investigate what additional challenges this may present. Finnish seemed a good choice, in that it is well-resourced with Wikipedia content that can be used for enrichment, and we also have a Wiktionary evaluation data set and baseline results available from Levy et al.

5.2 Evaluation Tasks

Evaluation was focused on the lexical induction task, re-using the benchmarks from (Levy, Søgaard, and Goldberg 2017). This is the more intrinsic of the two baseline evaluation tasks provided in their work. Compared to the word alignment problem, it isolates the evaluation task down to the cross-lingual word similarity problem alone, without any of the idiomatic and morphological complications introduced by the word alignment task.

5.3 Evaluation Metrics

The evaluation metric used by (Levy, Søgaard, and Goldberg 2017) was 'Precision@1' (P@1)i.e. the proportion of words for which the model output word vectors whose cosine similarity measure found a single nearest neighbour word that exactly matched the translation word given in the Wiktionary data set. The tests were initially recreated using the same metric so that benchmarks could be compared to ensure an accurate recreation of the experiment.

However, for practical purposes, the P@1 metric is perhaps overly stringent and not necessarily a good indicator of the usefulness of the model, for the following reasons.

The evaluation dataset is derived by navigating the Wiktionary inter-language ontology, so that it produces an evaluation entry for each different sense of polysemous words in the source language. Each evaluation entry challenges the model to find the target language translation for the particular intended sense, without giving it any source language sentence context to help it identify which sense is meant. The P@1 metric, however, will only ever look at the single nearest neighbour target language word vector defined by the embedding model, without any notion of the intended sense in the source language. It will always come up with the same result. To take into account polysemous words, we need to evaluate against a wider neighbourhood of translation words.

In the English-Finnish translations in particular (but also to a lesser extent between all language pairs), we observe that the 10 nearest word vectors tend to include various different inflections of the same word stem. In fact, it would often be impossible in a unigram lexical induction task to get the correct Finnish inflection, since English often requires many words to express the same meaning as a single inflected Finnish word. For many practical tasks in which we might use this kind of rough lexical induction model, we would not require the model to get the inflection exactly right.

Whereas, using a 'Precision@10' (P@10) allows us to apportion some value to the model if it manages to position the intended sense/inflection of the word somewhere within a wider range of the source word i.e. within the 10 nearest neighbour word vectors.

5.4 Significance Testing

When designing and evaluating the performance of the algorithm, hypothesis testing was performed wherever possible on quantitative results, to evaluate the statistical significance of any change in the result versus a prior baseline. A two-tailed hypothesis test (Student's t-test) was performed using `scipy.stats.ttest_1samp()`, which test for the null hypothesis that the expected value (mean) of a sample of independent observations is equal to the given population mean. In this case, the population is an infinite set of all the possible test statistics that could be collected for these benchmarks, and we are trying to establish whether the sample of statistics that we have observed is likely to be representative of the population as a whole.

For example, when comparing the initial Bible corpus baseline Precision@1 results for INV words obtained in our recreation of their experiment against the results published in (Levy, Søgaard, and Goldberg 2017) we take null and alternative hypotheses:

$$H_0 : \mu_{pop} \approx \mu_{Levy} \tag{5.1}$$

$$H_a : \mu_{pop} \neq \mu_{Levy} \tag{5.2}$$

i.e.
H_0: The recreation of the experiment is consistent with Levy et al's, and the published results of Levy et al's experiments are a good approximation of the true population mean.
H_a: Either the published results of Levy et al's experiments are not a good approximation of the true population mean, or the recreation of the experiment differs in some significant way.

$$\text{p-value} = P(\text{observed experimental results or a more extreme outcome} \,|H_0 \text{ true}) \tag{5.3}$$

In a two-tailed test at the 5% significance level, a p-value of 0.025 or less indicates that we should consider rejecting the null hypothesis H_0 and look at whether an alternative hypothesis H_a may be a more accurate approximation of the ground truth.

5.5 Qualitative Evaluation

In addition to quantitative evaluation, the top 10 translations for a selection of words were inspected qualitatively, to investigate the linguistic characteristics of some of the inaccuracies and the particular factors that might cause the model to perform less well on certain words or language pairs.

The frequency of the words within the training corpus seems a relevant criterion for selecting which words to investigate, based on an intuition that there may be a minimum threshold frequency required to deduce generalised translations from the data. Two samples of words were chosen: a 'biblical' sample that occur in the Bible corpus and a 'modern' sample that do not, where we are reliant on the enrichment process to populate the embeddings in the model. The frequencies of these words in the corpora are shown in table (5.1).

Note that the frequencies of the Finnish words are considerably lower than in the other languages. This is likely to be a result of the polysynthetic and highly inflected nature of the language; fewer words are generally required to express the same meaning, and instances of a particular word stem will more frequently arise with inflectional prefixes and suffixes than in the other languages.

w_{en}	C_B	C_{W_r}	C_T	C_{W_k}	w_{fr}	C_B	C_{W_r}	C_T	C_{W_k}
man	2351	2430	5186	3301	homme	1559	1630	1902	2151
woman	313	370	1377	674	femme	669	726	1027	1128
lamb	78	78	147	102	agneau	77	77	80	85
shepherd	37	37	90	70	berger	24	24	30	50
sea	345	437	959	613	mer	340	377	434	632
car	0	40		215	voiture	0	13		74
phone	0	16		74	téléphone	0	2		47
film	0	274		814	film	0	181		323
newspaper	0	32		41	journal	0	31		156
national	0	378		856	nationale	0	106		339
w_{es}	C_B	C_{W_r}	C_T	C_{W_k}	w_{fi}	C_B	C_{W_r}	C_T	C_{W_k}
hombre	1405	1454	1769	1857	mies	456	482	456	636
mujer	592	655	903	906	nainen	70	81	788	161
cordero	79	79	91	98	karitsa	26	26	175	26
pastor	37	39	65	58	paimen	15	17	27	24
mar	316	384	445	692	meri	38	44	52	57
coche	0	10		43	auto	0	14		23
teléfono	0	14		21	puhelin	0	3		19
película	0	85		181	elokuva	0	54		99
periódico	0	23		58	sanomalehti	0	6		10
nacional	0	202		923	kansallinen	0	7		22

Table 5.1: Word frequencies of the set of 'biblical' and 'modern' words used for qualitative evaluation in the Bible corpus C_B and the Wikipedia enriched corpora C_{W_r} (random) and C_{W_k} (ranked).

Chapter 6

Results

6.1 Quantitative Benchmarks

6.1.1 In-Vocabulary (INV) Evaluation Words

Corpus	Metric	en-es	en-fi	en-fr	es-en	fi-en	fr-en
C_B Levy Baseline	μ(P@1)	0.351	0.159	0.330	0.387	0.258	0.389
C_B Recreated	μ(P@1)	0.352	0.157	0.332	0.385	0.258	0.391
	p-value vs Levy BL	0.057	0.040	0.013	0.057	0.129	0.062
	μ(P@10)	0.555	0.401	0.552	0.569	0.426	0.570

Table 6.1: Summarised results for Wiktionary evaluation for in-vocabulary (INV) words only and language pairs {en-es, en-fi, en-fr}. P@1 and P@10 scores are means over a sample of 10 independent training executions.

The recreation of the Levy baseline results were generally slightly better than the results reported in the paper, although the en-fr test had a p-value in the statistically significant range (¡0.025), suggesting there may have been something about the set-up or the data that was inconsistent with Levy et al's experiments.

The Precision@10 results were substantially better than the Precision@1 results, which shows that the model was effective in locating word translation pairs cross-lingually within a neighbourhood.

The highest precision scores were achieved on the fr-en evaluation set, whilst the lowest were on the en-fi set. This could be explained by the the degree of inflection of the respective languages. Finnish is highly inflected, whilst English is much less so. On a unigram lexical induction task without any sentence context around the word, there will inevitably be information missing in English when translating into Finnish, such that it would often be impossible to predict the required inflection. On the Precision@1 test metric, there is no credit given for getting the word stem right, but the wrong inflection. Whereas on the Precision@10 metric the en-fi test fares better, since the required inflection is more likely to be found in the top 10 nearest neighbours.

6.1.2 Out-of-Vocabulary (OOV) Evaluation Words

Corpus	Metric	en-es	en-fi	en-fr	es-en	fi-en	fr-en
C_B Baseline	μ(P@1)	0.0416	0.0097	0.0361	0.0455	0.0159	0.0425
	μ(P@10)	0.0656	0.0247	0.0601	0.0675	0.0267	0.0622
$C_B \cup Randomized(C_W)$	μ(P@1)	0.0412	0.0097	0.0360	0.0458	0.0159	0.0418
	p-value vs C_B	0.0034	0.6583	0.6238	0.0046	0.4765	0.0000
	μ(P@10)	0.0657	0.0246	0.0598	0.0672	0.0255	0.0615
	p-value vs C_B	0.7188	0.1479	0.0007	0.0025	0.0000	0.0000
$C_B \cup Search(C_T)$	P@1	0.0417	0.0096	0.0365	0.0457	0.0156	0.0415
$C_B \cup Ranked(C_W)$	μ(P@1)	0.0412	0.0097	0.0360	0.0458	0.0159	0.0418
	p-value vs C_B	0.0034	0.6583	0.6238	0.0046	0.4765	0.0000
	μ(P@10)	0.0655	0.0247	0.0596	0.0671	0.0258	0.0614
	p-value vs C_B	0.0003	0.6488	0.0001	0.0002	0.0000	0.0000

Table 6.2: Summarised results for Wiktionary evaluation including out-of-vocabulary (OOV) words for language pairs {en-es, en-fi, en-fr}. P@1 and P@10 scores are means over a sample of 10 independent training executions (apart from the Twitter trial $C_B \cup Search(C_T)$, which was executed only once to get an indicative result and without incurring long execution time).

The P@1 and P@10 results for OOV evaluation words showed that the proposed enrichment process did not achieve any significant improvement on the baseline scores. Furthermore, in all cases where the p-value demonstrated a statistically significant difference from the baseline, the result was in fact slightly worse than the baseline.

6.2 Qualitative Results Analysis

6.2.1 In-Vocabulary (INV) Evaluation Words

Table (6.3) shows qualitative results from the baseline Multilingual SID-SGNS model for a hand-picked sample INV word, which occurred in the Bible corpus with frequencies for each of the four languages in the range 15-37. All translation tests apart from en-fi would have scored successfully on a P@1 metric, whereas the en-fi translation returned the correct word stem, but with the wrong inflection. The correct inflection was second in the list, so en-fi would have scored successfully on the P@10 metric.

In most of the translations, the top 10 nearest neighbours are populated with words that are topically related to the subject of *shepherd*, such as *flock, sheep, fold, pasture, graze*. However, the en-fi top 10 are dominated by multiple different inflections of only two topically related word stems, *paimen [shepherd]* and *lampaat [sheep]*. There are a few results in the top 10 that are somewhat unexpected e.g. *buttocks* and *bishop*. The latter is related to a specific biblical domain-related sense of the word *shepherd*.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	berger [shepherd]	en	shepherd	0.853	en	shepherd	fr	berger [shepherd]	0.853
			flock	0.621				pasteur [pastor]	0.809
			shepherds	0.600				pasteurs [pastors]	0.654
			sheep	0.570				troupeau [herd]	0.652
			pastors	0.531				bergers [shepherds]	0.614
			fold	0.529				paître [graze]	0.581
			flocks	0.523				brebis [ewe]	0.556
			wolf	0.509				dispersent [disperse]	0.534
			feed	0.498				loup [wolf]	0.533
			catcheth	0.496				pâturage [pasture]	0.533
es	pastor [shepherd]	en	shepherd	0.948	en	shepherd	es	pastor [shepherd]	0.948
			flock	0.663				pastores [shepherds]	0.666
			shepherds	0.629				dispersadas [scattered]	0.625
			pastors	0.585				rebaño [flock]	0.620
			sheep	0.578				ovejas [sheep]	0.583
			flocks	0.549				apacentará [will graze]	0.564
			wolf	0.529				pastizal [pasture]	0.525
			pasture	0.511				lobo [wolf]	0.519
			feed	0.502				obispo [bishop]	0.508
			catcheth	0.483				perenne [perennial]	0.506
fi	paimen [shepherd]	en	shepherd	0.742	en	shepherd	fi	paimenen [shepherd]	0.785
			fold	0.549				paimen [shepherd]	0.742
			catcheth	0.544				paimenta [shepherd]	0.717
			wolf	0.538				lampaat [sheep]	0.637
			flock	0.527				paimenia [shepherds]	0.600
			sheep	0.515				paimenensa [shepherd]	0.593
			flocks	0.492				lammasten [sheep]	0.588
			brown	0.470				laumansa [his flock]	0.576
			shepherds	0.462				lampaistani [sheep]	0.569
			buttocks	0.457				paimeneni [shepherd]	0.568

Table 6.3: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the word *shepherd*. The English translations in square brackets are obtained from Google Translate[1].

By contrast, table (6.4) shows the results from the Multilingual SID-SGNS with Ranked Wikipedia Enrichment model for the same word *shepherd*. The top 10 lists of words and the similarity scores have changed, which demonstrates that the enrichment is having some effect on the generated embeddings. Qualitatively, the results seem about as good as the baseline.

[1] https://translate.google.co.uk/

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	berger [shepherd]	en	shepherd	0.858	en	shepherd	fr	berger [shepherd]	0.858
			flock	0.611				pasteur [pastor]	0.818
			sheep	0.601				pasteurs [pastors]	0.660
			fold	0.571				troupeau [herd]	0.651
			shepherds	0.570				bergers [shepherds]	0.605
			flocks	0.532				brebis [ewe]	0.586
			pastors	0.523				paître [graze]	0.568
			catcheth	0.520				pâturage [pasture]	0.555
			wolf	0.501				boucherie [butchery]	0.529
			feed	0.488				dispersées [dispersed]	0.525
es	pastor [shepherd]	en	shepherd	0.952	en	shepherd	es	pastor [shepherd]	0.952
			flock	0.645				pastores [shepherds]	0.699
			pastors	0.631				dispersadas [scattered]	0.626
			shepherds	0.616				rebaño [flock]	0.624
			sheep	0.613				ovejas [sheep]	0.600
			fold	0.532				descarriadas [gone astray]	0.552
			flocks	0.528				apacentará [will graze]	0.551
			wolf	0.525				rebaños [flocks]	0.550
			pasture	0.496				pastizal [pasture]	0.548
			feed	0.485				apacentaré [will graze]	0.544
fi	paimen [shepherd]	en	shepherd	0.778	en	shepherd	fi	paimenen [shepherd]	0.780
			fold	0.599				paimen [shepherd]	0.778
			catcheth	0.595				paimenta [shepherd]	0.732
			wolf	0.573				lampaat [sheep]	0.644
			flock	0.527				paimenensa [shepherd]	0.614
			brown	0.515				laumansa [his flock]	0.602
			sheep	0.496				kaitsevat [shall feed]	0.586
			treadeth	0.494				lammasten [sheep]	0.584
			pastors	0.488				paimenia [shepherds]	0.580
			leopard	0.479				paimeneni [shepherd]	0.570

Table 6.4: Multilingual SID-SGNS with Ranked Wikipedia Enrichment - translations of the word *shepherd*.

6.2.2 Out-of-Vocabulary (OOV) Evaluation Words

In the Supplementary Results appendix, table (B.9) shows the results from the baseline Multilingual SID-SGNS model for a hand-picked OOV word. This demonstrates what happens in an intentionally negative test scenario. All results have a zero similarity score and the same list of words is returned in every test.

Whereas table (6.5) shows the results obtained from the Multilingual SID-SGNS with Ranked Wikipedia Enrichment model. The similarity scores are much lower than for the INV Bible words. Qualitatively, there appears to be almost no relevance of the predicted translations words to the source word. In some cases, the translation words returned are mis-spellings, proper nouns, numeric values, non-words, or even in the wrong language (although the enrichment process guarantees that they were sourced from the Wikipedia site of the appropriate target language).

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	film [film]	en	must	0.196	en	film	fr	compositeur [composer]	0.182
			pallars	0.174				non-p [-]	0.168
			middlesbrough	0.174				égoïsme [selfishness]	0.166
			wainwright	0.161				conversos [-]	0.163
			wei	0.159				conduise [lead]	0.161
			nettle	0.158				prôné [advocated]	0.159
			cnblue	0.157				qu'une [only one]	0.158
			fascination	0.155				ccmm [-]	0.156
			lesion	0.155				conditionne [conditioned]	0.155
			skin	0.154				grand-duché [Grand Duchy]	0.155
es	película [film]	en	parodic	0.173	en	film	es	corredera [slide]	0.181
			guénon	0.171				cedidos [ceded]	0.170
			four-year	0.169				aristide [-]	0.168
			france	0.166				milans [-]	0.160
			clockwise	0.165				encantado [charmed]	0.159
			epitome	0.164				mexiquense [-]	0.158
			conquered	0.163				porlo [-]	0.157
			first-time	0.163				elegían [they chose]	0.156
			exacerbated	0.157				leitmotivs [-]	0.154
			mobs	0.155				capacitado [capable]	0.154
fi	elokuva [film]	en	auf	0.177	en	film	fi	anteeksianto [forgiveness]	0.196
			s11	0.167				uhaksi [a threat]	0.166
			$2\frac{1}{2}$	0.167				1942 [-]	0.165
			victoria	0.166				kuulan [ball]	0.163
			18.2-79	0.165				saavutat [you reach]	0.162
			niobium	0.163				dôme [-]	0.160
			valves	0.161				cliff [-]	0.159
			exceeds	0.158				maasälpää [feldspar]	0.157
			lou	0.154				asettamista [placing]	0.155
			neighbouring	0.149				vastuuton [irresponsible]	0.154

Table 6.5: Multilingual SID-SGNS with Ranked Wikipedia Enrichment - translations of the OOV word *film*.

Results for the Twitter search enrichment on the OOV word *car* are shown in the appendix (table B.14). These appear to be even more noisy than the Wikipedia results. The lists include Twitter hashtags and handles, URLs, numerical values, words in the wrong language, non-words, etc. However, there are a couple of results that show some qualitative relevance to the source word: the 4th translation for *voiture* in French is *turboshaft*, and the 8th translation for *auto* in Finnish is *crankshaft*. There is also a similar word returned within the top 10 list for French (*bilha*), Spanish (*bilha*) and Finnish (*bilhan*). Bilha is the name of a character in the Bible, however it is not clear why has been correlated with *car*.

Further raw results tables are shown in appendix (B.2 Supplementary Results).

Chapter 7

Discussion

The monolingual enrichment approach described here is a simple extension of the (Levy, Søgaard, and Goldberg 2017) baseline Bible corpus Multilingual SID-SGNS model. Intuitively, it seemed a reasonable hypothesis that this could have some positive impact on the bilingual dictionary induction task in increasing the vocabulary coverage. However, the trials undertaken here achieved no significant improvement when evaluated against the full vocabulary of the Wiktionary dataset. There could be several factors involved in this, which warrant further consideration before abandoning this approach.

7.1 Enrichment Datasets

The quality of the data retrieved from the two enrichment sources, Twitter and Wikipedia, was variable. In both cases, additional pre-processing steps could be added, to cleanse the data more thoroughly. For example, regular expressions could be used to strip out Twitter hashtags and handles, URLs, numerical values and residual Wikimedia mark-up.

Subjectively, it appeared that the Wikipedia data was of better quality and more likely to be able to be cleansed into a usable corpus. This could be quantitatively evaluated by analysing statistics on the proportion of mis-spellings, non-words and words from a different language from the page language. One advantage of Twitter is that the service inherently limits tweet length to 280 characters, so the sentence-ID feature is guaranteed to provide good localisation of co-occurrence.

It would be worthwhile investigating the published Wikipedia corpus used in the original BilBOWA experiments (Gouws, Bengio, and Corrado 2014; Al-Rfou, Perozzi, and Skiena 2013). Whilst it was a useful educational experience to attempt to build a corpus from scratch, the data cleansing applied to the published corpus may be more effective than that achieved in this short project.

Although BilBOWA takes a different approach to the problem, the motivations behind it are the same as in this project i.e. to use large monolingual corpora to extend the vocabulary coverage of a much smaller cross-lingual parallel corpus. A comparative study of BilBOWA versus the Enriched Multilingual SID-SGNS approach described here might help to establish whether the data inputs or the language model was the more significant factor in the model's poor precision. Both models could be trained with both the published Wikipedia corpus and the home-grown Wikipedia corpus. This would differentiate between data quality factors and the relative efficacy of the two models.

7.2 Enrichment Sentence Ranking

The vocabulary frequency distributions in figure (4.2g & 4.2h) show that the ranking has boosted occurrence of words in the middle frequency range, as desired. However, the summation of the TF-IDF score over the sentences had the unintentional side-effect of favouring very long sentences. The sentence length distributions are plotted in figure (7.1). The majority of top ranked sentences had length >50 words, and the maximum lengths for each language were in the 100s. Consequently, some very anomalous sentences have been selected, e.g. long lists of words, rather than high quality grammatical prose in a coherent subject domain. This greatly reduces the co-occurrence information provided by the sentence-ID feature; with such long sentences, many words co-occur with one another and fine semantic distinctions cannot be derived from the data.

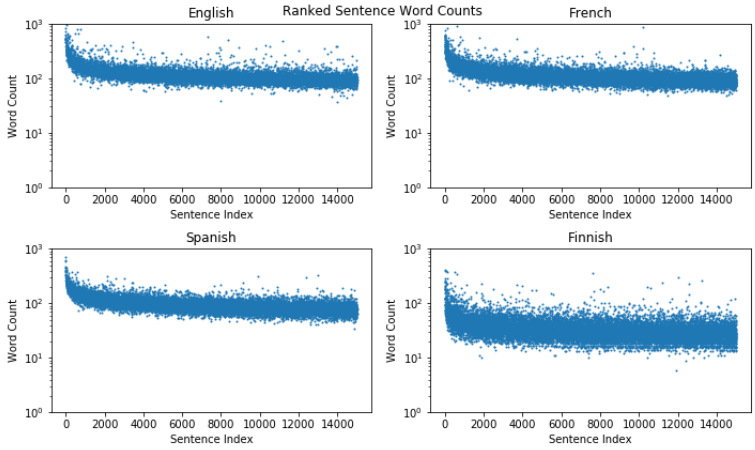


Figure 7.1: Wikipedia Ranked enrichment data - the distribution of sentence word counts when ranked by *score*. Maximum word counts: ['en': 1840, 'es': 722, 'fi': 409, 'fr': 910].

The alternative normalized score was proposed to counter this problem. However, a quick qualitative observation of the top ranked sentences showed that it had the opposite effect. The ranking now favoured very short sentences (sometimes only one word) made up of the rarest words. These sentences also provide little useful co-occurrence signal to the model. Due to project time constraints, there was only time to train and evaluate the model with one of these datasets, and the first was arbitrarily selected.

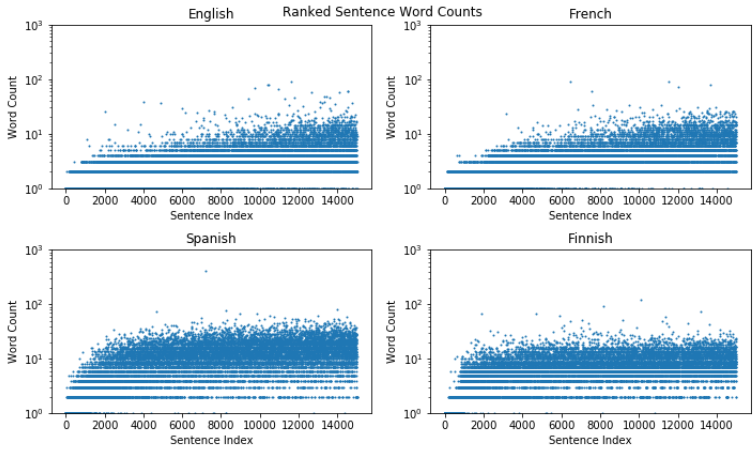


Figure 7.2: Wikipedia Ranked enrichment data - the distribution of sentence word counts when ranked by *normalized score*. Maximum word counts: ['en': 92, 'es': 404, 'fi': 122, 'fr': 92].

In hindsight, having plotted the sentence length distributions for both scoring methods (figure 7.2), the normalized score appears to return a high proportion of sentences in a length range more typical of good quality prose (8-20 words), so this method may have been a better choice to use. This could be combined with a simple heuristic that filters out sentences of length <8, for example. It would also be worth analysing the sentence characteristics (word count and TF-IDF) of the Bible corpus, since some degree of success was achieved in the baseline using this data.

7.3 Linguistic Characteristics of Languages

Observations have been made throughout this report on the different characteristics of the languages under test. Finnish, in particular, stands out from the other languages, due to its polysynthetic morphology and the many different inflections that arise from the same word stem. It would be interesting to investigate whether some pre-processing improves the lexical induction precision scores for languages like Finnish, such as stemming or tokenizing into morphemes rather than words.

7.4 Processing Performance

The method described here was shown to be feasible to run on single instance commodity infrastructure, although some processes did take a long time to run. Depending on the size of the enrichment data and the number of languages being enriched, training the Enriched Multilingual SID-SGNS model took about 1.5 to 3 hours, multiplied by 10 trials for each experiment. The Wiktionary evaluation of 3 language pairs over 10 trials typically took about 24-32 hours.

The Twitter API rate limit throttled the rate at which tweets could be downloaded (720 tweets/hour). Although the Wikipedia API has no rate limits, the extract ran at about the same page rate per hour, because for each page a large proportion of the downloaded data was parsed and discarded in the HTML scraping process. However, each Wikipedia page contained on average about 15-30 usable sentences, so it was possible to generate a much larger corpus in the same amount time.

The most time-consuming part of the project was the ranking process. Loading the full set of 19 million sentences into MongoDB took about two weeks, with the aggregation and scoring steps outlined in (§4.4) taking a further two weeks. As with most database solutions, it was particularly important to ensure that the appropriate indexes had been set up in the database.

Many of these steps could be parallelised on a clustered environment to achieve improvements in execution time.

Chapter 8

Conclusion

Word count: 12866

Bibliography

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena (2013). "Polyglot: Distributed Word Representations for Multilingual NLP". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 183–192. ISBN: 1532-4435. DOI: 10.1007/s10479-011-0841-3. arXiv: 1307.1662.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A Neural Probabilistic Language Model". In: *The Journal of Machine Learning Research* 3, pp. 1137–1155. ISSN: 15324435. DOI: 10.1162/153244303322533223. arXiv: arXiv:1301.3781v3.

Bond, Francis and Ryan Foster (2013). "Linking and extending an open multilingual Wordnet". In: *Proceedings of ACL 2013*, pp. 1352–1362.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). "The mathematics of statistical machine translation: Parameter estimation". In: *Computational linguistics* 19, pp. 263–311. ISSN: 08912017.

Çakmak, Mehmet Talha, Süleyman Acar, and Gülşen Eryiğit (2012). "Word Alignment for English-Turkish Language Pair". In: *Lrec*, pp. 2177–2180.

Chandar, Sarath A P, Stanislas Lauly, Hugo Larochelle, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha (2014). "An Autoencoder Approach to Learning Bilingual Word Representations". In: ISSN: 10495258. arXiv: 1402.1454.

Christodouloupoulos, Christos and Mark Steedman (2015). "A massively parallel corpus: the Bible in 100 languages". In: *Language Resources and Evaluation* 49.2, pp. 375–395. ISSN: 15728412. DOI: 10.1007/s10579-014-9287-y.

Cieri, Christopher, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey (2016). "Selection Criteria for Low Resource Language Programs". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4543–4549.

Collobert, Ronan and Jason Weston (2008). "A unified architecture for natural language processing". In: *Proceedings of the 25th international conference on Machine learning - ICML '08*, pp. 160–167. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. arXiv: 1603.06111.

Faruqui, Manaal and Chris Dyer (2014). "Improving Vector Space Word Representations Using Multilingual Correlation". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471. ISBN: 9781632663962. DOI: 10.3115/v1/E14-1049.

Goldberg, Yoav and Omer Levy (2014). "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: ISSN: 0003-6951. DOI: 10.1162/jmlr.2003.3.4-5.951. arXiv: 1402.3722.

Gouws, Stephan, Yoshua Bengio, and Greg Corrado (2014). "BilBOWA: Fast Bilingual Distributed Representations without Word Alignments". In: arXiv: 1410.2455.

Graca, Joao, Joana Paulo Pardal, Luisa Coheur, and Diamantino Caseiro (2008). "Building a Golden Collection of Parallel Multi-Language Word Alignment". In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. ISBN: 2-9517408-4-0.

Holmqvist, Maria and Lars Ahrenberg (2011). "A Gold Standard for English-Swedish Word Alignment". In: *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*. 11, pp. 106–113.

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai (2012). "Inducing Crosslingual Distributed Representations of Words". In: *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers (2012)* December, pp. 1459–1474. ISSN: 15324435. DOI: 10.1162/153244303322533223. arXiv: 1301.3781.

Koehn, P. (2010). "Word-based Models". In: *Statistical Machine Translation*. Cambridge University Press. Chap. 4, pp. 81–125. ISBN: 978-0-511-68759-4.

Koehn, Philipp (2005). "Europarl : A Parallel Corpus for Statistical Machine Translation". In: *MT Summit* 11, pp. 79–86. ISSN: 9747431262. DOI: 10.3115/1626355.1626380.

Lambert, Patrik, Adrià De Gispert, Rafael Banchs, and José B. Mariño (2005). "Guidelines for word alignment evaluation and manual alignment". In: *Language Resources and Evaluation* 39.4, pp. 267–285. ISSN: 1574020X. DOI: 10.1007/s10579-005-4822-5.

Levy, Omer, Anders Søgaard, and Yoav Goldberg (2017). "A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments". In: *EACL*.

Mihalcea, Rada and Ted Pedersen (2003). "An evaluation exercise for word alignment". In: *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts data driven machine translation and beyond* - 3.June, pp. 1–10. DOI: 10.3115/1118905.1118906.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation". In: ISSN: 10495258. DOI: 10.1162/153244303322533223. arXiv: 1309.4168.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations". In: *Proceedings of NAACL-HLT* June, pp. 746–751.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Distributed Representations of Words and Phrases and their Compositionality". In: ISSN: 10495258. DOI: 10.1162/jmlr.2003.3.4-5.951. arXiv: 1310.4546.

Mikolov, Tomas, Greg Corrado, Kai Chen, and Jeffrey Dean (2013b). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12. ISSN: 15324435. DOI: 10.1162/153244303322533223. arXiv: arXiv:1301.3781v3.

Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational Linguistics* 29.1, pp. 19–51. ISSN: 0891-2017. DOI: 10.1162/089120103321337421.

Rowling, J. K. (1997). *Harry Potter and the Philosopher's Stone*. 1st ed. Vol. 1. London: Bloomsbury Publishing. ISBN: 978-0747532699.

Ruder, Sebastian (2016). "A survey of cross-lingual embedding models". In: 6. arXiv: 1706.04902.

Søgaard, Anders, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen (2015). "Inverted indexing for cross-lingual NLP". 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)*, pp. 1713–1722.

Upadhyay, Shyam, Manaal Faruqui, Chris Dyer, and Dan Roth (2016). "Cross-lingual Models of Word Embeddings: An Empirical Comparison". In: *ACL*, pp. 1661–1670. ISBN: 9781510827585. arXiv: 1604.00425.

Vulić, Ivan and Marie Francine Moens (2016). "Bilingual distributed word representations from document-aligned comparable data". In: *Journal of Artificial Intelligence Research* 55, pp. 953–994. ISSN: 10769757. arXiv: 1509.07308.

Xiao, Min and Yuhong Guo (2014). "Distributed Word Representation Learning for Cross-Lingual Dependency Parsing". In: *CoNLL*, pp. 119–129.

Zou, Will Y, Richard Socher, Daniel Cer, and Christopher D Manning (2013). "Bilingual Word Embeddings for Phrase-Based Machine Translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)* October, pp. 1393–1398. ISSN: 9781937284978.

Appendices

Appendix A

Ethical Review

A.1 Self-Assessment Checklist

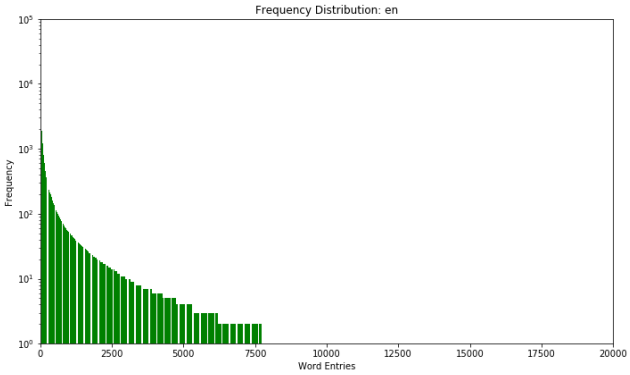
Question	Answer
1.a Will the research project involve human participants, with or without their knowledge or consent at the time? (This includes yourself if you are the main subject of the research).	No
b. Will the research project involve animals	No
2. Is the research project likely to expose any person, whether or not a participant, to physical or psychological harm?	No
3. Will you have access to personal information that allows you to identify individuals or to confidential corporate or company information?	No
4. Does the research project present a significant risk to the environment or society?	No
5. Are there any ethical issues raised by this research project that require further ethical review?	No

Appendix B

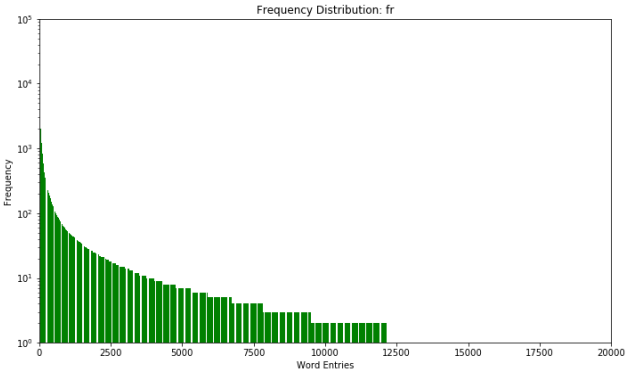
Supplementary Results

B.1 Vocabulary Frequency Distributions

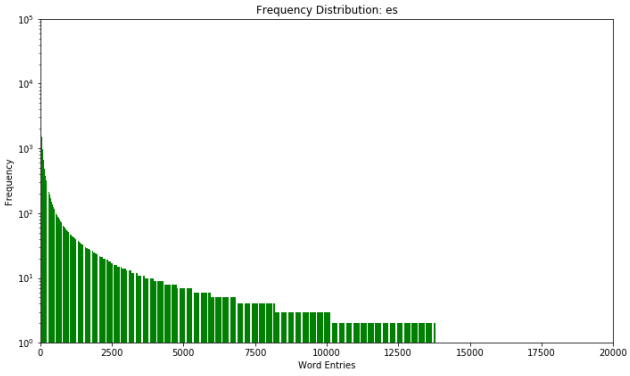
B.1.1 Bible Corpus Baseline



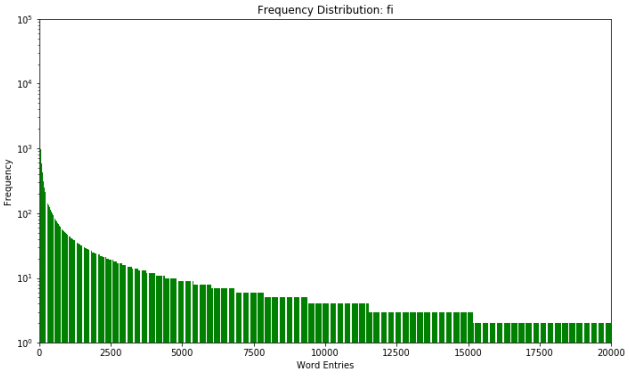
(a) English: $freq(w) : w \in C_{B_{en}}$



(b) French: $freq(w) : w \in C_{B_{fr}}$



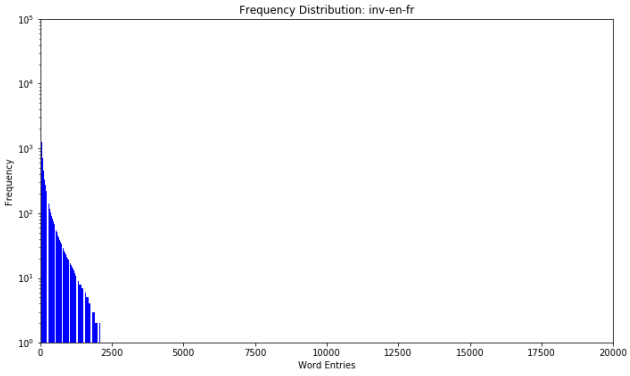
(c) Spanish: $freq(w) : w \in C_{B_{es}}$



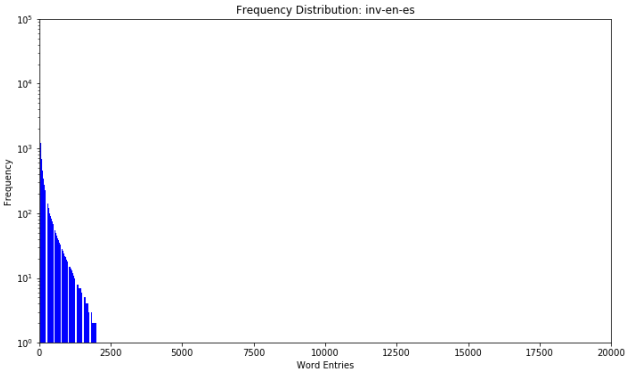
(d) Finnish: $freq(w) : w \in C_{B_{fi}}$

Figure B.1: Plots showing the vocabulary frequency distributions for the respective languages in the Bible corpus C_{B_l}.

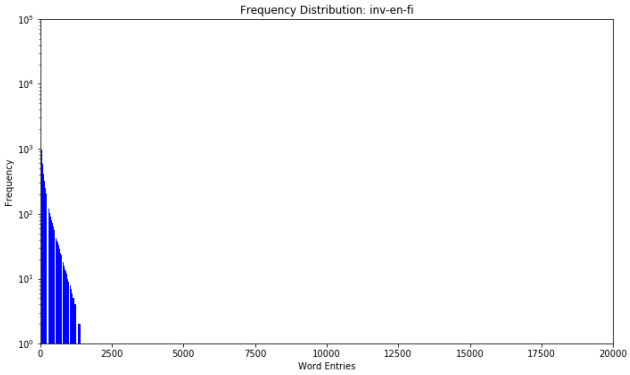
B.1.2 Bible Corpus In-Vocabulary (INV)



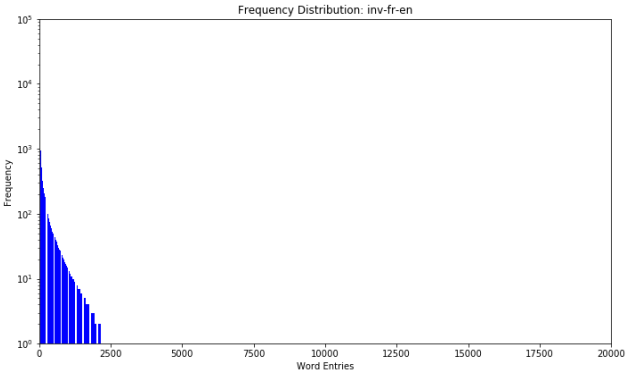
(a) English/French: $freq(w) : w \in \{C_{B_{en}} \cap E_{W_{en-fr}}\}$



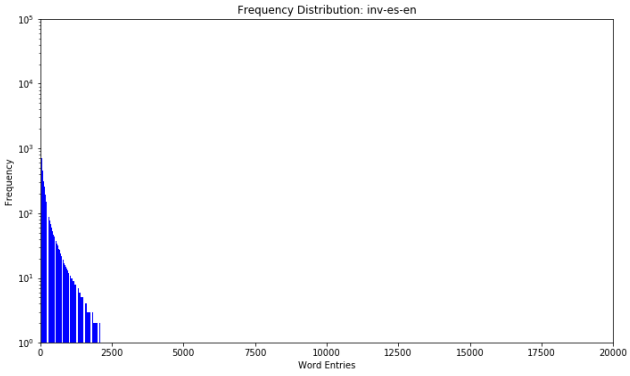
(b) English/Spanish: $freq(w) : w \in \{C_{B_{en}} \cap E_{W_{en-es}}\}$



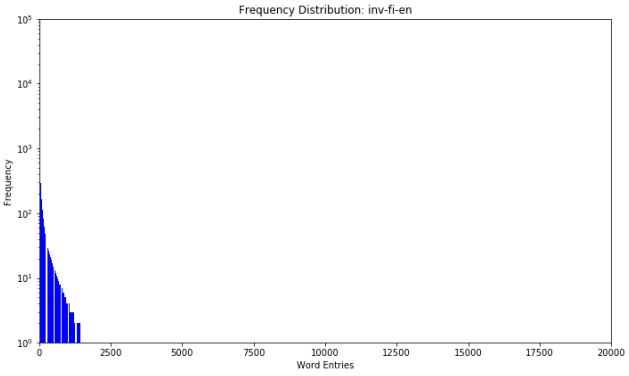
(c) English/Finnish: $freq(w) : w \in \{C_{B_{en}} \cap E_{W_{en-fi}}\}$



(d) French/English: $freq(w) : w \in \{C_{B_{fr}} \cap E_{W_{fr-en}}\}$



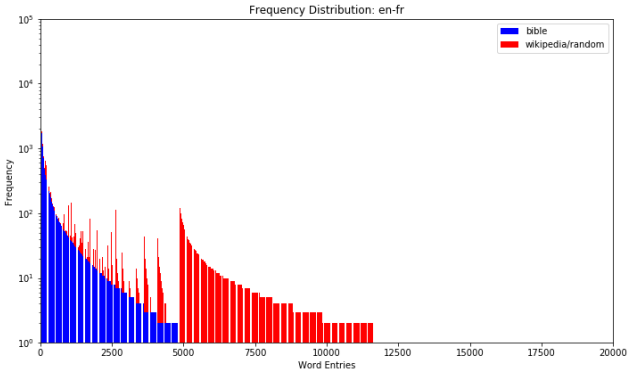
(e) Spanish/English: $freq(w) : w \in \{C_{B_{es}} \cap E_{W_{es-en}}\}$



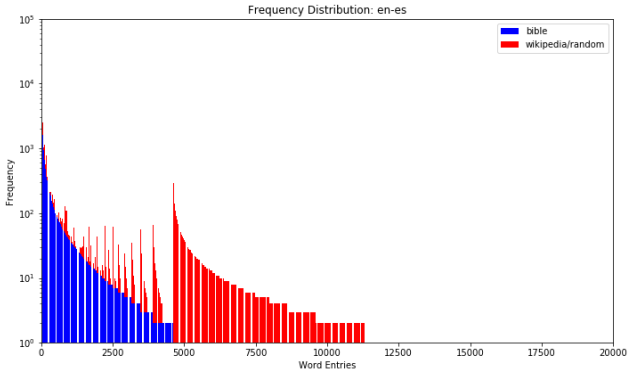
(f) Finnish/English: $freq(w) : w \in \{C_{B_{fi}} \cap E_{W_{fi-en}}\}$

Figure B.2: Plots showing frequency distributions of the subset of Bible corpus words $C_{B_{l_1}}$ that are INV in each of the Wiktionary evaluation sets $E_{W_{l_1-l_2}}$.

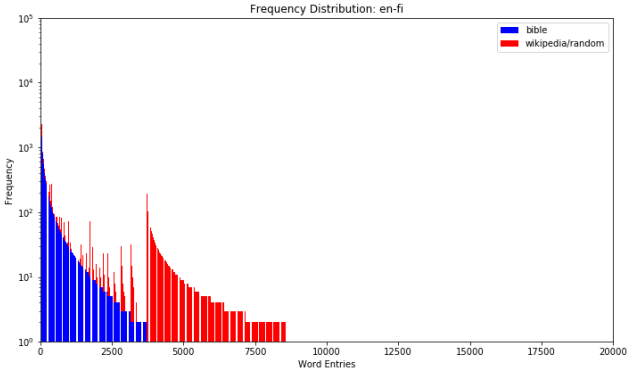
B.1.3 Bible Corpus Enriched with Randomized Wikipedia Corpus



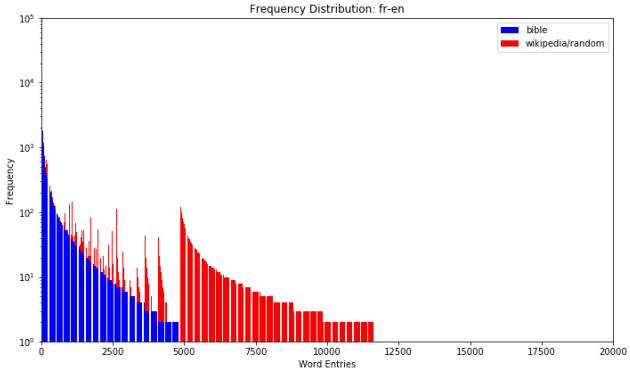
(a) English/French:
$freq(w) : w \in \{\{C_{B_{en}} \cup Randomized(C_{W_{en}})\} \cap E_{W_{en-fr}}\}$



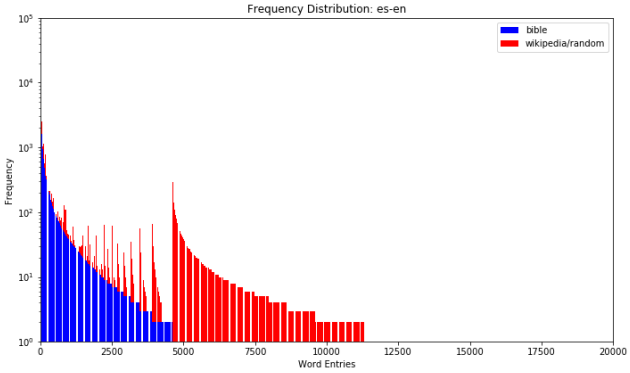
(b) English/Spanish:
$freq(w) : w \in \{\{C_{B_{en}} \cup Randomized(C_{W_{en}})\} \cap E_{W_{en-es}}\}$



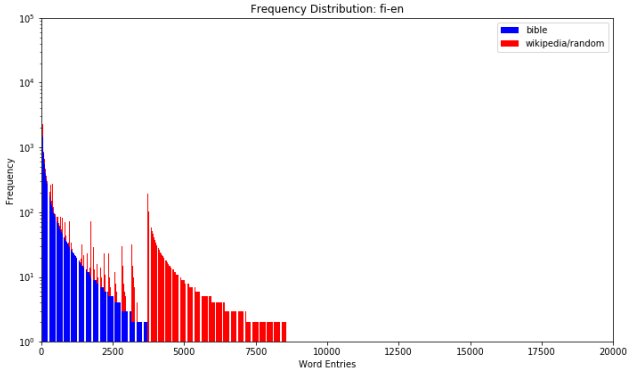
(c) English/Finnish:
$freq(w) : w \in \{\{C_{B_{en}} \cup Randomized(C_{W_{en}})\} \cap E_{W_{en-fi}}\}$



(d) French/English:
$freq(w) : w \in \{\{C_{B_{fr}} \cup Randomized(C_{W_{fr}})\} \cap E_{W_{fr-en}}\}$



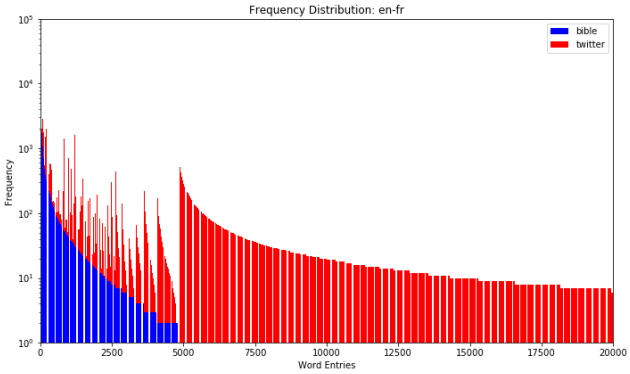
(e) Spanish/English:
$freq(w) : w \in \{\{C_{B_{es}} \cup Randomized(C_{W_{es}})\} \cap E_{W_{es-en}}\}$



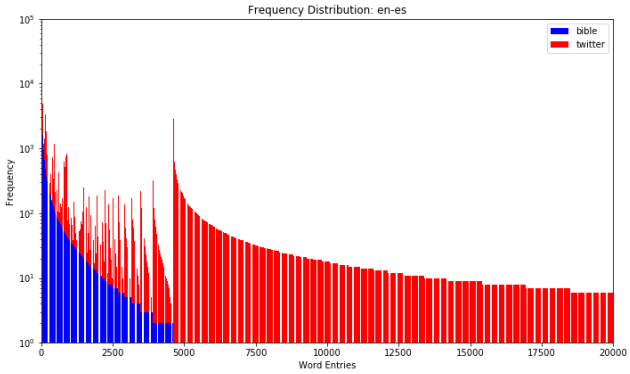
(f) Finnish/English:
$freq(w) : w \in \{\{C_{B_{fi}} \cup Randomized(C_{W_{fi}})\} \cap E_{W_{fi-en}}\}$

Figure B.3: Plots showing frequency distributions of the subset of words from the union of the Bible corpus $C_{B_{l_1}}$ and the randomized Wikipedia enrichment corpus $Randomized(C_{W_{l_1}})$ that are INV in each of the Wiktionary evaluation sets $E_{W_{l_1-l_2}}$.

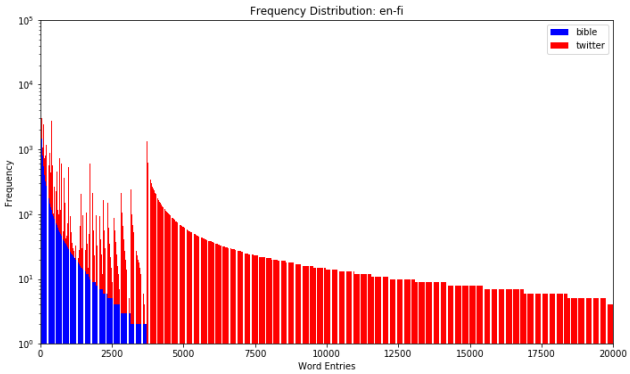
B.1.4 Bible Corpus Enriched with Twitter Search Results



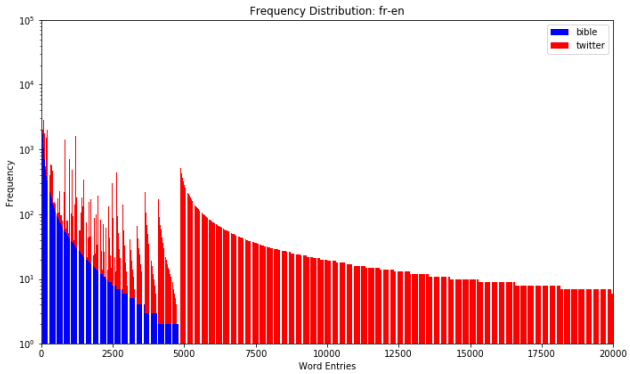
(a) English/French:
$freq(w) : w \in \{\{C_{B_{en}} \cup C_{T_{en}}\} \cap E_{W_{en-fr}}\}$



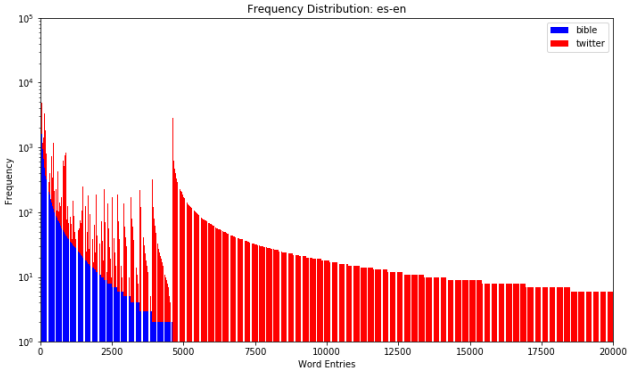
(b) English/Spanish:
$freq(w) : w \in \{\{C_{B_{en}} \cup C_{T_{en}}\} \cap E_{W_{en-es}}\}$



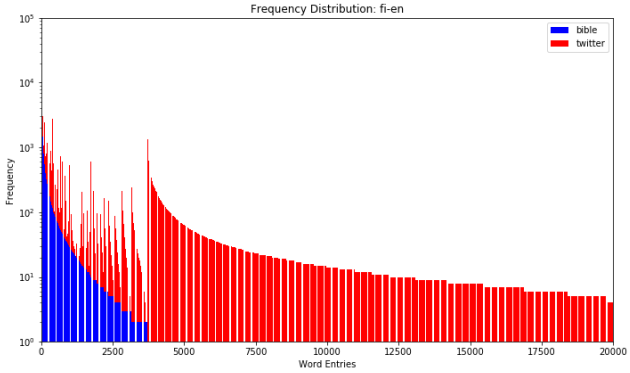
(c) English/Finnish:
$freq(w) : w \in \{\{C_{B_{en}} \cup C_{T_{en}}\} \cap E_{W_{en-fi}}\}$



(d) French/English:
$freq(w) : w \in \{\{C_{B_{fr}} \cup C_{T_{fr}}\} \cap E_{W_{fr-en}}\}$



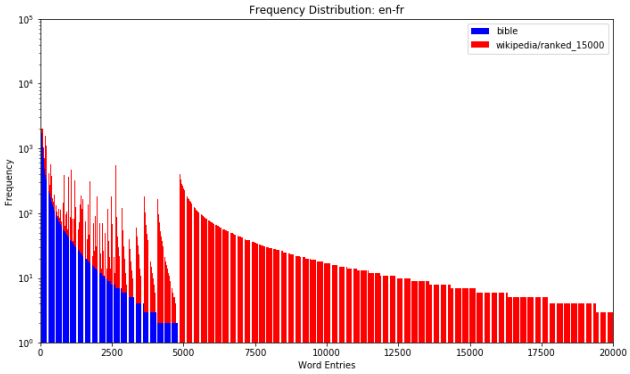
(e) Spanish/English:
$freq(w) : w \in \{\{C_{B_{es}} \cup C_{T_{es}}\} \cap E_{W_{es-en}}\}$



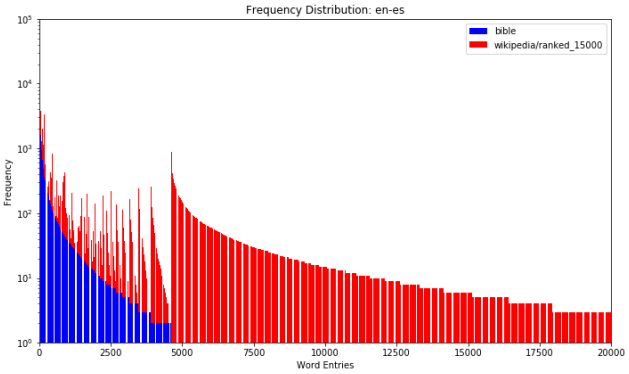
(f) Finnish/English:
$freq(w) : w \in \{\{C_{B_{fi}} \cup C_{T_{fi}}\} \cap E_{W_{fi-en}}\}$

Figure B.4: Plots showing frequency distributions of the subset of words from the union of the Bible corpus $C_{B_{l_1}}$ and the Twitter enrichment corpus $C_{T_{l_1}}$ that are INV in each of the Wiktionary evaluation sets $E_{W_{l_1-l_2}}$.

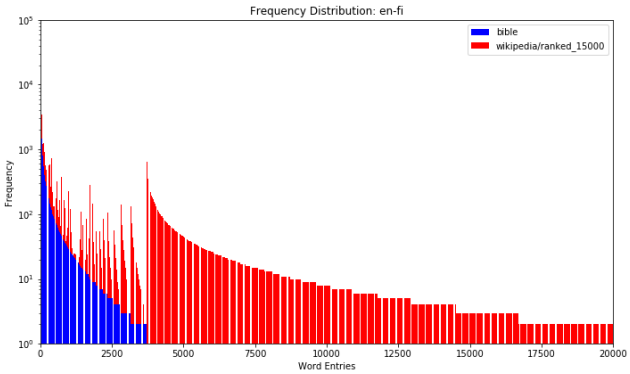
B.1.5 Bible Corpus Enriched with Ranked Wikipedia Corpus



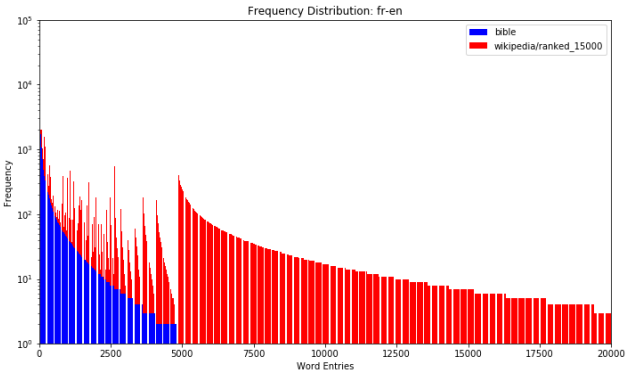
(a) English/French:
$freq(w) : w \in \{\{C_{B_{en}} \cup Ranked(C_{W_{en}})\} \cap E_{W_{en-fr}}\}$



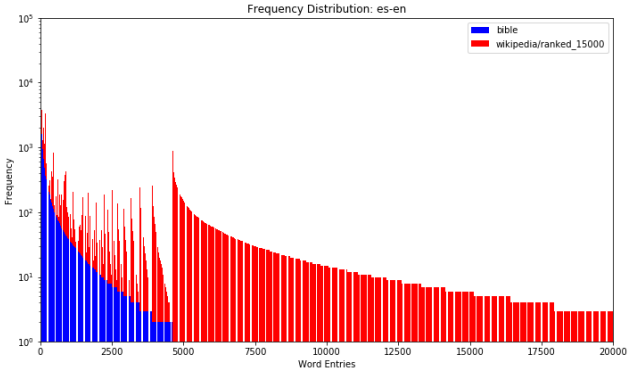
(b) English/Spanish:
$freq(w) : w \in \{\{C_{B_{en}} \cup Ranked(C_{W_{en}})\} \cap E_{W_{en-es}}\}$



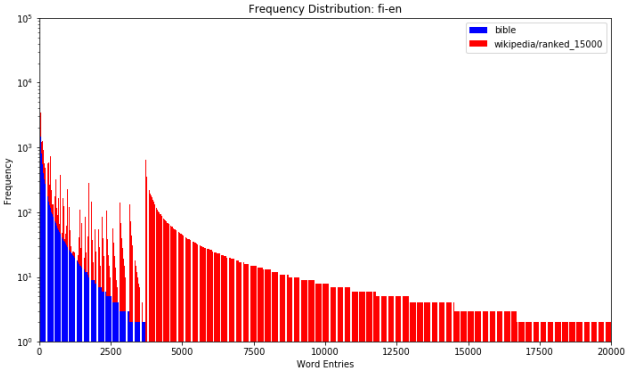
(c) English/Finnish:
$freq(w) : w \in \{\{C_{B_{en}} \cup Ranked(C_{W_{en}})\} \cap E_{W_{en-fi}}\}$



(d) French/English:
$freq(w) : w \in \{\{C_{B_{fr}} \cup Ranked(C_{W_{fr}})\} \cap E_{W_{fr-en}}\}$



(e) Spanish/English:
$freq(w) : w \in \{\{C_{B_{es}} \cup Ranked(C_{W_{es}})\} \cap E_{W_{es-en}}\}$



(f) Finnish/English:
$freq(w) : w \in \{\{C_{B_{fi}} \cup Ranked(C_{W_{fi}})\} \cap E_{W_{fi-en}}\}$

Figure B.5: Plots showing frequency distributions of the subset of words from the union of the Bible corpus $Ranked(C_{B_{l_1}})$ and the ranked Wikipedia enrichment corpus $C_{W_{l_1}}$ that are INV in each of the Wiktionary evaluation sets $E_{W_{l_1-l_2}}$.

B.2 Multilingual SID-SGNS Baseline

B.2.1 In-Vocabulary (INV) Evaluation Words

B.2.1.1 Quantitative Benchmark Results

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.351	0.162	0.334	0.387	0.258	0.388
1	0.351	0.158	0.333	0.385	0.256	0.394
2	0.353	0.156	0.333	0.383	0.255	0.389
3	0.350	0.156	0.331	0.387	0.259	0.391
4	0.354	0.157	0.329	0.388	0.259	0.392
5	0.352	0.159	0.331	0.386	0.257	0.388
6	0.350	0.157	0.331	0.382	0.257	0.389
7	0.352	0.156	0.331	0.385	0.259	0.394
8	0.352	0.154	0.333	0.385	0.257	0.392
9	0.353	0.159	0.331	0.388	0.260	0.393
μ_{sample}	0.352	0.157	0.332	0.385	0.258	0.391
μ_{H0}	0.351	0.159	0.330	0.387	0.258	0.389
t-statistic	2.182	-2.391	3.068	-2.178	-1.673	2.129
p-value	0.057	0.040	0.013	0.057	0.129	0.062

Table B.1: Raw Precision@1 results against Wiktionary lexical induction evaluation set, for 10 training instances recreating the baseline multilingual SID-SGNS trained on the Bible corpus. The sample mean μ_{sample} is compared against the null hypothesis μ_{H0}, the results given in (Levy, Søgaard, and Goldberg 2017), which are assumed to be the mean of several trials (although this is not explicitly stated in their paper). Note: The baseline evaluation ignored any OOV words in the evaluation set.

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.556	0.400	0.555	0.569	0.427	0.569
1	0.556	0.402	0.555	0.574	0.427	0.573
2	0.556	0.401	0.551	0.567	0.429	0.569
3	0.555	0.401	0.551	0.569	0.429	0.571
4	0.552	0.401	0.549	0.568	0.421	0.570
5	0.554	0.401	0.553	0.571	0.430	0.572
6	0.555	0.401	0.553	0.571	0.425	0.571
7	0.554	0.402	0.550	0.568	0.425	0.570
8	0.554	0.400	0.551	0.567	0.424	0.568
9	0.554	0.402	0.555	0.571	0.427	0.572
μ_{sample}	0.555	0.401	0.552	0.569	0.426	0.570

Table B.2: Raw Precision@10 results against Wiktionary lexical induction evaluation set, for 10 training instances recreating the baseline multilingual SID-SGNS trained on the Bible corpus. Levy et al did not provide Precision@10 results, so we have no null hypothesis to compare against. However, these results will form the baseline for later comparative tests.

B.2.1.2 Qualitative Analysis Results

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	mer [sea]	en	sea	0.940	en	sea	fr	mer [sea]	0.940
			shore	0.621				rouge [red]	0.580
			seas	0.608				mers [seas]	0.565
			waves	0.580				côte [coast]	0.562
			sand	0.571				sable [sand]	0.560
			navy	0.490				occidentale [western]	0.522
			calm	0.480				flots [waves]	0.487
			roar	0.477				rivage [shore]	0.479
			red	0.475				mugissent [roar]	0.477
			ships	0.469				aboutir [to arrive at, to lead to]	0.476
es	mar [sea]	en	sea	0.916	en	sea	es	mar [sea]	0.916
			seas	0.593				mares [seas]	0.626
			shore	0.592				arena [sand]	0.574
			waves	0.575				rojo [red]	0.543
			sand	0.543				olas [waves]	0.514
			ships	0.485				peces [fish]	0.491
			navy	0.482				costa [coast]	0.487
			red	0.473				navíos [ships]	0.481
			calm	0.464				orilla [shore]	0.474
			shipmen	0.452				ondas [waves]	0.473
fi	meri [sea]	en	sea	0.765	en	sea	fi	meren [marine]	0.858
			waves	0.607				meri [sea]	0.765
			seas	0.601				merta [fish trap]	0.743
			shore	0.595				mereen [sea]	0.722
			calm	0.546				merestä [sea]	0.653
			raging	0.539				meressä [in the sea]	0.648
			roar	0.531				järven [lake]	0.606
			sand	0.528				kaislameren [reed sea]	0.603
			tempestuous	0.491				rannalla [onshore]	0.587
			moveth	0.475				merellä [at sea]	0.575

Table B.3: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the word *sea*.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	homme [man]	en	man	0.750	en	man	fr	homme [man]	0.750
			fellow	0.357				personne [no one]	0.377
			adam	0.347				chacun [each]	0.365
			mankind	0.324				humaine [human]	0.336
			imputeth	0.311				vaillant [valiant]	0.335
			winebibber	0.300				vieillard [old man]	0.327
			delighteth	0.295				quelqu [somebody]	0.324
			discerned	0.294				acception [sense, meaning]	0.317
			pondereth	0.293				arrogant [arrogant]	0.297
			unlawful	0.286				inutile [useless]	0.296
es	hombre [man]	en	man	0.740	en	man	es	hombre [man]	0.740
			adam	0.389				alguien [someone]	0.436
			fellow	0.372				nadie [no one]	0.417
			mankind	0.362				humano [human]	0.412
			imputeth	0.330				varón [male]	0.367
			husband	0.325				humana [human]	0.305
			winebibber	0.325				humanos [humans]	0.301
			delighteth	0.324				inteligente [intelligent]	0.299
			pondereth	0.322				alguno [any]	0.297
			unlawful	0.312				hombres [men]	0.290
fi	mies [man]	en	man	0.566	en	man	fi	mies [man]	0.566
			husband	0.428				ihminen [man, human]	0.535
			fellow	0.370				miehen [male]	0.509
			adam	0.321				ihmisen [man, human]	0.496
			adulteress	0.317				miehelle [the man]	0.480
			householder	0.309				ihmistä [men, humans]	0.479
			delighteth	0.307				ihmiselle [men, humans]	0.466
			cornelius	0.304				mieheksi [man]	0.417
			ruling	0.304				miestä [a man]	0.413
			eloquent	0.303				ihmislapsi [son of man]	0.409

Table B.4: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the word *man*.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	femme [woman]	en	woman	0.754	en	woman	fr	femme [woman]	0.754
			wife	0.732				femmes [women]	0.537
			wives	0.475				mari [husband]	0.511
			women	0.469				maîtresse [mistress]	0.482
			married	0.445				mariée [married]	0.452
			husband	0.435				amant [lover]	0.440
			maid	0.415				prostituée [prostitute]	0.440
			adulteress	0.399				répudiée [repudiated]	0.432
			concubine	0.395				servante [maid]	0.426
			female	0.395				rasée [shaved]	0.420
es	mujer [woman]	en	woman	0.792	en	woman	es	mujer [woman]	0.792
			wife	0.756				mujeres [women]	0.542
			wives	0.483				esposa [wife]	0.531
			women	0.472				marido [husband]	0.493
			husband	0.465				encinta [pregnant]	0.484
			married	0.451				prostituta [prostitute]	0.484
			female	0.411				casada [married]	0.468
			adulteress	0.399				esposo [husband]	0.468
			concubine	0.398				sentada [sitting]	0.443
			maid	0.393				profetisa [prophetess]	0.431
fi	nainen [woman]	en	woman	0.749	en	woman	fi	nainen [woman]	0.749
			adulteress	0.544				vaimo [wife]	0.708
			husband	0.504				naisen [the woman]	0.673
			female	0.500				vaimon [a wife]	0.656
			maid	0.499				vaimolle [the wife]	0.612
			women	0.495				naista [a woman]	0.600
			wife	0.474				vaimoa [a wife]	0.565
			bondmaid	0.467				naiset [ladies]	0.539
			whore	0.467				naiseen [lady]	0.533
			adulterer	0.447				tyttö [girl]	0.506

Table B.5: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the word *woman*.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	agneau [lamb]	en	lamb	0.925	en	lamb	fr	agneau [lamb]	0.925
			lambs	0.543				agneaux [lambs]	0.652
			goats	0.516				brebis [ewe]	0.504
			sheep	0.513				bélier [ram]	0.500
			kid	0.512				immolé [immolated]	0.495
			firstling	0.479				béliers [rams]	0.491
			rams	0.476				bouc [billy goat]	0.472
			flock	0.468				femelle [female]	0.472
			ram	0.467				culpabilité [guilt]	0.465
			ewe	0.464				chèvre [goat]	0.462
es	cordero [lamb]	en	lamb	0.900	en	lamb	es	cordero [lamb]	0.900
			lambs	0.569				corderos [lambs]	0.634
			kid	0.524				inmolado [immolated]	0.534
			goats	0.494				carneros [rams]	0.521
			ram	0.491				oveja [sheep]	0.503
			ewe	0.482				rebaño [flock]	0.500
			flock	0.480				cabrito [kid]	0.490
			rams	0.477				machos [males]	0.489
			sheep	0.472				macho [male]	0.488
			bullocks	0.469				carnero [ram]	0.487
fi	karitsa [lamb]	en	lamb	0.789	en	lamb	fi	karitsan [lamb]	0.885
			lambs	0.552				karitsa [lamb]	0.789
			ewe	0.545				karitsaa [lambs]	0.680
			kid	0.518				lampaan [sheep]	0.544
			goat	0.486				lammas [sheep]	0.522
			goats	0.475				syntiuhriksi [sacrifice for a sin]	0.507
			blemish	0.475				teurastettu [slaughtered]	0.502
			rams	0.468				vuohista [goats]	0.501
			flock	0.462				virheettömän [flawless]	0.501
			calf	0.449				uuhikaritsan [female lamb]	0.498

Table B.6: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the word *lamb*.

B.2.2 Out-of-Vocabulary (OOV) Evaluation Words

B.2.2.1 Quantitative Benchmark Results

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0415	0.0100	0.0363	0.0457	0.0159	0.0422
1	0.0415	0.0097	0.0362	0.0454	0.0157	0.0428
2	0.0417	0.0096	0.0363	0.0452	0.0157	0.0423
3	0.0414	0.0096	0.0360	0.0457	0.0159	0.0425
4	0.0418	0.0096	0.0358	0.0459	0.0159	0.0426
5	0.0416	0.0098	0.0360	0.0456	0.0158	0.0422
6	0.0414	0.0097	0.0360	0.0452	0.0158	0.0423
7	0.0416	0.0096	0.0360	0.0455	0.0159	0.0428
8	0.0416	0.0095	0.0363	0.0454	0.0158	0.0427
9	0.0417	0.0098	0.0360	0.0459	0.0160	0.0427
μ_{sample}	0.0416	0.0097	0.0361	0.0455	0.0159	0.0425

Table B.7: Raw Precision@1 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, including OOV words in the evaluation set. Levy et al did not provide results for OOV words, so we have no null hypothesis to compare against. However, these results will form the baseline for later comparative tests.

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0658	0.0246	0.0604	0.0674	0.0267	0.0620
1	0.0658	0.0248	0.0604	0.0681	0.0267	0.0625
2	0.0659	0.0247	0.0599	0.0672	0.0268	0.0621
3	0.0657	0.0247	0.0599	0.0674	0.0269	0.0623
4	0.0654	0.0247	0.0597	0.0673	0.0263	0.0622
5	0.0655	0.0247	0.0601	0.0677	0.0269	0.0624
6	0.0657	0.0247	0.0602	0.0677	0.0266	0.0623
7	0.0656	0.0247	0.0598	0.0674	0.0266	0.0622
8	0.0656	0.0246	0.0600	0.0673	0.0265	0.0619
9	0.0656	0.0247	0.0603	0.0677	0.0267	0.0624
μ_{sample}	0.0656	0.0247	0.0601	0.0675	0.0267	0.0622

Table B.8: Raw Precision@10 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, including OOV words in the evaluation set. Levy et al did not provide results for OOV words, so we have no null hypothesis to compare against. However, these results will form the baseline for later comparative tests.

B.2.2.2 Qualitative Analysis Results

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	film [film]	en	zur	0.000	en	film	fr	ôtée [removed]	0.000
			flames	0.000				espèrent [they hope]	0.000
			flow	0.000				espéré [hoped]	0.000
			flowed	0.000				espérons [we hope]	0.000
			flower	0.000				espérions [we hoped]	0.000
			flowers	0.000				espéreront [they will hope]	0.000
			floweth	0.000				espérer [to hope]	0.000
			flowing	0.000				espérances [hopes]	0.000
			flute	0.000				espérance [hope]	0.000
			fly	0.000				espérait [he hoped]	0.000
es	película [film]	en	zur	0.000	en	film	es	útiles [tools]	0.000
			flames	0.000				egipto [Egypt]	0.000
			flow	0.000				egipcio [Egyptian]	0.000
			flowed	0.000				egipcia [Egyptian]	0.000
			flower	0.000				efrón [Ephron]	0.000
			flowers	0.000				efraín [Ephrathan]	0.000
			floweth	0.000				efrateo [of Ephrath]	0.000
			flowing	0.000				efrata [Ephrath]	0.000
			flute	0.000				efod [ephod]	0.000
			fly	0.000				eficiente [efficient]	0.000
fi	elokuva [film]	en	zur	0.000	en	film	fi	öljyä [oil]	0.000
			flames	0.000				kunnioittakaa [to honour]	0.000
			flow	0.000				kunniansa [glory]	0.000
			flowed	0.000				kunniasi [glory]	0.000
			flower	0.000				kunniassa [glory]	0.000
			flowers	0.000				kunniasta [glory]	0.000
			floweth	0.000				kunnioita [honour]	0.000
			flowing	0.000				kunnioitan [to honour]	0.000
			flute	0.000				kunnioitat [to honour]	0.000
			fly	0.000				kunnioittaa [to honour]	0.000

Table B.9: Baseline multilingual SID-SGNS trained on the Bible corpus - translations of the OOV word *film*.

B.3 Multilingual SID-SGNS with Randomised Wikipedia Enrichment

B.3.1 Out-of-Vocabulary (OOV) Evaluation Words

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0407	0.0094	0.0361	0.0460	0.0161	0.0420
1	0.0409	0.0101	0.0357	0.0459	0.0157	0.0418
2	0.0416	0.0097	0.0361	0.0459	0.0159	0.0418
3	0.0414	0.0095	0.0360	0.0456	0.0159	0.0418
4	0.0416	0.0098	0.0357	0.0457	0.0160	0.0421
5	0.0411	0.0096	0.0358	0.0458	0.0158	0.0419
6	0.0412	0.0096	0.0363	0.0453	0.0157	0.0417
7	0.0411	0.0096	0.0362	0.0460	0.0159	0.0416
8	0.0412	0.0099	0.0362	0.0459	0.0159	0.0417
9	0.0412	0.0099	0.0364	0.0458	0.0160	0.0419
μ_{sample}	0.0412	0.0097	0.0360	0.0458	0.0159	0.0418
μ_{H0}	0.0416	0.0097	0.0361	0.0455	0.0159	0.0425
t-statistic	-3.9332	0.4573	-0.5078	3.7425	0.7429	-13.6739
p-value	0.0034	0.6583	0.6238	0.0046	0.4765	0.0000

Table B.10: Raw Precision@1 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, enriched with randomised Wikipedia sentences. The sample mean μ_{sample} is compared against the null hypothesis μ_{H0}, the mean result obtained earlier when training against the Bible corpus only.

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0658	0.0246	0.0598	0.0674	0.0260	0.0615
1	0.0658	0.0246	0.0596	0.0670	0.0256	0.0614
2	0.0655	0.0247	0.0600	0.0671	0.0255	0.0615
3	0.0660	0.0245	0.0595	0.0674	0.0255	0.0614
4	0.0654	0.0249	0.0599	0.0671	0.0254	0.0616
5	0.0654	0.0242	0.0601	0.0671	0.0252	0.0617
6	0.0656	0.0247	0.0597	0.0673	0.0255	0.0617
7	0.0659	0.0248	0.0599	0.0675	0.0256	0.0614
8	0.0655	0.0245	0.0598	0.0667	0.0255	0.0615
9	0.0658	0.0246	0.0596	0.0675	0.0254	0.0615
μ_{sample}	0.0657	0.0246	0.0598	0.0672	0.0255	0.0615
μ_{H0}	0.0656	0.0247	0.0601	0.0675	0.0267	0.0622
t-statistic	0.3716	-1.5830	-5.0711	-4.1402	-18.2758	-21.0837
p-value	0.7188	0.1479	0.0007	0.0025	0.0000	0.0000

Table B.11: Raw Precision@10 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, enriched with randomised Wikipedia sentences. The sample mean μ_{sample} is compared against the null hypothesis μ_{H0}, the mean result obtained earlier when training against the Bible corpus only.

B.4 Multilingual SID-SGNS with Twitter Search Enrichment

B.4.1 Out-of-Vocabulary (OOV) Evaluation Words

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0417	0.0096	0.0365	0.0457	0.0156	0.0415
μ_{H0}	0.0416	0.0097	0.0361	0.0455	0.0159	0.0425

Table B.12: Raw Precision@1 results against Wiktionary lexical induction evaluation set, for 1 training instance of multilingual SID-SGNS trained on the Bible corpus, enriched with tweets obtained via the Twitter search API. The single set of sample statistics is compared against the mean result obtained earlier when training against the Bible corpus only. The number of samples was reduced on this experiment to save test execution time - this was only intended as a rough indication of performance relative to the baseline.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	film [film]	en	incomparable	0.191	en	film	fr	#programming	0.173
			cannibals	0.187				https://t.co/t89xcyfcyp	0.169
			homestuck	0.183				chauds [hot]	0.168
			sacramento	0.182				parallélisme [parallelism]	0.166
			https://t.co/xo4c3lxdil	0.182				basses [low]	0.166
			https://t.co/27hdrrls7s	0.177				08:29	0.162
			cashing	0.172				anthem [-]	0.162
			zyrtec	0.170				presstalis [-]	0.160
			#hallowfic	0.169				climat [climate]	0.159
			new-difficult	0.168				passives [passive]	0.157
es	película [film]	en	diphthongs	0.190	en	film	es	presidirá [will preside]	0.188
			@yg_trece	0.189				acudiendo [coming]	0.184
			@mrkernow	0.188				@leopoldolopez	0.179
			#utsunomiya	0.186				una [a]	0.178
			suggestive	0.183				otw [-]	0.173
			retrofit	0.178				mapuche [-]	0.173
			#literary	0.176				https://t.co/peyooz9ntr	0.164
			#historical	0.176				ffccss [-]	0.160
			labourers	0.176				conicet [-]	0.160
			dicotyledon	0.176				https://t.co/kk3onbol0n	0.160
fi	elokuva [film]	en	#pacificrimuprising	0.190	en	film	fi	avauksessa [statement]	0.232
			https://t.co/46yafclbey	0.184				terassilla [terrace]	0.198
			employmen	0.177				juttu [thing]	0.185
			wikipedian	0.177				@timoaro	0.172
			bassett	0.174				peltirasia [-]	0.167
			home-made	0.168				helpoin [easiest]	0.166
			sharaga	0.168				103	0.165
			pubertal	0.167				ulkoministeriö [ministry]	0.163
			gauge	0.165				#luomuruoka [-]	0.160
			campaigns	0.165				kodikas [cozy]	0.159

Table B.13: Multilingual SID-SGNS with Twitter Search Enrichment - translations of the OOV word *film*.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	voiture [car]	en	3/1	0.188	en	car	fr	masqué [masked]	0.188
			cuticular	0.188				allemands [Germans]	0.183
			@solarteam_be	0.182				css [-]	0.175
			turboshaft	0.179				roquefort [-]	0.170
			jujube	0.177				bilha [-]	0.169
			diocletian	0.177				seas [-]	0.164
			alteplase	0.176				l'acteur [actor]	0.163
			@sami_kivela	0.172				gallois [Welsh]	0.163
			imperceptible	0.172				productrice [producer]	0.161
			@emartineeez	0.170				https://t.co/gy2vpmta36	0.161
es	coche [car]	en	cat's	0.203	en	car	es	bilha [-]	0.186
			liga	0.195				ela [-]	0.174
			@outfy	0.190				vtv [-]	0.172
			bills	0.184				hacian [they made]	0.170
			@cntraveler	0.177				@irenecido	0.165
			zzzz	0.175				util [useful]	0.165
			@randomportion	0.171				27/9	0.164
			@brine_gildchaff	0.170				anclado [anchored]	0.162
			#hashbased	0.170				ricaurte [-]	0.161
			10:00	0.167				elote [corn cob]	0.159
fi	auto [car]	en	ulcers	0.177	en	car	fi	metallinen [metal]	0.191
			10mins	0.177				bilhan [-]	0.189
			@bft_podcast	0.170				lähileipomon [bakery]	0.181
			@clydesdalearc	0.166				mikro-fuckin'-aalto...	0.173
			conroy	0.165				https://t.co/hp4fikf4h1	0.173
			acumen	0.165				42.4	0.171
			https://t.co/fl3pi16pwx	0.163				myöhästyy [late]	0.170
			crankshaft	0.163				@symbaaliapina	0.169
			bajrang	0.162				napsirkeun [-]	0.165
			contributions	0.162				eela [-]	0.163

Table B.14: Multilingual SID-SGNS with Twitter Search Enrichment - translations of the OOV word *car*.

B.5 Multilingual SID-SGNS with Ranked Wikipedia Enrichment

B.5.1 Out-of-Vocabulary (OOV) Evaluation Words

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0407	0.0094	0.0361	0.0460	0.0161	0.0420
1	0.0409	0.0101	0.0357	0.0459	0.0157	0.0418
2	0.0416	0.0097	0.0361	0.0459	0.0159	0.0418
3	0.0414	0.0095	0.0360	0.0456	0.0159	0.0418
4	0.0416	0.0098	0.0357	0.0457	0.0160	0.0421
5	0.0411	0.0096	0.0358	0.0458	0.0158	0.0419
6	0.0412	0.0096	0.0363	0.0453	0.0157	0.0417
7	0.0411	0.0096	0.0362	0.0460	0.0159	0.0416
8	0.0412	0.0099	0.0362	0.0459	0.0159	0.0417
9	0.0412	0.0099	0.0364	0.0458	0.0160	0.0419
μ_{sample}	0.0412	0.0097	0.0360	0.0458	0.0159	0.0418
μ_{H0}	0.0416	0.0097	0.0361	0.0455	0.0159	0.0425
t-statistic	-3.9332	0.4573	-0.5078	3.7425	0.7429	-13.6739
p-value	0.0034	0.6583	0.6238	0.0046	0.4765	0.0000

Table B.15: Raw Precision@1 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, enriched with ranked Wikipedia sentences. The sample mean μ_{sample} is compared against the null hypothesis μ_{H0}, the mean result obtained earlier when training against the Bible corpus only.

	en-es	en-fi	en-fr	es-en	fi-en	fr-en
0	0.0653	0.0246	0.0596	0.0670	0.0258	0.0612
1	0.0655	0.0246	0.0593	0.0666	0.0259	0.0615
2	0.0654	0.0247	0.0595	0.0667	0.0260	0.0614
3	0.0656	0.0247	0.0594	0.0671	0.0259	0.0612
4	0.0656	0.0249	0.0600	0.0670	0.0258	0.0614
5	0.0654	0.0248	0.0597	0.0673	0.0256	0.0613
6	0.0654	0.0245	0.0596	0.0673	0.0256	0.0614
7	0.0654	0.0247	0.0596	0.0671	0.0255	0.0615
8	0.0655	0.0248	0.0599	0.0673	0.0258	0.0614
9	0.0655	0.0248	0.0598	0.0671	0.0256	0.0612
μ_{sample}	0.0655	0.0247	0.0596	0.0671	0.0258	0.0614
μ_{H0}	0.0656	0.0247	0.0601	0.0675	0.0267	0.0622
t-statistic	-5.7864	0.4711	-7.0056	-6.1623	-17.3009	-24.7919
p-value	0.0003	0.6488	0.0001	0.0002	0.0000	0.0000

Table B.16: Raw Precision@10 results against Wiktionary lexical induction evaluation set, for 10 training instances of multilingual SID-SGNS trained on the Bible corpus, enriched with ranked Wikipedia sentences. The sample mean μ_{sample} is compared against the null hypothesis μ_{H0}, the mean result obtained earlier when training against the Bible corpus only.

l_s	w_s	l_t	w_t	$sim(w_s, w_t)$	l_s	w_s	l_t	w_t	$sim(w_s, w_t)$
fr	voiture [car]	en	bracket	0.170	en	car	fr	paléozoïque [paleozoic]	0.173
			couldn	0.161				ż [-]	0.168
			multidrug	0.160				disjonction [disjunction]	0.166
			follows	0.159				hennin [-]	0.165
			birthmark	0.158				magic [-]	0.164
			mannerism	0.156				tumultueuse [stormy]	0.162
			tundra	0.155				réflexive [reflexive]	0.161
			som	0.155				cessait [ceased]	0.158
			refusing	0.155				concentre [concentrated]	0.155
			montagu	0.153				patriotes [patriots]	0.154
es	coche [car]	en	sophia	0.182	en	car	es	torácica [thoracic]	0.185
			paprika	0.177				disparaban [they shot]	0.182
			stablished	0.171				dance [-]	0.165
			fast-acting	0.168				anticipando [anticipating]	0.165
			ayala	0.167				socialización [socialization]	0.158
			odysseus	0.163				duplicó [doubled]	0.157
			abelard	0.163				solicitaría [I would request]	0.155
			depresses	0.163				mg/dl [-]	0.154
			tubules	0.162				transponer [transpose]	0.150
			roar	0.160				mediterránea [Mediterranean]	0.149
fi	auto [car]	en	licensed	0.199	en	car	fi	annos [dose]	0.188
			persistence	0.180				kortit [cards]	0.177
			phosphoric	0.170				toinen [second]	0.173
			now-defunct	0.168				työpaikan [the workplace]	0.168
			enacts	0.164				suhtautuminen [attitude]	0.163
			pants	0.163				sitä [the]	0.155
			11–12	0.163				soutu [rowing]	0.154
			sankeertana	0.159				graalin [-]	0.153
			disclose	0.159				synestesia [synesthesia]	0.152
			axon	0.158				engels [-]	0.152

Table B.17: Multilingual SID-SGNS with Ranked Wikipedia Enrichment - translations of the OOV word *car*.