

Cross-lingual word embeddings for low-resource and morphologically-rich languages

by

Ali Hakimi Parizi

**Master in Computer Engineering, Razi University, 2015
Bachelor in Computer Engineering, University of Sistan and
Bluchestan, 2010**

**A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF**

Doctor of Philosophy

In the Graduate Academic Unit of Computer Science

Supervisor(s): Paul Cook, Ph.D., Department of Computer Science, UNB
Examining Board: Huajie Zhang, Ph.D., Department of Computer Science, UNB
Mike Fleming, Ph.D., Department of Computer Science, UNB
Wladyslaw Cichocki, Ph.D., Department of French, UNB
External Examiner: Diana Inkpen, Ph.D, Department of Computer Science, UOttawa

This dissertation is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

April, 2021

© Ali Hakimi Parizi, 2021

Abstract

Despite recent advances in natural language processing, there is still a gap in state-of-the-art methods to address problems related to low-resource and morphologically-rich languages. These methods are data-hungry, and due to the scarcity of training data for low-resource and morphologically-rich languages, developing NLP tools for them is a challenging task. Approaches for forming cross-lingual embeddings and transferring knowledge from a rich- to a low-resource language have emerged to overcome the lack of training data. Although in recent years we have seen major improvements in cross-lingual methods, these methods still have some limitations that have not been addressed properly. An important problem is the out-of-vocabulary word (OOV) problem, i.e., words that occur in a document being processed, but that the model did not observe during training. The OOV problem is more significant in the case of low-resource languages, since there is relatively little training data available for them, and also in the case of morphologically-rich languages, since it is very likely that we do not observe a considerable number of their word forms in the training data. Approaches to learning

sub-word embeddings have been proposed to address the OOV problem in monolingual models, but most prior work has not considered sub-word embeddings in cross-lingual models. The hypothesis of this thesis is that it is possible to leverage sub-word information to overcome the OOV problem in low-resource and morphologically-rich languages. This thesis presents a novel bilingual lexicon induction task to demonstrate the effectiveness of sub-word information in the cross-lingual space and how it can be employed to overcome the OOV problem. Moreover, this thesis presents a novel cross-lingual word representation method that incorporates sub-word information during the training process to learn a better cross-lingual shared space and also better represent OOVs in the shared space. This method is particularly suitable for low-resource scenarios and this claim is proven through a series of experiments on bilingual lexicon induction, monolingual word similarity, and a downstream task, document classification. More specifically, it is shown that this method is suitable for low-resource languages by conducting bilingual lexicon induction on twelve low-resource and morphologically-rich languages.

Dedication

To my father and mother
who did everything so I can be here

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance that I want to take a moment to thank them.

I would first like to thank my supervisor, Professor Paul Cook, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to acknowledge my lab mates for their wonderful collaboration. I want to thank you for your support and for all of the help I was given to further my research.

In addition, I would like to thank my family for their wise counsel and sympathetic ear. To accomplish this goal, I missed many events and was distracted all these years, but you did not stop loving and supporting me through all these years.

Table of Contents

Abstract	ii
Dedication	iv
Acknowledgments	v
Table of Contents	viii
List of Tables	xii
List of Figures	xiii
1 Introduction	1
1.1 Research Questions	6
1.2 Contributions	8
1.3 Thesis Structure	10
2 Related Work	12
2.1 Word Embeddings	12
2.2 Cross-Lingual Word Representations	20

2.2.1	Word Alignment	22
2.2.2	Sentence Alignment	30
2.2.3	Document Alignment	35
2.3	Low-Resource Languages	37
3	Evaluating Sub-word Embeddings in Cross-lingual Models	48
3.1	Corpora	50
3.2	Evaluation Datasets	52
3.3	Results	54
3.3.1	OOV Bilingual Lexicon Induction	56
3.3.2	Combined In-vocabulary and OOV Test Set	60
3.3.3	Interpolation with String Similarity	65
3.3.4	Low-resource Language Experiments	66
3.4	Summary	69
4	Sub-word Level Cross-lingual Word Representations	70
4.1	Joint Training with Pseudo-bilingual Corpora	72
4.2	Joint Training Incorporating Sub-word Information	75
4.3	Experiments	76
4.3.1	Intrinsic Evaluation	77
4.3.1.1	Bilingual Lexicon Induction	77
4.3.1.2	Monolingual Word Similarity	82
4.3.1.3	Summary	84

4.3.2	Extrinsic Evaluation	85
4.3.2.1	Resources	85
4.3.2.2	Zero-Shot Document Classification	87
4.3.2.3	Supervised Document Classification	89
4.3.2.4	Summary	90
4.3.3	Low-Resource and Morphologically- Rich Languages	90
4.3.3.1	Resources	92
4.3.3.2	BLI for In-Vocabulary Words	93
4.3.3.3	BLI for OOVs	94
4.3.3.4	Summary	100
5	Conclusion	101
5.1	Contributions	101
5.2	Research questions revisited	104
5.3	Future Work	105
	Bibliography	133
	Vita	

List of Tables

3.1	The size of the full corpus, and sample, for each language, in terms of the number of tokens, types, and resulting embeddings (Emb’s). Language families are also shown.	52
3.2	The number of translation pairs in each test set, for each language, with English as both the source and target language.	53
3.3	Precision@ N for bilingual lexicon induction for the dataset of translation pairs with OOV source language words. The method is indicated by “supervision+embeddings”, where supervision is Supervised, Semi-supervised or Unsupervised, and embeddings is FT for fastText or BPE for the approach of Zhu et al. (2019) using byte pair encoding. Results for the copy baseline are also shown. The best precision@ N , for each language and translation direction, is shown in boldface. . . .	57

3.4	Precision@ N for bilingual lexicon induction for the test set containing both in-vocabulary and out-of-vocabulary words. Results are shown for each language, with English as both the source and target language, for cross-lingual embeddings formed using each level of supervision. “Copy” refers to handling OOVs using the copy baseline, while “FT” indicates employing fastText sub-word embeddings, and “BPE” means using BPE sub-word embeddings to find translations for OOVs. The best precision@ N , for each language, translation direction, level of supervision and the subword embedding method, is shown in boldface.	62
3.5	Precision@ N incorporating edit distance for OOV source. The best precision@ N , for each language, translation direction and evaluation measure, is shown in boldface.	67
3.6	Precision@ N for Cherokee, using English as the target language, for each level of supervision. The best precision@ N is shown in boldface.	68
4.1	The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.	79

4.2	Precision@ N for bilingual lexicon induction. The best performance, for each dataset and evaluation measure, is shown in boldface.	80
4.3	Spearman’s correlation for monolingual similarity on each dataset, for each method considered. The best performance on each dataset is shown in boldface.	84
4.4	The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.	86
4.5	Accuracy on the MLDoc zero-shot cross-lingual document classification task, for each model and target language, with English as the source language. The average accuracy over all target languages is also shown.	88
4.6	Accuracy on the MLDoc supervised document classification task, for each model and language. The average accuracy over all languages is also shown. The highest accuracy in each case is shown in boldface.	90

4.7	The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.	92
4.8	Precision@ N for BLI for in-vocabulary words. The best precision for each dataset and evaluation measure is shown in boldface.	95
4.9	The number of translation pairs and unique pairs (i.e., the source language word is unique) in each test set, for each language.	96
4.10	Precision@ N for bilingual lexicon induction for the dataset of translation pairs with OOV source language words. The best precision@ N , for each language and methodology, is shown in boldface.	97

List of Figures

2.1	The word2vec architecture for CBOW and skipgram (Mikolov et al., 2013a)	15
2.2	A 2D projection of cross-lingual word embeddings for two languages, English and German (Luong et al., 2015)	21

Chapter 1

Introduction

Most of the natural language processing (NLP) tools and research are focusing on developing new NLP methods for a limited number of well-resourced languages, such as English, French, German, and Spanish. There are around 7000 languages in the world (Bird and Chiang, 2012) and according to Alyafeai et al. (2020) there are more than 310 languages in the world that have at least one million native speakers. However, there has not been enough research on these languages and not enough resources have been prepared to be utilized in developing NLP tools for these languages. Joshi et al. (2020) introduce six groups of languages based on the availability of their labelled and unlabelled data and they range from languages with almost none or very little available data, class 1 (e.g., Dahalo and Warlpiri), to well-resourced languages, class 6 (e.g., English and Spanish). Languages belonging to class 1 have more than 1 billion speakers around the world, but very little NLP research has been

dedicated to these languages.

Low-resource languages are languages that suffer a lack or absence of annotated data, and also do not have abundant unannotated data to be used for training NLP methods. In order to tackle the problem of low-resource languages, first, we should define and specify how low is considered as low-resource. This matter differs by changing the task and the language. To give an example, a part of speech tagger is a tool that assigns parts of speech (e.g., verb, noun, adjective) to each word in a sentence. It can reach a satisfactory performance by having access to a few thousand labelled tokens. However, another task like text summarization needs at least hundreds of thousands of labelled training instances (Alyafeai et al., 2020). It is also dependent on the language. For instance, a language that has a fixed structure, like English subject-verb-object, and does not have a complex morphological structure has a more modest data requirement than a freer word order language such as Persian, or a morphologically-rich language such as Inuktitut.

Morphemes are the smallest meaningful units of a language and cannot be split any further. We can divide morphemes into two classes, derivational and inflectional. Derivational morphemes are those that make a fundamental change in meaning if added to another morpheme, whereas inflectional morphemes are those which are used to indicate grammatical information. For example, if we consider the English word *cats*, this word contains two morphemes, a root *cat*, and an inflectional morpheme *s* that indicates that the word is plural. On the other hand if we consider the word *action*, this

word also has two morphemes, a root *act*, and a derivational morpheme *ion*, which create a new word, here a noun, out of the root.

We can also divide languages into two categories based on their use of morphology, analytic languages and synthetic languages. Analytic languages are those languages for which each word only has one morpheme. Purely analytical language, referred to as isolating languages, do not use affixes, and they instead use separate words to convey meanings and grammatical roles. An example of an analytic language is Mandarin Chinese. English is also an analytic language. Although English uses affixes and some words change form based on their grammatical role (e.g., the word *we* changes its form to *us* when it is used as an object), word order is very important and is used to show the function of a word. For example, in the sentence *the child is playing* the only indication that the *child* is the subject is its position in the sentence (Dawson et al., 2016). In contrast, in synthetic languages, words are made of several morphemes to convey a meaning or to show a particular function in a sentence. Synthetic languages are classified into three groups: agglutinative languages, fusional languages, and polysynthetic languages. In agglutinative languages, morphemes can be added to each other relatively loosely and it is easy to find the boundaries between morphemes, for example as in Hungarian. In fusional languages, we have a similar form to agglutinative languages, and morphemes are added together to form new meanings. However, in fusional languages, it is not possible to draw a line between morphemes and it is not clear where one morpheme ends and the other one starts. Spanish and Russian

are examples of fusional languages. In polysynthetic languages, it is possible to form very complex words by just adding many morphemes together. For instance in Inuktitut, which is a polysynthetic language, it is possible to convey the meaning of an English sentence by combining morphemes (Dawson et al., 2016). Below you can see two examples, one in English and the other one in Inuktitut.

1. English:

- *act - ion*
- *act - N*
- ‘action’

2. Inuktitut:

- *tusaa - tsiaq - junnaq - nngit - tu - alu -u - junga*
- *to hear - well - be able to - NEG - 3sg - very - be - 1sgg*
- ‘I cannot hear very well’

As the language gets more complex and shows a more complicated morphological structure, such as Inuktitut (Joanis et al., 2020), it requires more data so that conventional word-based models can capture and learn its structure. State-of-the-art approaches in NLP are data-hungry. Some of them require a substantial amount of human-annotated data (e.g, for part of speech tagging, dependency parsing) or they need a huge amount of unannotated text

containing millions of tokens to be trained (e.g, methods for learning word embeddings). This is problematic when working with low-resource languages due to the lack of training data. This can also cause a problem while working with morphologically-rich languages since many of their word forms do not appear in the training data. One way to solve this problem is to transfer knowledge from a rich-resource language to a low-resource one. Word embeddings, i.e., vector representations which can capture the semantics of words, have recently become a key feature in various NLP tasks such as named entity recognition (Pennington et al., 2014),¹ sentiment analysis (Schnabel et al., 2015)² and part of speech tagging (Al-Rfou' et al., 2013). So if we transfer the knowledge captured in embeddings of a rich-resource language to another low-resource language, developing NLP tools could become more feasible for low-resource languages. This idea has led to the development of new models, specifically cross-lingual word embeddings methods, to form word embeddings for two or more languages in a shared embeddings space (Mikolov et al., 2013b; Ruder et al., 2019). The goal of generating cross-lingual word embeddings is to provide a shared space to induce semantic word vectors in a multilingual context. Cross-lingual word embeddings can be used as a bridge to transfer knowledge between languages, especially from a rich-resource language to a low-resource one (Ruder et al., 2019). Although there has

¹This task tries to identify and locate named entities in a sentence, such as proper names, names of places, dates, etc.

²This task's goal is to recognize the sentiment of a sentence, i.e., whether the sentence has a positive or negative sentiment.

been much research carried out in this area to introduce advanced methods, which can generate high-quality cross-lingual embeddings, there are still some limitations that have not been addressed properly. When working with a low-resource language or a morphologically-rich language, it is very likely that we do not observe a considerable number of the language’s words in the training set. In the case of low-resource languages, the reason is the lack of training data, and in the case of morphologically-rich languages, the reason is the absence of many word forms in the training data. This problem is referred to as the out-of-vocabulary (OOV) problem. This is very important because in this case, we do not have an embedding for these words and the trained models do not know how to represent them. Since the number of OOVs for low-resource and morphologically-rich languages is high, the current models would show poor performance when employed for these languages.

1.1 Research Questions

In this thesis, I would like to address the absence of large training corpora and the existence of OOV words in low-resource and morphologically-rich languages and propose a method that is more reliable when applied to low-resource languages. Sub-word level embeddings (e.g., Bojanowski et al., 2017; Sennrich et al., 2016) — i.e., embeddings for units smaller than words, such as character sequences — have been proposed to address this limitation concerning OOV words for monolingual embedding models, but little prior

work — with the notable exception of Braune et al. (2018) which only considers low-frequency words and not OOVs — has considered sub-word embeddings in cross-lingual models. Therefore, understanding how we can get the most from a limited amount of data and finding a solution for OOV words by leveraging the sub-word information is critical. My goal is to address and find answers to the following questions.

1. Can we leverage sub-word embeddings in cross-lingual models to address the OOV problem in the cross-lingual domain? In this work, I evaluate whether the existing sub-word representation methods can be leveraged in the established cross-lingual word representation methods and whether the current methods are suitable to provide a representation for OOVs in a shared space. The goal here is to show that sub-word embeddings are able to provide a representation for an OOV word in the cross-lingual domain, as they can in monolingual settings.
2. Can cross-lingual word representations be improved by incorporating sub-word information in the process of training cross-lingual word representations and does incorporating sub-word information impact the performance in downstream tasks? In the previous question, I aim to verify my hypothesis that sub-word embeddings can be used in the cross-lingual domain, and here I argue that to get the most out of sub-word information, sub-word information should be incorporated in the training phase. In this way, not only do the sub-word embeddings

carry the monolingual information but also they obtain some cross-lingual knowledge. Therefore, we end up having embeddings with higher quality and also the ability to form representations for OOV words in a cross-lingual shared space.

3. Does the proposed method for learning cross-lingual embeddings by incorporating knowledge of sub-word information during training improve over benchmark approaches for learning cross-lingual embeddings on truly low-resource and morphologically-rich languages? In answering the previous question, the low-resource languages are simulated. However, as mentioned earlier, languages vary based on their morphological complexity, and it is crucial to show that a proposed method for low-resource and morphologically-rich languages actually works for these languages and it is not just limited to the case of simulated low-resource languages.

1.2 Contributions

In this thesis, first, I describe a step by step approach to justify the importance of sub-word embeddings in the cross-lingual domain and then introduce a method to employ this information more effectively. In the end, I provide results on low-resource languages as proof of my claim that incorporating sub-word information in the training phase improves the quality of cross-lingual word embeddings for low-resource languages. Below is a break down of the

contributions and their details.

- I evaluate and investigate whether sub-word embeddings can be leveraged in cross-lingual models. For this evaluation, I consider two widely used methods that provide sub-word information, character n-grams and byte-pair-encoding. I compare these two methods in a simulated low-resource scenario to find out which one of these methods is more suitable for the case of low-resource languages. In addition, I consider the case of a truly low resource language, Cherokee, to demonstrate the same trend, as in the simulated situations, exists in a non-simulated, actual low-resource language.
- I consider a novel bilingual lexicon induction task in which an in-vocabulary target language translation is found for an OOV source language word, where the representation of the source language word is constructed from sub-word embeddings.
- I propose a novel approach to employ sub-word information during the training phase to improve the performance of cross-lingual word representations. The proposed approach uses a modest size of monolingual training corpora and a bilingual dictionary as the cross-lingual signal.
- The proposed method is evaluated on two intrinsic tasks, bilingual lexicon induction and monolingual word similarity, and one extrinsic

task, document classification. The achieved results show that the proposed method performs on par with some strong benchmarks.

- The proposed method is evaluated on twelve truly low-resource and morphologically-rich languages. The evaluation is conducted by considering the task of bilingual lexicon induction for in-vocabulary words and also another bilingual lexicon induction task focusing on OOV words. The results show that the proposed method is indeed effective on the low-resource and morphologically-rich languages. In the case of in-vocabulary words, the proposed method performs on par with the current state-of-the-art method, and in the case of OOV words, it outperforms the same method by a great margin.

1.3 Thesis Structure

The rest of this thesis is organized as follows. In Chapter 2, I describe a history of word embedding methods and their advances and then discuss various cross-lingual embedding methods. Then at the end of Chapter 2, I describe low-resource languages and methods that leverage cross-lingual embeddings for down-stream tasks in low-resource languages. In chapter 3, I answer research question 1 by investigating the effect of leveraging sub-word embedding in cross-lingual models. More specifically, I compare two different methods of providing sub-word embeddings and various methods of forming cross-lingual word embeddings in the case of representing OOV words. This

chapter is an extended version of Hakimi Parizi and Cook (2020a). In order to answer research questions 2 and 3, Chapter 4 is dedicated to my proposed method to incorporate sub-word information in the training process. In this chapter, I also present the results of evaluating the proposed method in several cross-lingual and monolingual scenarios and also consider a down-stream task to demonstrate its performance in real-world applications. This part of Chapter 4 is an extended version of Hakimi Parizi and Cook (2020b). At the end of Chapter 4, I describe the experiment conducted on 12 low-resource and morphologically-rich languages. The last chapter, Chapter 5, presents a summary of my contributions and depicts a path for possible future work.

Chapter 2

Related Work

In this chapter I present an overview of the past and recent works on word embeddings and cross-lingual word embeddings, and discuss the importance of cross-lingual embeddings for low-resource and morphologically-rich languages. At the end of this chapter I also give a brief description of recent works on downstream tasks — document classification, part-of-speech (POS) tagging, and dependency parsing — targeting low-resource and morphologically-rich languages.

2.1 Word Embeddings

One of the main research areas in NLP is learning representations of words. One of the most well-known methods to form a semantic space is the vector space model (VSM) (Salton et al., 1975). The earliest models consider a

vector for a document with the dimensionality equal to the vocabulary size. Each cell of this vector is filled up by the weight of the corresponding word in the document. One way to calculate the weight is to calculate the word frequency in the document (Chowdhury, 2010). Aside from the usefulness of this method that provides a way to compare two documents or words together, this method has some drawbacks. The main one is the size of each vector, which is equal to the size of the vocabulary and it is mostly sparse. Therefore, other methods have been introduced to reduce the dimensionality of the vectors and make them denser. One of the earliest methods that I can point to is latent semantic analysis (LSA) (Deerwester et al., 1990). The main goal of LSA is to map from high dimensional count vectors to a low dimensional latent semantic space. It reduces the dimensionality by using singular value decomposition (SVD) to separate a matrix into its singular values and singular vectors.

To the best of our knowledge, the earliest attempt to form distributional representation for words, i.e., word embeddings, by neural networks is the method proposed by Bengio et al. (2003). They reduce the dimensionality of word vectors by utilizing a deep language model, a neural network language model with several layers, to jointly learn a vector for each word and also parameters of the probability function for the language model. Collobert and Weston (2008) employ a deep convolutional network to train word embeddings jointly with a language model. They also demonstrate the effectiveness of word embeddings in several downstream NLP tasks. However, these deep

architectures are very complex and computationally very expensive.

Mikolov et al. (2013a) propose word2vec, which has revolutionized the way of generating word embeddings and impacted almost all downstream NLP tasks. Word2vec contains two models to generate word embeddings, continuous bag-of-words (CBOW) and skip-gram. The architecture of these two models is shown in Figure 2.1. CBOW uses its surrounding words to predict the center word and skip-gram uses the center word to predict its surroundings. One other novelty of word2vec is the use of negative sampling. The intuition behind negative sampling is to select several words randomly from the vocabulary in the training corpus, and then the model tries to learn to distinguish words selected randomly and words that belong to the context window. So, negative sampling changes a multi-class classifier, predicting a target word among all the possible words in the vocabulary, to a binary classifier that only has to decide whether a word is from the training sample or the negative sample. The basic CBOW model computes word representations by using the probability distribution function defined below:

$$P(w_i|w_{i\pm k\setminus i}) = \frac{\exp(u_{w_i}^T h_i)}{\sum_{w \in W} \exp(u_w^T h_i)} \quad (2.1)$$

where h_i is the average of context vectors surrounding the target word, u_{w_i} is the vector representation of the target word, k is the window size, and W is all the words in the vocabulary of the training corpus. Calculating the denominator is a very expensive task, thus, Mikolov et al. (2013c) suggest

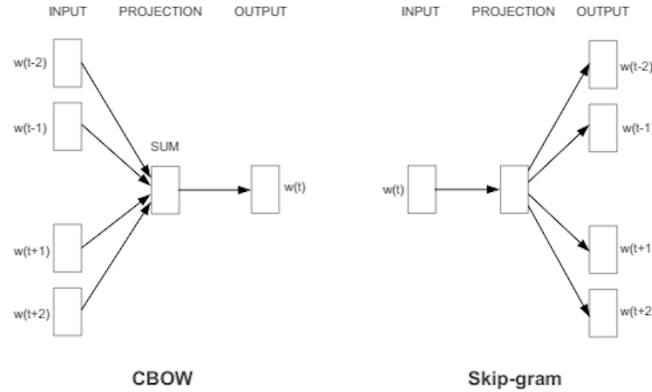


Figure 2.1: The word2vec architecture for CBOW and skipgram (Mikolov et al., 2013a)

instead of computing the Softmax function, to use negative sampling as shown in Equation 2.2.

$$O = \sum_{i \in D} (\log \sigma(u_{w_i}^T h_i) + \sum_{j=1}^p \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i)) \quad (2.2)$$

where D is the training data, p is the number of negative samples, and P_n is the noise distribution to draw negative samples.

Later Pennington et al. (2014) introduce a different method based on matrix factorization to learn word embeddings that employs the co-occurrence matrix of the whole corpus, in contrast to word2vec which uses just local contexts. They also use a weighted least square function in their objective function to give less weight to rare words. This approach is called GloVe.

These models that I have described so far consider a vector for each word and they do not consider morphology. As an example, in English, word2vec

only gives a vector for “*coming*” and it does not consider different parts of the words, for instance, it does not give a vector for “*ing*”. This is a bigger problem for morphologically-rich languages, such as Finnish, which is an agglutinative language, since plenty of words do not appear in the training corpus, hence we will not have a vector for these rare words. Below is an example in Finnish.

- *kahvi - n - juo - ja - lle - kin*
- *coffee - of - drink - -er - for - also*
- ‘also for coffee drinker’

Morphologically-rich languages encode much information at the word level as in the Inuktitut example in Chapter 1 (page 4). Their word forms are a function of their grammatical role, their relation to the surrounding words, pronominal clitics, and so on (Tsarfaty et al., 2013). For instance, in Arabic, which is a fusional language, most verbs are derived from three letter roots and they can be transferred into 15 different forms based on their grammatical role. For example, using the root *ktb* meaning ‘writing-related’, we can have *katab* meaning ‘to write’, *kaAtib* meaning ‘writer’, and *maktuwb* meaning ‘writing’ (Habash and Rambow, 2006). Ignoring the internal structure of words and their morphemes is also problematic in the case of low-resource languages since the available resources for these languages are very scarce and many of the word forms might not appear in the limited training data, and therefore there would not be vector representations for these missing words.

Sub-word level information is one way to solve this issue. Sub-word embeddings make the model more suitable for low-resource and morphologically-rich languages since they can construct an embedding for missing words to represent their semantics even though they have not been seen in the training corpus. Bojanowski et al. (2017) introduce a novel method, fastText, based on word2vec. Each word in the training corpus is augmented with special beginning and end of word markers. Each word is then represented as a bag of character sequences (i.e., sub-words), for example sequences of length 3–6 characters. They additionally include the entire word itself (with beginning and end of word markers) among the sub-words. In this model, a word is represented as the sum of its n -gram vectors. Therefore, Equation 2.2 is modified to Equation 2.3:

$$O = \sum_{i \in D} (\log S(w_i, c) + \sum_{j=1}^p \mathbb{E}_{w_j \sim P_n(w)} \log -S(w_j, c)) \quad (2.3)$$

where c is the context. S , shown in Equation 2.4, measures the similarity between a word and context, taking into account sub-words:

$$S(w, c) = \sum_{g \in G_w} z_g^T v_c \quad (2.4)$$

where G_w is the set of sub-words appearing in w , and z_g is the sub-word embedding for g . To calculate v_c , Bojanowski et al. sum representations for each word appearing in c , where each word is represented by the sum of its sub-word embeddings.

Byte Pair Encoding (BPE) is another method that can help provide sub-word knowledge. To work with BPE, we first need to define a vocabulary size beforehand. Then, the BPE model tries to extract the most frequent sequences of characters from the training corpus until it reaches the predefined vocabulary size. The vocabulary size could be as low as 26 for English, equal to the size of the alphabet considering only letters and not numbers, punctuations, etc., but it is normally defined as around several thousand (Sennrich et al., 2016). Zhu et al. (2019b) investigate different aspects for forming sub-word informed word representations such as different methods to break a word into segments, like morphemes or BPE, and various ways to compose these sub-word representations together to form word embeddings, e.g., averaging or a self-attention mechanism which is a learnable weighted addition. They find that the choice of the sub-word model is mostly dependent on the task, and that no one model is the solution for every situation and problem. They also demonstrate that BPE is not effective for forming a representation for OOVs.

One problem related to these classical word embedding methods is that they only assign one vector to each word and ignore the various meanings that a word can take by appearing in different contexts. This problem is very challenging when learning an embedding for words that have multiple meanings, i.e., polysemous words. For example, the word *bank* can mean “a financial institution” or it can mean “the land alongside of a river”. However, standard word representation methods only consider one vector for this word.

Contextual word embeddings go beyond a single semantic representation and generate an embedding for a word based on the surrounding context. ELMO (Peters et al., 2018) (Embeddings from Language MOdel) is one of the recent attempts to develop a language model that can generate context-dependent embeddings. Its architecture consists of L layers of forward LSTM (Long-Short Term Memory) and L layers of backward LSTM. As an input, we can feed it embeddings from classical word embeddings methods (e.g., word2vec or Glove) or we can use a character-level convolution neural network to incorporate sub-word information in the language model. At each layer, a representation for a word can be computed by concatenating the hidden states of the forward and backward LSTM in that layer. The final contextualized embedding is an aggregation of all the representations from each layer. After pre-training the language model, it can just simply serve as the first layer in supervised methods for other NLP tasks (e.g., document classification, named entity recognition). The input text is fed into the language model and the supervised method has the responsibility to learn parameters for the aggregation method to reach an embedding for each token. GPT (Radford et al., 2018) (Generative Pre-Training) and GPT2 (Radford et al., 2019) are two other examples of contextualized word representation methods. GPT is a deep language model that employs a multi-layer transformer decoder (Liu et al., 2018) to capture long-range dependencies in contrast to ELMO, which utilizes LSTM networks. Furthermore GPT uses BPE to encode sub-word information. Another feature of GPT is that it only captures left-to-right

dependencies and ignores the context which comes after a token. GPT2 architecture mostly follows GPT and it is only trained on larger corpora and has more parameters. BERT (Devlin et al., 2019) (Bidirectional Encoder Representation from Transformers) employs a different objective function and architecture that can consider all contexts surrounding a token when generating its representation. ELMO only uses a simple concatenation of forward and backward LSTM and GPT takes only left context into account. BERT is built on multi-layer bidirectional transformer encoder (Vaswani et al., 2017) and a masked language model (MLM) objective function, which incorporates both the left and the right context to predict masked tokens in a sentence. Furthermore, BERT has another objective function, next sentence prediction, to reach a better understanding of relations between sentences. Approaches to forming contextualized embeddings such as BERT have achieved substantial improvements in many NLP tasks such as named entity recognition (Hu and Verberne, 2020), document classification (Hoang and Vu, 2020), and question answering (Yang et al., 2019).

2.2 Cross-Lingual Word Representations

The word embedding methods that have been discussed are designed for a monolingual setting. However, the need for methods to represent words in a cross-lingual setting to be able to transfer knowledge between languages has emerged. This has led to development of new models to form cross-

I use the typology that is introduced by Ruder et al. (2019). They argued that the existing cross-lingual embedding methods optimize a similar objective function, and the only difference between them is the signal, the source of data, and its alignment type, that they use to generate embeddings. The signals are divided based on two features:

- Type of alignment;
- Comparability.

The first feature identifies the amount of supervision, strong or weak, and the second one categorizes the signals into two groups, parallel signals that are exact translations and comparable signals that are data which is similar in some way, such as the Wikipedia page of a concept but in two different languages.

Three types of alignment can be identified: word, sentence and document. I will explain each of them separately and describe some of the papers and methods proposed for each type.

2.2.1 Word Alignment

Most approaches introduced to form a cross-lingual representation employed parallel data in the form of bilingual dictionaries. These methods can be categorized into two main groups (Ruder et al., 2019):

- Mapping-based approaches;

- Joint models and pseudo-multilingual corpus-based approaches.

Methods using comparable signals use, for example, images as their signals. I do not discuss them here and instead try to describe the two aforementioned groups in more depth.

Mapping-Based Approaches: These methods assume we already have access to high-quality monolingual word embeddings for the source and target languages. Then, they try to find a mapping between the source vector space and the target vector space by using a bilingual dictionary. Mikolov et al. (2013b) show that there is a linear relationship between the vector spaces of two languages. If we consider the first language as A, and the second language as B, by solving Equation (2.5), we get a transformation matrix, W.

$$\Omega_W = \sum_{i=1}^n \|A_i W - B_i\|^2 \quad (2.5)$$

The transformation matrix maps the vectors of the language A to the vector space of the language B. Mikolov et al. (2013b) employ 5000 of the most frequent words in the source language and their translation in the target language, as the cross-lingual signal to find the transformation matrix by minimizing the mean squared error (MSE). Xing et al. (2015) argue there are inconsistencies between the objective function to learn word embeddings, the distance measurement for word vectors, and the objective function to learn the linear transformation. Thus, they propose to normalize all word vectors

to be unit length located on a hypersphere. They also suggest applying an orthogonal constraint on the transformation matrix. Faruqui and Dyer (2014) use two monolingual corpora to get word representations for each language, and then apply canonical correlation analysis to find a transformation matrix to map both languages to a shared space. Joulin et al. (2018) introduce a new loss function to find the mapping between the vector space of two languages. They use the squared loss to optimize the mapping matrix, W , and also consider an orthogonality constraint on matrix W . They propose to use a modified version of cross-domain similarity local scaling (CSLS) (Lample et al., 2018) to optimize and find the mapping function explicitly for the bilingual lexicon induction task.

Most of the prior methods assume that there is a high-quality bilingual signal, which the model requires in the form of a bilingual dictionary, readily available. Vulić and Korhonen (2016) argue that the bilingual seed lexicon plays a crucial role in the quality of the induced cross-lingual word embeddings. Therefore, to make sure that the bilingual seed lexicon contains high-quality translation pairs, they introduce a hybrid method that first finds a shared cross-lingual space by using an existing cross-lingual word embedding method — which does not require a bilingual seed lexicon for training — and extract a high-quality seed lexicon from this shared space to then use to train and form a second shared cross-lingual space using an existing mapping-based method. To ensure the quality of the seed lexicons, they consider a symmetric constraint and only accept those pairs that are mutual nearest neighbors.

Artetxe et al. (2017) propose a method that can work with a small seed lexicon, as low as 25 pairs. The main idea is to start from a small number of pairs and expand automatically. These methods can especially be useful in the case of dealing with a very low-resource language. They solve the same optimization problem as Mikolov et al. (2013b), and in a process of self-learning and in several rounds of bootstrapping add more translation pairs to the bilingual dictionary.

Recently, there have been a vast number of attempts to make the bilingual lexicon induction task feasible in an unsupervised fashion, i.e., without requiring any seed lexicon translation pairs for training. Hauer et al. (2017) introduce a method that works in two steps. In the first step, it extracts some seed pairs from non-parallel monolingual documents based on words' orthographic similarity and their frequency. In the second step, they expand their lexicon by employing a method similar to Mikolov et al. (2013b) with one difference: instead of only finding a transformation matrix from the source language to the target language, they also find a transformation matrix from the target language to the source language. Lample et al. (2018) argue that most of the current methods to form cross-lingual word representations require parallel data, some in the form of bilingual dictionaries and some in the form of parallel corpora, and that it is hard to acquire such data for all languages. Thus, they propose a fully unsupervised method that does not rely on any cross-lingual signal and is not dependent on character-level word similarities between languages. Their method consists of two main steps. The

first is an adversarial training strategy, which tries to find a mapping matrix between the embedding space of the source and target language. Then to increase the quality of the shared space, it finds the translation for the most frequent words and uses them as anchors, and in an iterative form tries to enhance the precision of the mapping matrix. In another work, Artetxe et al. (2018b) propose a fully unsupervised method that can learn the mapping between the vector spaces of two languages iteratively without the need for a bilingual signal. They employ the same self-learning method as Artetxe et al. (2017); however, they introduce a fully unsupervised initialization method that exploits the similarity distributions of words in the two languages to find a set of word pairs to start the learning phase.

Artetxe et al. (2020b) investigate the different scenarios that unsupervised methods take to find an ideal mapping between the embedding spaces of two languages. They argue that the assumption which these unsupervised methods make, i.e., having a substantial amount of monolingual data and lack of parallel data in any form (Lample et al., 2018; Artetxe et al., 2018b), is unrealistic because having access to some sort of parallel data in the form of bilingual dictionaries or parallel corpora is more common than having a huge amount of monolingual data, especially in the case of low-resource languages. Moreover, it has been shown that fully unsupervised methods do not perform well across all languages, especially in the case of morphologically-rich languages, and when the monolingual embeddings do not come from the same domain (Søgaard et al., 2018; Vulić et al., 2019). Ormazabal et al. (2019)

show that the isomorphism assumption — i.e., that embeddings for different languages have a similar geometric arrangement, which is key to the success of mapping-based models — does not always hold. Even though the mapping-based methods, especially unsupervised methods, are the current trend to form cross-lingual word representations, their experiments demonstrate that methods which jointly learn the embedding space for the source and the target languages are superior to mapping-based methods.

Joint Models and Pseudo-Bilingual Corpora These types of methods try to train embeddings for the source and the target language in a shared cross-lingual space. Some of these methods (e.g., Klementiev et al., 2012) exploit parallel corpora to train embeddings jointly, and others (e.g., Xiao and Guo, 2014; Gouws and Søgaard, 2015; Duong et al., 2016) relax the assumption of having parallel corpora and construct a pseudo-bilingual corpus from monolingual data and a bilingual dictionary and then train embeddings jointly.

Klementiev et al. (2012) introduce a method to induce cross-lingual representations by training a language model on the source and target languages and optimizing their objective function jointly. The process of optimization is such that every time a vector related to a word has to be updated, all its possible translations are also updated.

The goal of pseudo-bilingual corpus methods is to construct a bilingual corpus by simply replacing words in a source language corpus with their target

language translation randomly, instead of finding a transformation matrix between the source and target language using a bilingual dictionary. One of the earliest attempts to do so is Xiao and Guo (2014). First, they translate all the words in the source language corpus into the target language by using Wiktionary. Afterward, polysemous words and out of vocabulary words in the target language are removed. Finally, they employ the method of Collobert and Weston (2008) to form word representations by using sentences from both languages. They consider one vector for each translation pair, and in this way, they make sure that a word and its translation have similar vectors. Gouws and Søgaard (2015) concatenate and shuffle the source corpus and the target corpus. Then, they replace each word that is in the bilingual seed lexicon with one of its translations in a random manner, by flipping a coin (the chance of being replaced is the same as being left unchanged). To get the vector representation, they run CBOW on the constructed corpus. Similarly, Duong et al. (2016) also propose a method that replaces words in a pseudo-bilingual corpus with their translation during training. However, they further propose a way to handle polysemy by choosing the best translation for a word by considering its context using the expectation-maximization algorithm. Later, Adams et al. (2017) employ this method to train language models for low-resource languages.

Even though the methods discussed so far only consider one language as the source language and one as the target language, it is possible to train embeddings for several languages simultaneously. Duong et al. (2017) introduce

two ways to build multilingual word representations. First, they employ the method introduced by Duong et al. (2016) to train embeddings for several languages with a shared target language (e.g, English). Then, they use a simple mapping-based method (e.g., Mikolov et al., 2013b) to map the target embeddings of all these trained models to a shared space. Since the target language for all the trained models is English, there is no need for another bilingual seed lexicon for the mapping-based method. Then after finding the transformation matrix, they map the source languages’ embeddings to this shared space using the learned transformation matrix. They also propose a modification of the optimization function of Duong et al. (2016) to train embeddings jointly for several languages without the need for any post-processing steps and utilizing any mapping-based method. Ammar et al. (2016) propose an extension of Faruqui and Dyer (2014) to map embeddings of several languages to a shared space. They also introduce a new method, *Multicluster*, to learn multilingual embeddings for 50 languages. First, they form a cluster of words with similar meanings by using a bilingual dictionary. Then, they assign an ID to each cluster, and replace the words in the monolingual corpus with their assigned ID. In this way, they transform a corpus into a series of IDs, with each ID representing a cluster of words with similar meaning. Afterward, the monolingual corpora are concatenated to form a multilingual corpus. In the end, skip-gram is run over the concatenated corpus to learn an embedding for each ID.

2.2.2 Sentence Alignment

Another form of alignment is sentence alignment, which like word alignment, has two different degrees of alignment, parallel and comparable. This type of alignment is known as one of the most expensive ones. The parallel alignment can be categorized into five main groups (Ruder et al., 2019):

- Word-alignment based matrix factorization approaches
- Compositional sentence models
- Bilingual auto-encoder models
- Bilingual skip-gram models
- Contextual language models

Word-alignment based matrix factorization approaches: This group of methods utilizes information from alignment matrices in machine translation, which can be created from sentence-aligned parallel corpora using techniques such as *fast-align* (Dyer et al., 2013), as a signal to induce cross-lingual word representations. If a word in the target language is only aligned with one word in the source language, their representations are similar too. However, if a word is aligned with several words in the source language, its representation is a combination of those words. Zou et al. (2013) employ a neural network model and alignment matrices as the signal to form bilingual representations. Then, they try to optimize the objective function for inducing

word embeddings for each language jointly with the bilingual objective function. Guo et al. (2015) propose that the representation for a target language word is equal to the average of its related words in the source language. They also suggest a solution for out-of-vocabulary (OOV) words that are not in the parallel data. To form a representation for an OOV word, first, they find the nearest words to it based on edit distance. Then, the average of these words is selected as an embedding for the OOV word.

Compositional sentence models: Hermann and Blunsom (2013) argue that sometimes it is not enough to just have a representation for a word, and it is better to consider longer structures such as phrases or sentences. Therefore, they suggest a method that minimizes the distance between parallel sentences in a multilingual setting. They define the representation of a sentence as the sum of its constituents. Later, Hermann and Blunsom (2014) extend this idea by proposing a non-linearity function to form sentence representations over bigram pairs. They also argue that even having a representation for a sentence is not enough and that we need to have a representation for each document. Thus, they employed a recursive function to first form representations for sentences and then use these vectors at a higher level to form document representations.

Bilingual auto-encoder models: An auto-encoder is presented by Lauly et al. (2014) to form cross-lingual representations. The auto-encoder tries to accomplish four tasks:

1. Construct a sentence representation of the source language from its parallel sentence in the target language;
2. Construct a sentence representation of the target language from its parallel sentence in the source language;
3. Reconstruct the source language sentence representation from itself;
4. Reconstruct the target language representation from itself.

Similarly to Hermann and Blunsom (2013) they also define the representation for a sentence as the sum of its word embeddings.

Bilingual skip-gram models: Luong et al. (2015) propose a novel model to learn cross-lingual representations in a way to preserve both cross-lingual and monolingual features. To reach this goal, their model tries to learn word vectors by utilizing monolingual co-occurrences and jointly using the cross-lingual equivalences in parallel data. The proposed model employs the skip-gram model with negative sampling as the base method to learn word vectors. The skip-gram model learns the context of each word monolingually and also uses cross-lingual alignment to predict the context in the other language too. So, we can see their model as four skip-gram models such that two of them capture monolingual features, and the other two try to capture cross-lingual features. A similar method is also presented by Coulmance et al. (2015) that attempts to predict the context of a word in the source language and also in its aligned sentence in the target language. Pham et al.

(2015) introduce a method that forms a vector for each sentence, without just simply adding the vectors of its component words, by employing Paragraph Vector (Le and Mikolov, 2014). They consider one vector for each sentence in the source language and its aligned sentence in the target language to make similar sentences have similar representations.

Contextual Language Models As mentioned earlier, sentence alignment is a very expensive requirement, and having large parallel corpora, especially for low-resource languages, is either highly unlikely or it would be very expensive to build. In contrast, parallel data in the form of bilingual dictionaries are relatively-widely available. For example, Panlex (Baldwin et al., 2010) is a translation resource that combines many bilingual dictionaries and provides translations for 5700 languages. Furthermore, the recent trend of methods is to mostly exploit word alignment signals, and they are only capable of learning word-level embeddings. However, in some downstream tasks, it is crucial to have sentence representations instead of word representation (e.g. cross-lingual natural language inference and question answering). Thus, the current state-of-the-art methods, impacted by the advances in contextualized word representations and language modeling, tend to learn sentence representation either from parallel or non-parallel documents.

Conneau et al. (2018) introduce a new dataset to evaluate the performance of cross-lingual methods on the task of natural language inference (NLI). They also propose two cross-lingual baseline methods. The first one takes advantage

of the power of mapping-based methods. It first finds a shared space for two languages and then uses a sentence embedding method to learn sentence embeddings in a way that the sentence representation for a sentence in the target language becomes similar to its translation in the source language. The other method is a multilingual bidirectional LSTM (BiLSTM) sentence encoder which tries to learn to encode the source language sentences and their translations in the target language with an objective to locate a sentence and its translation nearby in the embedding space. Schwenk and Li (2018) also introduce a dataset to evaluate cross-lingual methods on the task of document classification. Along with the dataset, they also propose a baseline, which uses multiple sequence encoder-decoders to map the sentences of all the languages into a common shared space and it is trained on parallel corpora. Artetxe and Schwenk (2019) propose a bi-directional LSTM language model that is trained on a very large parallel corpus, containing 223 million parallel sentences, and jointly learns representations for 93 languages. More recently Conneau and Lample (2019) propose three cross-lingual language models with different objective functions. The first one is similar to common language models and tries to predict the next token based on the previous context. The second one is similar to the BERT (Devlin et al., 2019) objective function and learns to predict masked tokens in a sentence. The third one is an extension of masked language models and needs access to parallel corpora. It is similar to the BERT objective function, but instead of only predicting the masked token in the source language, it also tries to predict the masked token in its parallel

sentence in the target language. Their results suggest the combination of the masked objective function with the cross-lingual masked objective achieves the best results in cross-lingual downstream tasks. Multilingual BERT (mBERT) is a BERT language model trained on concatenated Wikipedia corpora for 105 languages. Wu and Dredze (2019) show that since mBERT uses a shared vocabulary for all languages, it can represent embeddings for all languages in a shared space, rather than representing each language in a separate space. This model is therefore able to learn deep contextualized cross-lingual word embeddings without any cross-lingual signal.

2.2.3 Document Alignment

As for word alignment and sentence alignment, for document alignment, we can divide the methods into two groups of parallel and comparable data. However, parallel documents are assumed to be sentence aligned too, so we can apply previous methods for sentence aligned data on them. Therefore, here I describe those methods, which utilize comparable documents. Comparable documents are also cheaper compared to parallel sentences. Methods presented in this subcategory can be categorized into two main groups (Ruder et al., 2019):

- Approaches based on pseudo-bilingual document-aligned corpora
- Concept-based methods

Approaches based on pseudo-bilingual document-aligned corpora

Vulić and Moens (2015) present a method to learn cross-lingual word vectors from non-parallel data. Their method first concatenates comparable documents together and after removing sentence boundaries, shuffles the concatenated document's words randomly so each word has a context of both languages. In the next step, it learns word embeddings by employing skip-gram. It is explained that if the window size is set to be big enough to contain context from both languages, it can lead to better bilingual representations.

Concept-based methods Søgaard et al. (2015) propose a novel method to build vector representations for languages based on Wikipedia topics. It is argued that Wikipedia contains a vast number of articles on the same topics written in different languages. Therefore, they propose to build representations for words based on the topics of the articles they are used in. In this way, words in different languages that are used to describe a similar topic are most likely similar. Therefore, they employ an inverted matrix to represent words based on the topics they are used in, instead of representing topics by their words. Last but not least, they utilize a dimensionality reduction technique (SVD) to reduce the dimension of vectors and sparsity.

2.3 Low-Resource Languages

Most of the tools presented for NLP tasks need strong supervision to perform adequately. To provide strong supervision, we need to have annotated text, which is very expensive and time-consuming to provide. Other methods, such as methods to form word representations, require a huge amount of unannotated text to effectively learn representations that encode word meanings. However, this type of data is also hard to obtain and for many languages impossible (i.e., languages that are not widely spoken and are endangered, like some Indigenous languages spoken in Canada). For these reasons, NLP tools only cover a limited set of languages, for example Google Translate only supports 104 languages. However, more than 200 languages were reported to be spoken just in Canada (Canada, 2018). Therefore, a vast amount of research has recently been done on low-resource languages to find a way to make current methods transferable to low-resource languages or develop approaches that are able to work with small amounts of data.

Before going any further and discussing the methods which make NLP tasks feasible for low-resource languages, it is essential to first discuss the features which make a language a low-resource language. If we want to define a low-resource language literally, it means a language that lacks resources for developing NLP tools. But, how scarce must the data be in order to call it low-resource, and what type of tools are we talking about? While some might assume a language is recognized as low-resource if it is on the verge of being

extinct, this is not the case and even the majority of European languages are recognized as low-resource, such as Lithuanian and Greek (Cieri et al., 2016). Strassel and Tracey (2016) consider a language as a low-resource one if it does not have any human language technology. Duong (2017) argues that the term human language technology is vague and there should be a more concrete definition to make it easier to identify a low-resource language. So they provide the following example:

“A language is considered low-resource for a given task if there is no algorithm using currently available data to automatically do the task with adequate performance.”

Most of the methods proposed for NLP tasks are for major languages such as English, German, Spanish, and French and these languages do not fall in the scope of the low-resource languages. However, since even accessing a small amount of training data for low-resource languages is sometimes difficult and it is even harder to find a proper benchmark to evaluate them, researchers often simulate the situation of a low-resource language by just decreasing the amount of training data for rich-resource languages, and then trying to observe how their proposed methods work in a simulated low-resource scenario by evaluating them on well-established benchmarks (e.g., Gu et al., 2018b; Zhu et al., 2019a; Mulcaire et al., 2019).

Various studies have been conducted on a variety of NLP tasks in a low-resource setting, such as machine translation (Gu et al., 2018a; Ramesh and Sankaranarayanan, 2018), sentiment analysis (Elming et al., 2014), language

modeling (Adams et al., 2017) and named entity recognition (Rahimi et al., 2019) to make them feasible for low-resource languages. Here, I describe some of the work that has recently been conducted on document classification, part-of-speech (POS) tagging, and dependency parsing. The first task, document classification, is the one that I consider in Chapter 4 to evaluate my proposed method. The other two tasks, POS tagging and dependency parsing are selected since they are widely addressed in the literature to evaluate cross-lingual word representations and their impact on downstream applications for low-resource languages.

Document Classification: One of the widely used extrinsic tests to evaluate cross-lingual models is document classification, i.e., classifying a document as one of a predefined set of classes. This task is motivated by the situation when sufficient labeled training data is not available for a low-resource language, however, there is a vast amount of training data available in a rich-resource language. A classifier is therefore trained on a rich-resource source language, such as English, and then directly applied to a low-resource target language. There are several datasets to perform cross-lingual document classification, but the one which is most widely used consists of news articles from the RCV1/ RCV2 dataset (Lewis et al., 2004). In this dataset, there are four topics: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). Each document can belong to one or several of these topics; however, in previous works the classification

focuses mostly on documents which belong to only one topic.

One of the early works on cross-lingual document classification is Klementiev et al. (2012). They only consider English and German for their experiments. First, they include a subset of RCV1/RCV2 in the training data so their model gains some domain knowledge. To train a classifier, a document is represented as the average of its word representations weighted by their *idf* score. *Idf*, inverse document frequency, shows how important a word is by assigning a low score to frequent words — e.g., *the*, *a*, *an* — and a high score to the words which are rare. An averaged perceptron algorithm is used as the classifier. Duong et al. (2016) represent documents as a bag of words weighted by *tf-idf* score. They train a classifier on 1000 documents in the source language, English, and applied it directly to 5000 documents in the target language. Similar to Klementiev et al. (2012) they only conduct experiments on English and German.

Schwenk and Li (2018) argue that selecting a subset of RCV1/RCV2 randomly would cause some difficulties that might not be desirable while evaluating a cross-lingual model. Specifically, a random subset would lead to imbalanced class distributions in our training, development, and test sets. Even though in reality we do not face a perfectly balanced class distribution, simulating balanced classes would make evaluating a cross-lingual model easier. Therefore, they introduce a subset of the RCV1/ RCV2 dataset (referred to as MLDoc) for 8 languages, Chinese, English, French, German, Italian, Japanese, Russian, and Spanish. It has 1000 documents in each of the training and development

sets, and 4000 documents in the test set, for each language. This dataset not only makes the class distribution balanced and provides more languages for the experiments but also provides a benchmark to make comparing results achieved from various methods easier. Artetxe and Schwenk (2019) evaluate a deep language model trained on 220M parallel sentences for 93 languages on MLDoc. They generate a representation for each document by passing it through the network and then feeding the representation to a feed-forward classifier. Devlin et al. (2019) and Conneau and Lample (2019) propose contextualized language models that are trained on a substantial amount of multilingual text. It is possible to just add a classifier to the last layer of these language models and perform zero-shot document classification (Wu and Dredze, 2019; Lai et al., 2019). Zero-shot document classification means to train a classifier on the source language and then apply it directly on the target language. The state-of-the-art result for this task on the MLDoc dataset belongs to Lai et al. They argue that apart from the cross-lingual gap — i.e., aligning the vector space of two languages — there is also a domain gap between languages. They argue that NLP tasks are not identical in different languages, for example, for sentiment analysis of customer reviews, people may express their feelings of a product differently in different languages. So, they propose to employ XLM (Conneau and Lample, 2019) along with an unsupervised data augmentation approach, i.e., fine tuning XLM on unlabeled data in the target language or generating noisy labeled data in an unsupervised fashion and training the classifier on these noisy data too in order to make

the classifier invariant to a small amount of noise, to fill the gap between domains and languages. These language models (Artetxe and Schwenk, 2019; Devlin et al., 2019; Conneau and Lample, 2019) are trained on a substantial amount of text data and in the case of Artetxe and Schwenk, and Conneau and Lample, they also employ parallel corpora which is the most expensive resource for building cross-lingual methods.

POS Tagging: This task aims to label each word in a sentence with its corresponding part of speech tag, i.e., noun, verb, adjective, etc. One line of research tries to take advantage of available parallel corpora. Duong et al. (2013) propose an unsupervised POS tagger that exploits parallel corpora between a rich-resource language and a low-resource language. Their method first uses the word alignment between the two corpora and tags the target corpus partially with the same tag as the alignment word in the source language. After this step, the top sentences in the target language are selected based on their alignment score to be used in a self-training and revision process. Agić et al. (2015) use just Bible translations to train a POS tagger in different languages. They argue that the translation of the Bible, or a part of it, is available for a large number of languages. For a subset of these languages, K , there is enough training data available to train a POS tagger. After tagging these languages, for the rest of the languages, they suggest using the majority vote between the K languages to create a tag dictionary to train a POS tagger for each other language. Kim et al. (2015) also leverage

the availability of the partial translation of the Bible. They assume that for each language they have $200K$ tokens and also that they have access to a reliable tagger for some of the rich-resource languages. Although they do not utilize any ancillary source of information, they employ a two-step canonical correlation analysis (CCA) method to form word representations that not only contain information about the word context and its projected tags from rich-resource languages, but that also is more general by incorporating a monolingual corpus from newswire for each language. In the end, they train a multi-class SVM to assign a tag to each word in a sentence.

Another line of research investigates the impact of providing a small amount of gold standard training data in the low-resource language and training a tagger using this data. Garrette et al. (2013) investigate the results of crafting training data for POS tagging for two truly low resource languages, Kinyarwanda and Malagasy, in a limited time. Their results demonstrate type annotation — i.e., annotating words out of context — is superior to token annotation — i.e., annotating word instances in context. By having a small set of type annotations provided by a linguist in a limited time, they achieve satisfactory results. Duong et al. (2014) try to develop a more accurate POS tagger by proposing a semi-supervised method. As in Duong et al. (2013), they use parallel data to transfer tags from a rich-resource language to a low-resource one. However, in this work, they suggest that incorporating the knowledge of a POS tagger, which has been trained only on 1000 tokens of annotated data on the low-resource language, can increase the

accuracy of the POS tagger. Fang and Cohn (2017) argue that we may not have access to readily available parallel data for many languages. Thus, they propose another method that employs cross-lingual word embeddings. Their method first learns cross-lingual embeddings for a rich-resource language and a low-resource one. Then, they train a BiLSTM tagger on the rich-resource language. The first layer of the network is the cross-lingual embeddings, and this shared cross-lingual space makes knowledge transfer possible. To improve the accuracy of the tagger, they employ gold standard annotations on the target language and train the tagger jointly using distant supervision, the rich-resource language, and the gold annotations for the low-resource language. Kim et al. (2017) introduce a method that can achieve high accuracy in low-resource scenarios by using only 1280 tagged sentences similar to fully-supervised methods on rich-resource languages. In the first layer, it has a BiLSTM model that provides character level embeddings. The concatenation of this network with word-level embeddings is fed to two other BiLSTM networks, a private BiLSTM, and a common BiLSTM. The common BiLSTM uses a language classifier objective function to provide language-agnostic sentence representations and it is shared between languages. However, the parameters for the private BiLSTM are not shared between languages. It is trained by concatenating its output with the common BLSTM and feeding it to a BiLSTM language model to capture the characteristics of each language and then to a softmax layer to predict part-of-speech tags.

With the recent advances in contextualized embeddings, and the introduction

of mBERT, several studies have applied it in downstream tasks including POS tagging. Wu and Dredze (2019) and Pires et al. (2019) employ mBERT in POS tagging in zero-shot and low-resource scenarios. Wu and Dredze compare their results with Kim et al. (2017) and even though mBERT outperforms Kim et al.’s method in zero-shot scenarios, it falls behind it in the low-resource scenarios, where a small amount of labeled data is available for the target language. However, assuming that we do not have access to any training data is a very strong assumption. As mentioned, Garrette et al. (2013) show that in only four hours, we can gather enough annotated data, 10K words, to train a cross-lingual tagger for a low-resource language, thus we can avoid zero-shot situations.

Dependency Parsing: The task of analyzing the grammatical structure of a sentence by identifying the syntactic relations between words is called dependency parsing. Designing a dependency parser also requires a large amount of training data, for example a treebank, which is available for only a limited number of languages. Thus, to be able to have a dependency parser for low-resource languages, approaches to transfer knowledge from a rich resource language to a low-resource one have been developed. A series of works have used the Universal Dependency Treebank (Nivre et al., 2020) to make this transfer possible. Duong et al. (2015b) propose a novel method to train a dependency parser jointly on a rich-resource language and a low-resource language, which only has a small amount of annotated data. By training a

neural network jointly on both languages, they create a shared cross-lingual space that enables knowledge transfer between languages. They also consider an embeddings matrix for each language in the objective function to preserve specific features related to each language. Another novelty of their work is that their model is able to incorporate a bilingual dictionary in the training process to make the representation of a word similar to its translation.

Agić (2017) introduces a model to select a delexicalized parser among several source languages for a target language. Three measures are introduced to select the best parser at runtime. The first measure considers the distribution of source and target tri-gram POS tags, the second one is a naive Bayes classifier using n -grams, and the last one gets help from resources that contain structured data for a large number of languages — i.e., WALS covers 2679 languages and for each language it has 202 features such as morphology, phonology and syntax (Dryer and Haspelmath, 2013).

Another way to transfer knowledge between languages is by using parallel data. Schlichtkrull and Søgaard (2017) suggest using Bible translations as a source to transfer knowledge between languages. First, they parse the Bible translation in the source (high-resource) language using a pre-trained graph-based parser. Then, by projecting labels to the low-resource language, a scoring edge matrix is created. Now, a parser can be trained for the low-resource language using this edge matrix. Contextualized embeddings are also utilized in building cross-lingual dependency parsers. Wu and Dredze (2019) use a graph-based model and the output of mBERT is fed to the graph

as its input. The results demonstrate the advantage of mBERT in capturing cross-lingual characteristics to build a cross-lingual dependency parser.

Chapter 3

Evaluating Sub-word Embeddings in Cross-lingual Models

Cross-lingual word embeddings provide a shared space for embeddings in two languages, enabling knowledge to be transferred between them. Cross-lingual word embeddings can be used for tasks such as bilingual lexicon induction, and can be leveraged to improve systems for natural language processing (NLP) for low-resource languages for tasks such as language modelling (Adams et al., 2017), part-of-speech tagging (Fang and Cohn, 2017), and dependency parsing (Duong et al., 2015a). In the case of out-of-vocabulary (OOV) words, however, no information is available. This could be particularly problematic for low-resource languages, where the number of words that embeddings are

learned for could be relatively low due to the relatively small amount of training data available, and for morphologically-rich languages, where many wordforms would not be observed while learning the embeddings.

This chapter, which is an extended version of Hakimi Parizi and Cook (2020a), presents a systematic evaluation of whether sub-word embeddings can be leveraged in cross-lingual models to address the OOV problem. Specifically, in Section 3.1, I first describe the training corpora used to train the cross-lingual embeddings and then in Section 3.2 I show the details of the evaluation data for the experiments, which focus on bilingual lexicon induction. Specifically, I propose a novel bilingual lexicon induction evaluation which focuses on finding an in-vocabulary target language translation for an OOV source language word. Section 3.3 is dedicated to presenting the results. More specifically, Section 3.3.1 shows the results on an OOV test-set by using three cross-lingual word embedding methods and employing two methods for splitting words into subwords. Section 3.3.2 presents the results achieved on a test set consisting of a combination of OOV and in-vocabulary words. Afterwards, in Section 3.3.3, I investigate the impact of interpolating string-based edit distance with the similarity score from cross-lingual embeddings, focusing again on OOV words. Section 3.3.4 considers the case of a truly low-resource language, Cherokee, and investigates the importance of sub-word embeddings for low-resource languages. In Section 3.4, I summarize and describe the contributions of this chapter.

3.1 Corpora

In these experiments, I consider the same languages that Adams et al. (2017) used in their paper on applying cross-lingual word embeddings for low-resource language modelling; specifically, I consider English, and the following languages which have varying degrees of similarity to English, and vary with respect to morphological complexity: Finnish, German, Japanese, Russian, and Spanish. The corpus for each language is a Wikipedia dump from 20 September 2018, except for Japanese, where I use a pre-processed Wikipedia dump (Al-Rfou’ et al., 2013). For English, the raw dump is preprocessed using wikifi (Bojanowski et al., 2017), and for the other non-Japanese languages I use WP2TXT.¹ Details of the corpora for each language are provided in Table 3.1 in the “Full” columns.

In preliminary experiments I observed that for the full Wikipedia corpora, relatively few words in the evaluation dataset (discussed in Section 3.2) were OOVs, yet OOVs are required for my experimental setup, since the goal is to observe the transferability of words that are not seen in the training corpus of the source language but are available in the embedding matrix of the target language. Therefore, following Adams et al. (2017), I carried out experiments in which I learned cross-lingual embeddings, but down-sized the size of the source language corpora, which increases the number of OOVs.

I conduct a series of preliminary experiments on the quality of cross-lingual

¹<https://github.com/yohasebe/wp2txt>

word representations and their dependency on the size of the target language training corpus. The goal of these preliminary experiments is to find the least amount of training data required to generate reasonably high-quality cross-lingual word representations. In these experiments, I only consider language pairs where English is the source or target language, and the other language is one of the five other languages (i.e., one of Finnish, German, Japanese, Russian, or Spanish). I begin by downsizing the target language training corpus into different size sub-corpora with varying numbers of tokens (e.g., 100K, 1M, 10M, 100M). These sub-corpora are used to train the target language word embeddings, and then these embeddings are employed in a bilingual lexicon induction task for in-vocabulary words. To learn the cross-lingual word embeddings, I employed the public implementation of a supervised method (Lample et al., 2018). My findings indicate that bilingual lexicon induction for in-vocabulary items performed with reasonably high accuracy (e.g., $> 60\%$) down to corpora of roughly 100M tokens. I therefore choose a randomly-selected 100M token portion of each corpus as a sample, except for Finnish, where the full corpus is less than 100M tokens. Details of these corpora are also shown in Table 4.1, in the “Sample” columns.

In the bilingual lexicon induction experiments in this chapter, I attempt to find an in-vocabulary target language translation for an OOV source language word. I therefore always use the full corpus for the target language — so that translations of many source language words will be in-vocabulary in the target language — and a sample for the source language — so that a

Language	Family	Full (target)			Sample (source)		
		Tokens	Types	Emb's	Tokens	Types	Emb's
English	Germanic	4500M	9.9M	2470k	100M	800k	210k
Finnish	Finnic	70M	3.8M	650k	70M	3800k	650k
German	Germanic	690M	10.2M	2030k	100M	2900k	550k
Japanese	Japanese	200M	2.2M	370k	100M	1100k	230k
Russian	Slavic	390M	8.7M	1550k	100M	3700k	650k
Spanish	Romance	500M	4.3M	810k	100M	1500k	310k

Table 3.1: The size of the full corpus, and sample, for each language, in terms of the number of tokens, types, and resulting embeddings (Emb's). Language families are also shown.

substantial number of gold-standard translations will be OOV in the source language, and to simulate a lower-resource source language.

3.2 Evaluation Datasets

Panlex (Baldwin et al., 2010) is a freely-available translation resource, built by combining many translation dictionaries, that covers thousands of languages and includes over 1 billion translations. I use Panlex to build gold-standard evaluation data.

In these experiments I only consider language pairs where English is the source or target language, and the other language is one of the five other languages (i.e., one of Finnish, German, Japanese, Russian, or Spanish). I begin by extracting all single-word translations from Panlex for these language pairs. For each language pair, I then create a gold-standard evaluation dataset by keeping only those translations for which the source language word is not in

Language	# of pairs	
	English source	English target
Finnish	13722	10723
German	23891	32473
Japanese	11100	50000
Russian	18299	72648
Spanish	15359	22433

Table 3.2: The number of translation pairs in each test set, for each language, with English as both the source and target language.

the embedding matrix for the source language corpus (i.e., OOV in the source language), and the target language word is in the embedding matrix for the target language (i.e., in-vocabulary for the target language). I observed that some translations in Panlex appear to be noisy. For example, in the case of the English - Spanish dictionary, some English entries consist of no Latin letters and appear to be non-English words (e.g., there is an entry for 炙る), while other entries consist entirely of non-alphabetic symbols (e.g., there is an entry for —'Auni). I therefore further eliminated any translation for which the source language word does not appear in the Aspell dictionary (version 0.6) for that language.² In this way I made sure that all remaining OOVs are valid and meaningful words. Details of the evaluation datasets are shown in Table 3.2.

²<http://aspell.net/>

3.3 Results

In this work, I use three approaches, a supervised, a semi-supervised and an unsupervised method, to learn cross-lingual embeddings in order to make a comprehensive comparison between methods with various degrees of supervision. For the supervised method I use a publicly-available implementation of Lample et al. (2018).³ For the semi-supervised and fully-unsupervised methods, I use publicly available implementations of the approaches of Artetxe et al. (2017) and Artetxe et al. (2018b), respectively.⁴ For all three models, I use their default settings to learn the transformation matrix, such as number of iterations and maximum vocabulary size. I stick to their default values, since these numbers are the optimal numbers used by their authors to keep a balance between performance and speed.

In order to evaluate the effectiveness of sub-word embeddings, two different approaches to forming word embeddings by using sub-word information are considered. For the first approach, I use fastText (Bojanowski et al., 2017) — discussed in Section 2.1 — to create an embedding matrix for each corpus. fastText is a method to form word embeddings based on word2vec (Mikolov et al., 2013a) with a noticeable difference that it uses sub-word information in the form of character n -grams to train word embeddings. I use the default fastText parameters, except for the number of dimensions for the embeddings, which is set to 300, since it is the embedding size for fastText pre-trained

³<https://github.com/facebookresearch/MUSE>

⁴<https://github.com/artetxem/vecmap>

embeddings.⁵

For the second approach, I use the method introduced by Zhu et al. (2019b), which provides a framework to investigate two components of forming sub-word informed word representations — segmentation of words into their sub-words, and the effect of different sub-word composition functions. I use byte pair encoding (BPE) (Sennrich et al., 2016) as the method which provides sub-word information. I select BPE, since it does not need any additional tools to be employed for various languages in contrast to approaches which require language specific tools such as a morpheme analyzer, and it also is widely used in many other NLP applications (e.g., machine translation (Sennrich et al., 2016) and language mode (Radford et al., 2019)). In contrast to fastText, which breaks a word into its character n -grams, BPE breaks a word into its most frequent consecutive sequences of characters. To train word embeddings using the Zhu et al. (2019b) framework, I use the default settings, which use addition as the composition function — similar to fastText — and do not include an embedding for the whole word itself in the composition — in contrast to fastText, which does include a representation for the whole word along with representations for its sub-words. I refer to this approach — which is based on Zhu et al. (2019b) and incorporates BPE — as BPE.

Results are presented in the following subsections. In subsection 3.3.1, the results for bilingual lexicon induction for OOVs are presented using the various approaches to representing OOVs and learning the transformation

⁵<https://fasttext.cc/docs/en/pretrained-vectors.html>

matrix. Subsection 3.3.2 describes a further experiment, in which the test data consists of both in-vocabulary and OOV source language words (as opposed to only OOVs). In subsection 3.3.3 I discuss incorporating information from edit distance, along with information from cross-lingual word embeddings, to find the best translation for OOVs. The last subsection presents results for bilingual lexicon induction for OOVs in Cherokee, a truly low-resource source language.

3.3.1 OOV Bilingual Lexicon Induction

In the case of the supervised method, given source and target language embeddings, a set of translations is required to learn the transformation matrix W in Equation 2.5 (page 23). Following previous work (e.g. Lample et al., 2018; Joulin et al., 2018; Jawanpuria et al., 2019), I use the MUSE training pairs provided by Lample et al. (2018), which include 5000 unique source words.⁶ For training the semi-supervised method, I take a random sample of 25 pairs from these training pairs. Given a gold-standard evaluation pair, I construct a representation for its (OOV) source language word by averaging its sub-word embeddings using fastText or BPE. I then transform this representation using W , and rank the target language words by the cosine similarity of their embeddings with this transformed representation of the source word. I report precision@ N — for $N = 1, 5$, and 10 — where the system is scored as correct if the gold-standard target word is amongst the

⁶<https://github.com/facebookresearch/MUSE>

Language	Method	Precision					
		English source			English target		
		@1	@5	@10	@1	@5	@10
Finnish	Supervised+FT	1.49	3.55	4.97	2.43	5.67	7.74
	Semi-supervised+FT	1.01	3.33	4.27	1.35	3.74	4.95
	Unsupervised+FT	1.10	3.24	4.25	1.17	3.65	4.68
	Supervised+BPE	0.22	0.64	0.94	0.50	1.13	1.35
	Semi-supervised+BPE	0.15	0.70	1.08	0.45	0.77	1.40
	Unsupervised+BPE	0.00	0.02	0.02	0.36	0.90	1.13
	Copy baseline	0.46	-	-	0.27	-	-
German	Supervised+FT	2.35	5.60	7.35	3.16	8.07	10.77
	Semi-supervised+FT	2.37	5.01	6.57	2.15	6.16	8.16
	Unsupervised+FT	2.32	5.15	6.42	2.01	5.80	8.06
	Supervised+BPE	0.25	0.63	0.99	0.25	0.73	1.15
	Semi-supervised+BPE	0.18	0.60	0.92	0.26	0.73	1.14
	Unsupervised+BPE	0.23	0.63	0.93	0	0	0.04
	Copy baseline	2.06	-	-	0.81	-	-
Japanese	Supervised+FT	0.45	1.61	2.17	0.67	1.73	2.33
	Semi-supervised+FT	0.95	2.62	3.65	0.36	1.07	1.47
	Unsupervised+FT	0.85	2.54	3.73	0.33	1.05	1.42
	Supervised+BPE	0.24	0.77	1.01	0.03	0.19	0.25
	Semi-supervised+BPE	0.21	0.63	0.98	0	0	0.01
	Unsupervised+BPE	0	0	0	0.04	0.13	0.21
	Copy baseline	0.13	-	-	0.73	-	-
Russian	Supervised+FT	2.11	5.14	6.85	3.86	9.19	12.07
	Semi-supervised+FT	1.32	3.49	4.74	2.45	6.21	8.35
	Unsupervised+FT	1.19	3.45	4.72	2.69	6.69	8.68
	Supervised+BPE	0.16	0.47	0.77	0.41	1.19	1.79
	Semi-supervised+BPE	0.17	0.52	0.84	0.36	1.18	1.75
	Unsupervised+BPE	0	0	0	0.45	1.12	1.69
	Copy baseline	0.09	-	-	0	-	-
Spanish	Supervised+FT	6.09	10.99	13.43	3.69	8.20	10.68
	Semi-supervised+FT	5.62	9.85	12.15	3.28	7.26	9.36
	Unsupervised+FT	5.63	9.86	12.23	2.93	6.98	9.23
	Supervised+BPE	0.63	2.12	2.81	0.30	0.82	1.11
	Semi-supervised+BPE	0.83	1.89	2.76	0.26	0.85	1.17
	Unsupervised+BPE	0.81	1.84	2.66	0.26	0.83	1.09
	Copy baseline	3.56	-	-	2.34	-	-

Table 3.3: Precision@ N for bilingual lexicon induction for the dataset of translation pairs with OOV source language words. The method is indicated by “supervision+embeddings”, where supervision is Supervised, Semi-supervised or Unsupervised, and embeddings is FT for fastText or BPE for the approach of Zhu et al. (2019) using byte pair encoding. Results for the copy baseline are also shown. The best precision@ N , for each language and translation direction, is shown in boldface.

top- N most similar target language words.

I compare against two baselines. I consider a random baseline, which randomly ranks the target language words for a given source language word. I also consider a second baseline motivated by a simple approach to handling OOVs in machine translation, in which the OOV source language word is copied into the target language. This approach could work well, particularly for some named entities and borrowings (Sennrich et al., 2016). I refer to this approach as the copy baseline. Note that the copy baseline only provides one target language translation for a given source language word, and as such, I only calculate precision@1 for this method.⁷

Results are shown in Table 3.3. For all languages and translation directions, when the source of sub-word information is fastText, precision@1 is higher using the supervised, semi-supervised and unsupervised methods than using the copy baseline, except for precision@1 in the case of Japanese with English as the target language.⁸ This inconsistent result for Japanese appears to be due to differences in the test data when English is the source, as opposed to target, language. When English is the source language, and Japanese is the target language, there are only 5 pairs in the test data where the source and target words are identical, i.e., cases where the copy baseline is correct. On

⁷Razmara et al. (2013) propose an approach to finding translations for OOVs based on graph propagation. Their method requires a phrase table derived from a parallel corpus. In contrast, the methods for bilingual lexicon induction considered in this chapter do not require a parallel corpus. Because of the substantially higher resource requirements of the method of Razmara et al. I do not compare against this approach.

⁸For all language pairs, and each value of N considered, precision@ N is 0% for the random baseline (results not shown).

the other hand, in the case of English being the target language, and Japanese being the source language, there are 270 pairs where the source and target words are identical. Overall these findings indicate that, for most languages considered, and any level of supervision, fastText embeddings outperform the copy baseline.

The results also indicate that fastText outperforms BPE for this task. In all cases, when comparing results for the same language, translation direction, level of supervision, and precision@ N , with the only difference being the source of sub-word information, fastText always outperforms BPE. Zhu et al. (2019b) noted that BPE is not effective for dealing with OOVs, and I observe the same here.

Focusing on approaches using fastText, and considering the differing levels of supervision, I observe that the supervised approach often performs better than the semi-supervised or unsupervised approaches, and indeed this is always the case for Finnish, Russian, and Spanish, but there are some exceptions for German and Japanese. I return to consider the level of supervision in Section 3.3.4 when I consider the case of a truly low-resource language.

I further observe that for each language, the accuracy is higher when English is used as the target language, than when English is used as the source language, except for the case of Spanish. Note that English has the largest corpus among the selected languages, and that I always use the full corpus for the target language, but a sample for the source language. Therefore, I expect the embeddings for English as the target language to be higher quality

than those for English as the source language, which could explain why the accuracy is higher when English is used as the target language than as the source language. The inconsistency of this finding in the case of Spanish could be due to the fact that the copy baseline has the highest accuracy in the case of Spanish as the target language. I also observe that the best accuracies are obtained with English as the source language, and Spanish as the target language when using fastText, and that this holds for all levels of supervision. Despite the relatively low precision@ N for OOV source language words, that the results are better than baselines indicates that sub-word level information is transferable across languages via cross-lingual embeddings. Moreover, these results suggest that this is the case even when the languages considered are in different language families and not closely related. This could potentially be applied to improve the handling of OOV words in NLP tasks that rely on cross-lingual word embeddings, such as low-resource document classification, POS tagging, and dependency parsing.

3.3.2 Combined In-vocabulary and OOV Test Set

In this subsection I consider an evaluation that considers both in-vocabulary and OOV source language words. I show that, although the precisions reported in Table 3.3 are relatively low (albeit better than a baseline), an approach that incorporates sub-word knowledge outperforms a method that does not, on a dataset consisting of both in-vocabulary and OOV source language words.

I build a new test dataset consisting of both in-vocabulary and OOV source

Beginning of Table 3.4

Language	Supervision	Method	Precision					
			English source			English target		
			@1	@5	@10	@1	@5	@10
Finnish	Supervised	Copy	12.01	23.30	27.43	25.92	37.04	39.76
		FT	12.18	24.28	28.71	25.92	37.42	40.53
		BPE	4.09	9.92	12.98	8.40	14.89	17.89
	Semi-sup	Copy	11.54	21.00	24.57	21.96	29.98	33.22
		FT	11.62	21.88	25.80	21.94	30.39	33.96
		BPE	3.32	8.64	11.32	6.58	11.69	13.98
	Unsupervised	Copy	11.71	20.97	25.17	22.05	30.84	33.41
		FT	11.92	21.92	26.39	21.99	31.20	34.16
		BPE	0	0.09	0.09	6.06	11.12	13.12
German	Supervised	Copy	15.33	25.18	27.90	18.72	26.32	28.46
		FT	15.29	26.31	29.63	19.02	27.69	30.56
		BPE	5.05	11.01	14.16	6.83	12.17	14.40
	Semi-sup	Copy	16.07	24.79	27.60	19.62	26.37	28.72
		FT	16.15	25.78	29.15	19.86	27.72	30.88
		BPE	5.44	11.79	13.99	6.02	10.68	12.86
	Unsupervised	Copy	16.33	24.97	27.26	16.45	23.42	25.51
		FT	16.45	26.12	28.76	16.79	24.73	27.64
		BPE	4.84	10.58	13.43	0.09	0.27	0.43
Japanese	Supervised	Copy	20.34	29.98	33.08	13.66	22.29	24.23
		FT	20.55	30.57	33.93	13.70	22.60	24.66
		BPE	6.45	11.12	14.23	4.43	10.21	13.87
	Semi-sup	Copy	19.79	28.75	30.79	10.14	16.22	18.28
		FT	19.87	29.21	31.46	10.25	16.74	18.88
		BPE	6.05	11.11	14.18	4.31	9.21	13.64
	Unsupervised	Copy	20.38	28.70	31.38	9.98	15.58	17.53
		FT	20.47	29.13	32.19	10.10	15.99	18.06
		BPE	5.68	9.45	12.32	4.20	10.13	12.44

Continuation of Table 3.4								
Russian	Supervised	Copy	14.73	25.52	28.22	22.25	29.71	31.57
		FT	14.90	25.87	28.88	22.57	30.74	33.44
		BPE	5.61	12.16	15.12	11.78	19.15	21.69
	Semi-sup	Copy	13.05	24.02	27.16	19.32	26.89	29.04
		FT	13.13	24.36	25.57	19.83	27.72	30.88
		BPE	5.13	10.96	13.79	11.18	17.17	19.74
	Unsupervised	Copy	12.91	24.28	27.33	19.99	27.37	29.59
		FT	12.95	24.62	27.93	20.50	28.59	31.25
		BPE	0.09	0.09	0.13	9.79	15.54	17.72
Spanish	Supervised	Copy	24.06	34.47	36.42	27.39	34.86	37.27
		FT	24.06	35.62	38.33	27.52	35.72	38.57
		BPE	11.20	19.40	22.10	11.77	18.93	21.82
	Semi-sup	Copy	25.17	34.20	36.11	27.35	34.30	36.97
		FT	25.24	35.47	37.95	27.57	35.16	38.27
		BPE	11.91	19.04	21.66	12.33	19.32	21.91
	Unsupervised	Copy	25.03	34.07	35.93	25.97	32.44	34.30
		FT	25.16	35.21	37.73	26.17	33.13	35.38
		BPE	11.38	18.51	21.30	11.64	18.24	20.27

Table 3.4: Precision@ N for bilingual lexicon induction for the test set containing both in-vocabulary and out-of-vocabulary words. Results are shown for each language, with English as both the source and target language, for cross-lingual embeddings formed using each level of supervision. “Copy” refers to handling OOVs using the copy baseline, while “FT” indicates employing fastText sub-word embeddings, and “BPE” means using BPE sub-word embeddings to find translations for OOVs. The best precision@ N , for each language, translation direction, level of supervision and the subword embedding method, is shown in boldface.

language words. For each language, I select 1500 test pairs that are in-vocabulary in both the source and target languages from the data of Lample

et al. (2018),⁹ and an equal number of test pairs from the previous test data, i.e., test pairs that are OOV for the source language, and in-vocabulary for the target language.

In these experiments, as in the previous experiments, I compose the representations of OOV source language words from their sub-word embeddings. In the case of in-vocabulary source language words, I use their embeddings as their representation. I then use the transformation matrix to find their target language translations. I refer to the approach which uses fastText’s sub-word information for OOVs as “FT”, and refer to the approach which uses BPE’s sub-word information for OOVs as “BPE”, I compare this against a method that does not use sub-word embeddings. For this latter method, I again use word embeddings to represent in-vocabulary words and use the transformation matrix to find their target language translations. However, in the case of OOV source words, I apply the copy baseline. I refer to this method — that has no knowledge of sub-words, and relies on the copy baseline to translate OOVs — as “Copy”. For this approach, I employ the fastText embeddings. Results are shown in Table 3.4. For all languages, with English as either the source or target language, and for every level of supervision, the FT approach outperforms the Copy approach, except in a small number of cases, specifically Finnish as the source language for precision@1 using the semi-supervised and unsupervised approaches, German as the target language for precision@1

⁹I found very few of the words in these translation pairs to be OOV in the training corpora, motivating the construction of the dataset described in Section 3.2.

using the supervised approach.

I carried out the experiments described in this sub-section using BPE, and the results are indeed relatively poor. For all languages, and levels of supervision, the Copy baseline and FT outperform BPE by a great margin. BPE especially performs quite poorly in the case of the unsupervised setting. These findings are inline with the results in Table 3.3 that BP performs very poorly compared to FT.

Overall these findings indicate that transferring information between languages using sub-word information is not dependent on the method for learning the transformation matrix, and that it is possible to make cross-lingual methods more accurate, and robust with respect to OOVs, by incorporating sub-word information.

I applied McNemar’s test with continuity correction to determine whether the results using the FT method were significantly different from those using the Copy method. To avoid carrying out an overly-large number of tests, I only conduct tests for the supervised method — which based on the findings in Table 3.4 often gives the best performance — and for precision@10 — where the accuracies are highest — although I do so for English as both the source and target language. In each case, the p value is well below the threshold of 0.05 ($p < 0.0002$ in each case) indicating that the difference between Copy and FT is significant in the case of the supervised method for precision@10.

3.3.3 Interpolation with String Similarity

Braune et al. (2018) considered English–German bilingual lexicon induction for rare words. Their approach incorporated sub-word embeddings, and also knowledge about the edit distance between two words. I therefore also considered incorporating edit distance into my approach for bilingual lexicon induction for OOVs. In these experiments, I only considered the supervised method to form cross-lingual word embeddings, and fastText embeddings as the source of sub-word information. I considered only these approaches because of the previous findings that fastText outperforms BPE, and that the supervised method often performs best.

In this approach, I rank target language words using the following linear combination:

$$\lambda \text{sim}(s, t) + (1 - \lambda) \text{NMED}(s, t) \quad (3.1)$$

where sim is cosine similarity, i.e., computed from cross-lingual embeddings; NMED is normalized minimum edit distance; and s and t are a source and target language word, respectively.

For these experiments I randomly sampled 1000 test pairs, and 1000 development pairs from each dataset, i.e., the datasets built from Panlex, in Table 3.2.¹⁰ In these experiments, as for subsection 3.3.1, all the source language words are OOVs and all the target language words are in-vocabulary. The development data was used to tune λ by grid search over the range of 0 to 1

¹⁰I downsampled the datasets due to the large number of edit distance calculations required in this evaluation.

in increments of 0.1. I did not consider Japanese and Russian because they do not use the Latin alphabet and the edit distance between these languages and English does not provide any useful information. For each language, $\lambda = 0.7$ gave the best results on the development data.

Results are shown in Table 3.5. Focusing on precision@1, in each case considered, combining knowledge from embeddings and edit distance improves over using either on its own, indicating that these two sources of information are complementary.¹¹ In terms of precision@5 and @10 I see the same trend, although there are some exceptions for Finnish. Nevertheless, I do not see the massive gains in absolute accuracy from incorporating edit distance reported by Braune et al. (2018), i.e., they achieve gains of 13.8 percentage points in some cases for precision@1, while the highest I achieve is 4.22 percentage points, suggesting that finding translations for OOVs might be particularly challenging compared to finding translations for low frequency words attested in a corpus.

3.3.4 Low-resource Language Experiments

So far I have simulated lower-resource languages by down-sampling the source language corpora. In this section, I consider the case of Cherokee, a true low-resource language. Cherokee is an endangered Iroquoian language, spoken in the United States, with approximately $2k$ speakers. It is a polysynthetic

¹¹The results for $\lambda = 1$ differ from those in Table 3.3 because they are for a sample of the full dataset.

Language	λ	Precision					
		English source			English target		
		@1	@5	@10	@1	@5	@10
Spanish	0	2.64	5.64	6.59	2.34	6.21	7.14
	0.7	6.12	9.11	11.03	5.27	9.60	10.19
	1	3.60	5.76	7.31	1.05	3.28	4.33
	Copy baseline	1.44	-	-	0.82	-	-
German	0	1.10	2.32	2.87	0.67	1.56	2.11
	0.7	2.21	5.51	6.39	2.78	6.56	8.12
	1	0.99	2.98	3.86	1.89	4.00	6.34
	Copy baseline	0.66	-	-	0.44	-	-
Finnish	0	0.35	0.70	0.93	0	0	0.37
	0.7	0.70	1.40	2.10	1.48	2.21	2.21
	1	0.58	1.40	1.87	0.74	2.95	3.32
	Copy baseline	0.00	-	-	0.00	-	-

Table 3.5: Precision@ N incorporating edit distance for OOV source. The best precision@ N , for each language, translation direction and evaluation measure, is shown in boldface.

language written using a syllabary. It is also a moribund language, which means it is an endangered language that is likely to be extinct in the near future.

In these experiments, I use Cherokee as the source language, and English as the target language. Because of the previous findings that fastText embeddings outperform BPE, I again only consider fastText embeddings in these experiments. Specifically, I use fastText embeddings pre-trained on Cherokee Wikipedia.¹² The embedding table for Cherokee contains only 7033 words. For English embeddings, I use the embeddings trained on the full English

¹²<https://fasttext.cc/docs/en/pretrained-vectors.html>

Language	Supervision	Precision		
		English target		
		@1	@5	@10
Cherokee	Supervised	0.75	2.12	2.49
	Semi-supervised	0.00	0.00	0.00
	Unsupervised	0.00	0.00	0.00
	Copy	0.00	-	-

Table 3.6: Precision@ N for Cherokee, using English as the target language, for each level of supervision. The best precision@ N is shown in boldface.

Wikipedia from the previous experiments.

I build training and test datasets from the English–Cherokee translations in Panlex. For the supervised method, I use all pairs for which both the English and Cherokee words are in-vocabulary for their respective word embedding models as training instances. This gives 1143 training instances. From these training pairs, a subset of 25 pairs is selected to train the semi-supervised method. For test data, I use all translation pairs for which the source Cherokee word is OOV, and the target English word is in-vocabulary, which gives a total of 1050 test instances.¹³

Results are shown in Table 3.6. These results indicate that, in the case of a morphologically-rich, truly low-resource language, sub-word embeddings — along with a supervised approach to learning a transformation matrix — provide information about translations for OOV words. However, the precision@ N for the semi-supervised and unsupervised methods is 0.00%,

¹³I do not consider the case of a test set consisting of both in-vocabulary and OOV source language words because only 1143 translation pairs are in-vocabulary for both languages, all of which are used for training the supervised approach.

which suggests that in the case of a truly low-resource language, these methods might not be capable of handling OOVs.¹⁴

3.4 Summary

In this chapter I considered whether sub-word embeddings can be leveraged in cross-lingual word embedding models. Specifically I evaluated sub-word embeddings in a novel bilingual lexicon induction task in which I identify target language translations for OOV source language words. The findings indicate that, although the accuracy is not high in absolute terms, sub-word embeddings nevertheless provide information that can be leveraged for identifying translations for OOV words, including in the case of a truly low-resource, morphologically-rich language, specifically Cherokee. I additionally showed that sub-word embeddings can be leveraged to find translations in the case that the test data consists of a mixture of both OOVs and in-vocabulary words. The findings further indicate that bilingual lexicon induction for OOVs can be improved by incorporating orthographic similarity.

¹⁴A random baseline is also considered but as in the previous sections, it yields a precision of 0.00%

Chapter 4

Sub-word Level Cross-lingual Word Representations

In this chapter, I propose a model that can learn cross-lingual word embeddings from a modest amount of monolingual data and a bilingual dictionary. I rely on bilingual dictionaries because they are relatively-widely available. For example, Panlex (Baldwin et al., 2010) is a translation resource that combines many bilingual dictionaries and provides translations for 5700 languages. The proposed model is an extension of the method proposed by Duong et al. (2016). In their work, they only considered word embeddings, and so their method is unable to form representations for OOVs and therefore is not expected to perform well for low-resource or morphologically-rich target languages. I extend the method of Duong et al. (2016) by incorporating sub-word information in the process of training cross-lingual word embeddings.

In this way, a shared embedding space is formed that not only contains embeddings for both source and target language words, but that has also been enriched with sub-word embeddings enabling representations to be formed for OOVs.

The structure of this chapter is as follows. In Section 4.1 I describe the base model and give a very detailed explanation of how the method of Duong et al. (2016) works. Then in Section 4.2 I demonstrate the proposed method and how sub-word information is incorporated during training. Section 4.3 presents the experiments and results. More specifically, in Section 4.3.1.1, I show the impact of incorporating sub-word information during training by evaluating the proposed method and Duong et al. in a bilingual lexicon induction task. Then to demonstrate that the proposed method is also effective in a monolingual space, I show the results for monolingual word similarity in Section 4.3.1.2. I further show the impact of the proposed method on downstream tasks in Section 4.3.2. In Section 4.3.3, in order to show the effectiveness of the proposed method in the case of low-resource and morphologically-rich languages, the results of a series of experiments, i.e., bilingual lexicon induction on in-vocabulary and OOV words, on 12 low-resource and morphologically-rich languages are shown. This chapter, except for Section 4.3.3, is an extended version of Hakimi Parizi and Cook (2020b).

4.1 Joint Training with Pseudo-bilingual Corpora

Duong et al. (2016) introduce an approach to learning cross-lingual word representations that can jointly learn representations for words in two languages — referred to as the source and target language — without requiring a parallel corpus. This method is an extension of CBOW (Mikolov et al., 2013a) that uses a pseudo-bilingual corpus which is built using two monolingual corpora and a bilingual dictionary. A prefix is added to each word in each monolingual corpus indicating its language. Then, the monolingual corpora are concatenated and the sentences are shuffled. Also, each center word is replaced by its translation on-the-fly during the CBOW training procedure. This process leads to the construction a pseudo-bilingual corpus. The CBOW objective function — shown in Equation 4.1 — is only capable of capturing monolingual similarities, thus Equation 4.2 is proposed to adapt it to cross-lingual settings.

$$O = \sum_{i \in D} (\log \sigma(u_{w_i}^T h_i) + \sum_{j=1}^p \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i)) \quad (4.1)$$

$$\begin{aligned}
O = \sum_{i \in D_s \cup D_t} & (\alpha \log \sigma(u_{w_i}^T h_i) + \\
& (1 - \alpha) \log \sigma(u_{\bar{w}_i}^T h_i) + \\
& \sum_{j=1}^p \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i))
\end{aligned} \tag{4.2}$$

where h_i encodes the context vector, \bar{w}_i is the translation of w_i , α is a weight parameter, p is the number of negative samples, and D_s and D_t are the source and target language vocabularies, respectively. In Equation 4.1, D is the vocabulary of the monolingual corpus. The intuition behind this approach is similar to CBOW. For each context window, it tries to predict the middle word, w_i , based on the context it appears in, h_i . The one difference here compared to CBOW is that not only does the method try to predict w_i , but also it replaces w_i with its translation, \bar{w}_i , and attempts to predict the translation based on h_i . In this way, since the word and its translation are learned by appearing in the same context, their embeddings will be located closer to each other in the vector space.

Duong et al. (2016) also propose an approach to finding the best translation for polysemous words using the expectation maximization algorithm. They argue a word’s meaning is dependent on the context that it appears in, and it is possible to find the correct translation for a polysemous word by looking at the context and finding a translation that conveys a similar meaning as the words in the context. It is done by computing the cosine similarity of the context window — the average of the embeddings for words in the context —

and the possible translations in the bilingual lexicon seed. The translation which achieves the highest similarity score is selected as the correct translation. Thus the translation for a word is selected based on its context. Equation 4.3 demonstrates their solution for dealing with polysemous words, where \bar{w}_i is a possible translation, h_i denotes the context and *cosine* is cosine similarity between the vector representation of the translation and the context.

$$translation = \max_{\bar{w}} cosine(\bar{w}_i, h_i) \quad (4.3)$$

Duong et al. (2016) further argue that each of the matrices V and U in word2vec encode different information: V is better for capturing monolingual characteristics, whereas U preserves cross-lingual information. In each update, the words in the context are updated in the matrix V and the target word and negative samples are updated in the matrix U . Therefore, in each update context words get closer to each other in V and the target word gets further away from the negative samples in U . In the method introduced by Duong et al. (2016), aside from the target word and its context, the translation word also gets updated in U . Therefore, the word and its translation get closer to each other in the matrix U and they get further away from the negative samples. Duong et al. (2016) achieve their best results in both monolingual and cross-lingual evaluations by combining V and U during the training phase using a regularization term, δ , in the objective function, Equation 4.4.

I refer to this approach as DUONG2016.

$$O' = O + \delta \sum_{w \in V_s \cup V_t} \|u_w - v_w\|_2^2 \quad (4.4)$$

4.2 Joint Training Incorporating Sub-word Information

Incorporating sub-word information in training word embeddings enhances the quality of the learned embeddings (Bojanowski et al., 2017). Moreover, because sub-word embeddings can be used to construct representations for OOVs, approaches that incorporate sub-word embeddings are better-suited for low-resource and morphologically-rich languages which are expected to have relatively high rates of OOVs. Therefore, I extend DUONG2016 by incorporating sub-word information during training.

To incorporate sub-word information, I follow a similar approach to Bojanowski et al. (2017). They introduce a novel word embeddings method, based on word2vec, that represents a word as the sum of its character n -grams and learns an embedding for each of these n -grams, which are called sub-words. Each word in the training corpus is augmented with special beginning and end of word markers. Each word is then represented as a bag of character sequences (i.e., sub-words); in my experiments I consider sequences of length 3–6 characters. I additionally include the entire word itself (with beginning and end of word markers) among the sub-words. The embedding for a word

is formed by averaging its sub-word embeddings. This gives the following objective function:

$$O = \sum_{i \in D_s \cup D_t} (\alpha \log S(w_i, h_i) + (1 - \alpha) \log S(\bar{w}_i, h_i)) + \sum_{j=1}^p \mathbb{E}_{w_j \sim P_n(w)} \log -S(w_j, h_i) \quad (4.5)$$

where S , shown in Equation 4.6, measures the similarity between a word and context, taking into account sub-words:

$$S(w, h) = \frac{1}{|G_w|} \sum_{g \in G_w} z_g^T h \quad (4.6)$$

where G_w is the set of sub-words appearing in w , and z_g is the sub-word embedding for g . To calculate h , I average representations for each word appearing in the context, where each word is represented by the average of its sub-word embeddings.

4.3 Experiments

In this section, I describe the experiments which are conducted to demonstrate the advantage of incorporating sub-word information in the process of training cross-lingual word embeddings. This section is divided into three subsections. In the first subsection, I describe preliminary experiments, which are conducted by following Duong et al. (2016), in order to show the advantages of the

proposed model in a comparable way to the original paper and also provide a way to systematically find the best parameter settings for each method. In the second subsection the results and details of an extrinsic evaluation, document classification, are shown to demonstrate the ability of the proposed model to give improvements on a downstream task. The last subsection is dedicated to prove my claim that this method is a suitable choice for low-resource and morphologically-rich languages. In each subsection different languages are employed for the experiments and consequently different resources, therefore the resources for each experiment are described separately in each subsection.

4.3.1 Intrinsic Evaluation

Following Duong et al. (2016), two intrinsic evaluations, bilingual lexicon induction (BLI) and monolingual word similarity, are chosen to demonstrate the performance of DUONG2016 and the proposed method in two different settings, cross-lingual and monolingual.

4.3.1.1 Bilingual Lexicon Induction

For evaluation, I begin by simulating lower-resource languages using 5 well-resourced languages: Dutch, English, German, Italian and Spanish. These languages are those considered by Duong et al. (2016). I only consider pairs of languages with English as either the source or target language, and one of the remaining 4 languages as the other language, to be able to use a widely-used evaluation data-set, MUSE (Lample et al., 2018), which only has pairs of

translations from English to other languages and other languages to English. To train word embeddings for these languages, I use pre-processed Wikipedia dumps (Al-Rfou' et al., 2013), which are already tokenized and cleaned. To simulate the case of lower-resource languages, following Duong et al. (2016), I randomly select 5 million sentences for each language from their Wikipedia dump. Table 4.1 shows the number of tokens and types in each corpus.

I use a bilingual dictionary extracted from Panlex as the cross-lingual signal in the proposed approach. My study builds on the work of Duong et al. (2016), and so I use the same dictionaries that they did, which were extracted from Panlex.¹ It should also be noted that before training the cross-lingual embeddings, the MUSE test pairs are removed from the Panlex training dictionary. Table 4.1 also shows the size of each dictionary, with English as the source language, and the other language as the target language.² The coverage of the dictionary with respect to the number of tokens, types, and embeddings learned is also shown. For example, 68% coverage for Italian tokens means that 68% of tokens in the Italian corpus occur in the bilingual dictionary.

Bilingual lexicon induction (BLI) is a standard task to evaluate the quality of cross-lingual word embeddings (Vulić and Moens, 2013; Artetxe et al., 2017; Ruder et al., 2019). In this task, I try to find the translation of a source language word in the target language by looking at its nearest neighbours in

¹<https://github.com/longdt219/XlingualEmb>

²The dictionary size for English is therefore not shown.

Language	Family	Tokens	Types	Embeddings	Dict. entries
Dutch	Germanic	84M (64%)	1.3M (8%)	303K (28%)	406K
English	Germanic	121M	1.1M	240K	-
German	Germanic	92M (68%)	1.8M (8%)	411K (25%)	964K
Italian	Romance	119M (68%)	1.2M (7%)	304K (22%)	560K
Spanish	Romance	130M (75%)	1.1M (7%)	279K (22%)	712K

Table 4.1: The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.

the shared cross-lingual space. Ideally, a word and its translation would be located close to each other in the shared cross-lingual word embedding space. Here I focus on comparing the proposed method with DUONG2016 and so in all cases, English is the target language and the other languages are treated as the source language.

Following previous work (e.g. Lample et al., 2018; Joulin et al., 2018; Jawanpuria et al., 2019), I consider MUSE test sets for evaluation. Word pairs occurring in both the MUSE test sets and my training dictionaries are removed from the training data before training the embeddings. I report $\text{precision}@N$ — for $N = 1, 5,$ and 10 — where the system is scored as correct if the gold-standard target word is amongst the top- N most similar target language words (Ruder et al., 2019). I use cosine as the similarity measure.

Results are shown in Table 4.2. I begin by considering DUONG2016 and my model using the best parameter settings from Duong et al. (2016), i.e., a learning rate of 0.025, 25 negative samples, a window size (c) of 48, an

Model	es-en			it-en			nl-en		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
DUONG2016 ($c = 48, d = 200$)	54.59	83.12	86.87	45.98	77.11	81.79	40.73	71.72	77.06
DUONG2016 ($c = 5, d = 200$)	28.20	70.26	76.36	21.08	60.78	67.47	24.36	55.07	62.65
DUONG2016 ($c = 20, d = 200$)	50.50	82.92	87.07	41.83	77.11	81.53	41.41	72.19	77.88
DUONG2016 ($c = 48, d = 300$)	50.90	83.86	87.54	44.24	77.44	82.33	38.16	71.31	77.67
Proposed Model ($c = 48, d = 200$)	60.15	79.84	84.26	54.62	73.83	78.92	42.25	67.39	72.80
Proposed Model ($c = 5, d = 200$)	41.39	78.63	85.06	36.21	72.42	79.45	36.54	69.15	76.25
Proposed Model ($c = 20, d = 200$)	59.14	83.12	87.27	54.02	77.64	82.00	47.56	73.00	78.69
Proposed Model ($c = 20, d = 300$)	60.21	84.53	89.28	55.15	80.12	84.94	46.21	74.83	80.11
VecMap	81.27	91.07	93.27	76.13	86.87	89.47	71.53	83.93	86.53

Table 4.2: Precision@ N for bilingual lexicon induction. The best performance, for each dataset and evaluation measure, is shown in boldface.

embedding size (d) of 200, sub-sampling of $1e^{-4}$, α of 0.5, and δ set to 0.01.³ In terms of precision@1, my model out-performs DUONG2016 for each language, but for precision@5 and precision@10, DUONG2016 performs better.

A window size of 48 takes into account a relatively large amount of context for the target word; however, when incorporating sub-words, as for the proposed model, this wide context could also add noise because of the large number of sub-words in the context, and the wide range of contexts in which sub-words occur. I therefore consider a window size of 5, the fastText default, and 20, which balances having a larger window size against introducing too much noise. Results are shown for this setup for both DUONG2016 and the proposed model. For both models, a window size of 5 performs relatively poorly. For DUONG2016, the original window size of 48 performs best in terms

³The differences between the results for DUONG2016 here and the numbers reported in Duong et al. (2016) are due to differences in the test set. I use the MUSE test set, which was not available in 2016, but is more widely used now.

of precision@1 for Spanish and Italian, but not Dutch. For my proposed model, the intermediate window size of 20 performs best, except for precision@1 for Spanish and Italian. These results suggest that a model including sub-word information might not be able to use information from a very wide context as effectively as a word-only model.

Next I consider increasing the embedding size to 300, which is commonly used for fastText (Bojanowski et al., 2017). I consider this for the best window size for each model, i.e., 48 for DUONG2016 and 20 for the proposed model.⁴ The proposed model with a window size of 20 and embedding size of 300 outperforms DUONG2016 for all parameter settings considered, for all languages and evaluation measures. The difference between the proposed model in this configuration, and DUONG2016 using its original parameter settings, is significant ($p < 4.31e^{-6}$) using a one-sided McNemar’s test with continuity correction. This demonstrates that incorporating sub-word knowledge during training of cross-lingual word embeddings enhances the quality of the resulting word representations.

These are not state-of-the-art results, where prior work has obtained higher precision. As a point of comparison, I also present results for VecMap (Artetxe et al., 2018a), a supervised mapping-based approach. These results for VecMap are achieved using fastText embeddings trained on full Wikipedia corpora for each language. My proposed model, on the other hand, is trained

⁴I also considered a window size of 20 and embedding size of 300 for DUONG2016, but this did not give improvements.

on substantially smaller corpora because I focus on approaches that could be applied to lower-resource languages. mBERT and Chaudhary et al. (2018) are further points of comparison that I do not include because of the resource requirements, and reliance on language-specific tools, respectively, of these methods.

For the rest of the chapter, the “proposed model” refers to the model with an embedding size of 300 and window size of 20. Since changing the window and embedding sizes does not consistently lead to improvements for DUONG2016, and has a negative impact on precision@1, I continue to use the best parameter settings from Duong et al. (2016) for this method.

4.3.1.2 Monolingual Word Similarity

Here I evaluate the quality of cross-lingual word representations in a monolingual setting. I compare cross-lingual embeddings from the proposed model and DUONG2016. I further consider monolingual embeddings from fastText, a well-known method to learn embeddings that uses sub-word information, as a baseline. I consider several parameter settings for fastText. In particular, I consider the best parameter settings for DUONG2016 (CBOW, $c = 48$, $d = 200$), the best parameter settings for the proposed model (CBOW, $c = 20$, $d = 300$), and commonly-used fastText settings (skipgram, $c = 5$, $d = 300$, and 5 negative samples). In addition, I consider three corpus sizes to train fastText: 5 million sentences (i.e., the same amount of monolingual text that DUONG2016 and the proposed method are trained on), 10 million sen-

tences (the total amount of text in both languages that DUONG2016 and the proposed method are trained on), and full Wikipedia corpora. For the full Wikipedia corpora I only consider the commonly-used parameter settings. For predicting word similarity, I use cosine as the similarity score. Following Duong et al. (2016), I consider English and German for these experiments. I use three datasets for evaluation: English WordSim353 (WS-en, Finkelstein et al., 2002), German WordSim353 (WS-de, Luong et al., 2015), and Stanford Rare Words (RW-en Luong et al., 2013). The number of OOVs in WS-en and WS-de is very low (none for WS-en, and two for WS-de). For these datasets, I therefore report results only for in-vocabulary items. For RW-en, however, roughly 25% of the test pairs include an OOV. For this dataset I therefore also report results considering both in-vocabulary words and OOVs (referred to as “RW-en+OOV”). Because DUONG2016 is not capable of forming representations for OOVs, in such cases I assign these test pairs the average cosine similarity score over test pairs that are in-vocabulary. Table 4.3 shows the results. For each dataset, the proposed model outperforms DUONG2016, and also fastText, in all configurations considered. These results indicate that a cross-lingual signal can be used to not only form a cross-lingual shared space, but also to enhance the quality of monolingual embeddings. Note that DUONG2016 improves over fastText on WS-en and WS-de, but not on RW-en (or RW-en+OOV). This indicates that sub-word information is particularly important for forming representations for low-frequency words.

Model	WS-en	WS-de	RW-en	RW-en+OOV
DUONG2016	74.46	69.72	44.06	37.68
Proposed Model	75.67	70.49	51.57	49.51
<i>trained on 5 million sentences</i>				
fastText (CBOW, $c = 48, d = 200$)	55.40	46.91	40.56	39.77
fastText (CBOW, $c = 20, d = 300$)	53.66	43.73	37.92	37.35
fastText (skipgram, $c = 5, d = 300$)	69.02	63.79	49.50	47.94
<i>trained on 10 million sentences</i>				
fastText (CBOW, $c = 48, d = 200$)	57.72	45.18	41.31	40.74
fastText (CBOW, $c = 20, d = 300$)	54.55	42.62	38.85	38.36
fastText (skipgram, $c = 5, d = 300$)	69.91	60.90	49.74	48.74
<i>trained on full Wikipedia corpora</i>				
fastText (skipgram, $c = 5, d = 300$)	73.77	66.63	48.61	48.09

Table 4.3: Spearman’s correlation for monolingual similarity on each dataset, for each method considered. The best performance on each dataset is shown in boldface.

4.3.1.3 Summary

In this section, I proposed an approach to learning cross-lingual word embeddings that incorporates sub-word information during training, and relies on only monolingual corpora and a bilingual dictionary for training. This approach could be particularly well-suited to lower-resource, morphologically-rich languages, for which large parallel corpora are not available. I evaluated the proposed approach, on a variety of simulated lower-resource languages, for the tasks of BLI and monolingual word similarity. The results on BLI and monolingual word similarity indicated that incorporating sub-word information during training enhances the quality of the resulting cross-lingual, as well as monolingual, representations. Now that I demonstrated the superiority of the proposed method on two intrinsic tasks and compared it against the

original method of Duong et al. (2016), the next section will be dedicated to showing the effectiveness of the proposed model in a downstream task.

4.3.2 Extrinsic Evaluation

Here I consider an extrinsic evaluation which uses cross-lingual word embeddings in a downstream task, specifically cross-lingual document classification. This task is motivated by the situation where sufficient labelled training data is not available for a low-resource language. I consider zero-shot classification, i.e., I train a classifier and tune parameters on a rich-resource source language, and then apply the classifier directly to documents in a low-resource target language. In order to show that the embeddings obtained from the proposed model are of a high quality, I also present the results for supervised document classification, in which a classifier is trained on the labelled data of a language and then applied on a test set of the same language.

4.3.2.1 Resources

Following previous work (e.g., Artetxe and Schwenk, 2019; Wu and Dredze, 2019), I use the MLDoc dataset (Schwenk and Li, 2018), which is a subset of the RCV1/RCV2 datasets (Lewis et al., 2004) with balanced classes for training, development, and test sets for the following languages: Chinese, English, French, German, Italian, Japanese, Spanish, and Russian. It has 1000 documents in each of the training and development sets, and 4000 documents in the test set, for each language. Following Artetxe and Schwenk, I use

Language	Family	Tokens	Types	Embeddings	Dict. entries
Chinese	Sino-Tibetan	30M (64%)	0.2M (20%)	86K (43%)	1983K
English	Germanic	121M	1.1M	240K	-
French	Romance	135M (80%)	1.1M (9%)	288K (30%)	1068K
German	Germanic	92M (68%)	1.8M (8%)	411K (25%)	964K
Italian	Romance	119M (68%)	1.2M (7%)	304K (22%)	560K
Japanese	Japanese	22M (76%)	0.3M (21%)	107K (47%)	736K
Russian	Slavic	84M (56%)	1.7M (7%)	445K (68%)	1594K
Spanish	Romance	130M (75%)	1.1M (7%)	279K (22%)	712K

Table 4.4: The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.

English as the source language, and the other languages as target languages. This direction is chosen since my goal is to train a classifier on a rich-resource language that has plenty of available training data and then directly apply it on a low-resource language that does not have any labeled training data available.

This study builds on the work of Duong et al. (2016), and so for languages that they consider — Dutch, German, Italian, Japanese, and Spanish — I use the same dictionaries that they did, which were extracted from Panlex.⁵ For Chinese, French, and Russian I extract dictionaries from Panlex using the same approach as Duong et al. To build corpora to train embeddings, again following previous work (Duong et al., 2016; Klementiev et al., 2012), I first randomly sample 400k sentences for each of the source and target language

⁵<https://github.com/longdt219/XlingualEmb>

from RCV1/RCV2,⁶ and then combine these in-domain corpora with larger Wikipedia corpora. I use the Wikipedia corpora described in Table 4.4.

4.3.2.2 Zero-Shot Document Classification

First I consider zero-shot classification, i.e., I train a classifier and tune parameters on the source language, and then apply the classifier directly to target language documents. Following Artetxe and Schwenk (2019), I use English as the source language, and use the other languages as target languages.

I represent documents as the average of their words' embeddings, where the embeddings are learned by the proposed approach from the corpora described above. I then use a feed-forward classifier (LASER, Artetxe and Schwenk, 2019), which has been previously applied to cross-lingual document classification, with one hidden layer of 10 hidden units, a learning rate of 0.001, dropout set to 0.2, and a batch-size of 12, as suggested by Artetxe and Schwenk.

I compare the proposed approach against several benchmarks. First I consider the same approach described above, but using embeddings from DUONG2016 instead of the proposed approach. In this case, embeddings for OOVs are not available, and so OOVs are simply ignored in forming document representations. I further consider two strong benchmark approaches — LASER

⁶I sampled 80k documents for both the source and target languages, and then sampled 400k sentences. For Spanish, Italian, Russian and Chinese I use all of their RCV2 documents because the total number of documents available for these languages is less than 80k.

Model	Target language							Average
	Chinese	French	German	Italian	Japanese	Russian	Spanish	
DUONG2016	54.12	87.82	86.95	73.88	71.12	50.15	77.90	71.71
LASER	70.98	78.03	86.25	70.20	60.95	67.25	79.30	73.28
mBERT	76.9	72.06	80.2	68.9	56.5	73.7	72.6	71.55
Proposed Model	69.55	86.45	90.22	72.90	74.62	53.30	78.47	75.07
XLM _{ft} UDA	93.32	96.05	96.95	-	-	89.07	96.8	-

Table 4.5: Accuracy on the MLDoc zero-shot cross-lingual document classification task, for each model and target language, with English as the source language. The average accuracy over all target languages is also shown.

(Artetxe and Schwenk, 2019) and mBERT (Devlin et al., 2019) — that are widely used for comparison (e.g., Wu and Dredze, 2019; Patidar et al., 2019; Keung et al., 2019). Artetxe and Schwenk recently improved their model, and reported updated results.⁷ I use these improved results for comparison. I use mBERT results reported by Wu and Dredze (2019).

Results are shown in Table 4.5. None of the approaches considered performs best for all languages. However, in terms of the average accuracy over all target languages, the proposed model performs better than DUONG2016, LASER and mBERT. It is worth noting that the proposed model is trained on only 5.4 million sentences in each language, and does not require a parallel corpus. LASER, the next best method in terms of average accuracy, on the other hand, is trained on 225 million parallel sentences. Furthermore, the proposed model outperforms mBERT — a very large language model-based approach — on average, and for all target languages except Chinese and Russian. The current state-of-the-art for MLDOC is XLM_{ft} UDA (Lai

⁷<https://github.com/facebookresearch/LASER/tree/master/tasks/mldoc>

et al., 2019). This model is pre-trained for 15 languages, but not Italian and Japanese, and so results are not available for these languages. XLM_{ft} UDA does however substantially out-perform the proposed model on the other languages, but also requires a large parallel corpus for training.

4.3.2.3 Supervised Document Classification

To evaluate the performance of the proposed model in a monolingual downstream task, I also consider a supervised document classification setup. In this setup, one classifier is trained separately for each language, using the training and validation sets in MLDOC. I use the same document representations and parameter settings as in the previous cross-lingual document classification experiments.

Results are shown in Table 4.6. In terms of average accuracy, mBERT performs best, although the accuracies for the proposed model and DUONG2016 are quite close, with the proposed model narrowly outperforming DUONG2016. For each language, the proposed model (as well as DUONG2016 and mBERT) outperforms LASER. Although the proposed model is relatively simple and inexpensive to train, and does not require a parallel corpus, it performs roughly on par with, or better than, more complex models that are trained on much larger corpora. The current state-of-the-art for supervised classification of MLDOC is MultiFiT (Eisenschlos et al., 2019). This model is a language model that can be trained on a small amount of raw text, as low as 100M tokens. Then it can be fine-tuned on different tasks such as document

Model	Chinese	French	German	Italian	Japanese	Russian	Spanish	Average
DUONG2016	89.90	93.55	94.70	86.28	88.95	85.68	94.15	90.46
LASER	88.98	90.80	92.70	85.93	85.15	84.65	88.75	88.14
mBERT	89.3	93.4	93.3	88	88.4	87.5	95.7	90.80
Proposed Model	89.93	93.72	94.78	86.80	88.47	85.85	94.25	90.54
MultiFiT	92.52	94.75	95.90	90.25	90.03	87.65	96.07	92.45

Table 4.6: Accuracy on the MLDoc supervised document classification task, for each model and language. The average accuracy over all languages is also shown. The highest accuracy in each case is shown in boldface.

classification.

4.3.2.4 Summary

Here, I demonstrated how we can utilize cross-lingual representations in order to perform document classification. Two different scenarios were considered, zero-shot document classification and supervised document classification. I compared the results from the proposed method with DUONG2016 and two powerful baselines, mBERT and LASER. The proposed method not only outperformed DUONG2016, but also even though the proposed method is simpler than these baselines and it requires less data to be trained, it performs on par or in some cases better.

4.3.3 Low-Resource and Morphologically-Rich Languages

DUONG2016 can learn cross-lingual word embeddings from modest size monolingual corpora, using a bilingual dictionary as the cross-lingual signal. Bilin-

gual dictionaries are available for many language pairs, e.g., Panlex (Baldwin et al., 2010) provides translations for roughly 5700 languages. These training resource requirements suggest this method could be well-suited to lower-resource languages. However, this word-level approach is unable to form representations for out-of-vocabulary (OOV) words, which could be particularly common in the case of low-resource, and morphologically-rich, languages. Here, I evaluate my proposed method which is an extension of DUONG2016 which incorporates sub-word information during the joint training process. Because my proposed method incorporates sub-word information, and is therefore able to form representations for OOVs, and additionally does not require parallel corpora for training, it could be particularly well-suited to lower-resource, and morphologically-rich, languages.

Most prior work on BLI focuses on in-vocabulary words and well-resourced languages (e.g., Artetxe et al., 2017; Ormazabal et al., 2019; Zhang et al., 2020), although there has been some work on low-frequency words (Braune et al., 2018), as well as low-resource languages (Anastasopoulos and Neubig, 2020).

In this section, I consider BLI for twelve lower-resource languages covering several language families, and also consider an evaluation focused on OOVs. The results indicate that the proposed method gives improvements, particularly for OOVs.

Language	Family	Tokens	Types	Embeddings	Dict. entries
Afrikaans	Germanic	25M (64%)	0.6M (2%)	114K (11%)	70K
Albanian	Albanian	21M (38%)	0.6M (1%)	127K (4%)	17K
Azerbaijani	Turkic	36M (35%)	1.3M (1%)	241K (3%)	25K
Bengali	Indic	26M (50%)	1.2M (2%)	179K (9%)	114K
Bosnian	Slavic	18M (28%)	0.7K (1%)	147K (4%)	23K
Croatian	Slavic	54M (60%)	1.3M (4%)	302K (15%)	388K
English(down-sized)	Germanic	121M	1.1M	240K	-
Estonian	Uralic	38M (47%)	1.7M (3%)	325K (11%)	201K
Greek	Greek	78M (60%)	1.4M (3%)	299K (10%)	253K
Hebrew	Semitic	143M (39%)	2.6M (1%)	520K (3%)	79K
Hindi	Indic	34M (74%)	0.9K (4%)	144K (18%)	296K
Hungarian	Uralic	133M (51%)	3.9M (2%)	686K (8%)	460K
Turkish	Turkic	79M (50%)	1.7M (2%)	354K (9%)	319K

Table 4.7: The size of the corpus for each language, in terms of the number of tokens and types. The language family, number of embeddings learned from each corpus, and number of entries in the bilingual dictionary, are also shown for each language. The parenthetical numbers indicate coverage in the dictionary.

4.3.3.1 Resources

I consider BLI from twelve lower-resource source languages to English. The languages (shown in Table 4.7) were selected to cover a variety of language families, while having small to medium size Wikipedias and BLI evaluation datasets available. I compare my proposed method with DUONG2016 and VECMAP (Section 2.2.1, page 22), a supervised mapping-based method (Artetxe et al., 2018a). In each case, I use cosine similarity to find the closest target language translations for a source language word. I evaluate using $\text{precision}@N$ (for $N = 1, 5, 10$) as in Section 4.3.1.1.

The corpus for each language is a Wikipedia dump from 27 July 2020, cleaned

using tools from Bojanowski et al. (2017). All corpora are tokenized using EuropolExtract (Ustaszewski, 2019), since it covers a broad range of languages that use the Latin, Cyrillic, Hebrew or Greek alphabet. For Bengali and Hindi, which are not part of EuropolExtract, the NLTK tokenizer package (Bird et al., 2009) is used.

DUONG2016 benefits from a relatively large training dictionary (Duong et al., 2016), whereas supervised mapping-based approaches tend to see a reduction in performance with seed lexicons larger than roughly 5k pairs (Vulić and Korhonen, 2016). Training translation pairs from MUSE (Lample et al., 2018) are therefore used to train VECMAP, except for Azerbaijani, which is not included in MUSE, where data from Anastasopoulos and Neubig (2020) is used, which consists of 2001 unique training translation pairs. For DUONG2016 and the proposed method I follow Duong et al. to create large training dictionaries by extracting translation pairs from Panlex and removing MUSE test pairs from the Panlex training dictionary. Details of the training corpora and Panlex dictionaries are shown in Table 4.7.

4.3.3.2 BLI for In-Vocabulary Words

For these experiments I use MUSE test data for all languages except Azerbaijani, where I use test data from Anastasopoulos and Neubig (2020). Because my focus here is on in-vocabulary words, I only consider translation pairs that are in-vocabulary with respect to the embedding matrices learned from the training corpora. Results are shown in Table 4.8.

For each language and evaluation measure, I see that the proposed method improves over DUONG2016. This finding indicates that DUONG2016 can indeed be improved by incorporating sub-word information during training. Comparing the proposed method and VECMAP, the results are more mixed. In terms of average precision@1, the results are comparable, although for precision@5 and precision@10 the proposed method substantially outperforms VECMAP. I now consider an evaluation focused on OOV source language words.

4.3.3.3 BLI for OOVs

For low-resource and morphologically-rich languages, large numbers of OOVs are inevitable, and must be handled. In the case of a low-resource language since we do not have access to large enough training corpora, it is certain that a substantial number of that language’s words will not be in the training data. In the case of a morphologically-rich language, we do not have access to all word forms in the training corpus and so we cannot learn any embeddings for them. Since the purpose of cross-lingual word embeddings is to overcome the shortcomings of not having access to enough training data, it is crucial for a cross-lingual method to be able to form representations for OOVs.

The current mapping methods do not consider sub-word embeddings during the training process and work at the word-level. However, if we use an embedding method that also learns sub-word embeddings, we can form an embedding for an OOV by composing its sub-word embeddings and then

Language	Method	% Precision		
		@1	@5	@10
Afrikaans	VECMAP	21.92	35.31	41.16
	DUONG2016	23.26	44.54	50.25
	Proposed Method	31.08	55.53	63.14
Albanian	VECMAP	31.03	47.66	52.83
	DUONG2016	8.04	18.31	22.50
	Proposed Method	18.31	34.52	40.67
Azerbaijani	VECMAP	30.77	46.15	46.15
	DUONG2016	23.08	38.46	53.85
	Proposed Method	38.46	61.54	61.54
Bengali	VECMAP	17.37	32.84	40.37
	DUONG2016	17.95	39.20	46.80
	Proposed Method	22.73	49.05	55.59
Bosnian	VECMAP	21.68	37.27	42.80
	DUONG2016	4.34	13.29	16.78
	Proposed Method	9.23	24.83	30.91
Croatian	VECMAP	33.38	48.68	54.31
	DUONG2016	32.81	62.56	67.54
	Proposed Method	43.35	69.18	75.52
Estonian	VECMAP	25.22	40.82	45.39
	DUONG2016	19.85	45.49	52.36
	Proposed Method	32.77	61.04	68.17
Greek	VECMAP	46.57	64.94	69.85
	DUONG2016	30.96	65.21	69.97
	Proposed Method	43.00	73.01	78.80
Hebrew	VECMAP	40.50	55.17	59.50
	DUONG2016	20.62	46.18	52.40
	Proposed Method	25.69	53.48	61.93
Hindi	VECMAP	28.22	46.16	51.94
	DUONG2016	31.32	61.22	67.21
	Proposed Method	37.72	63.97	70.65
Hungarian	VECMAP	39.02	55.22	61.58
	DUONG2016	27.51	57.43	62.92
	Proposed Method	35.81	64.79	71.82
Turkish	VECMAP	34.20	51.30	56.51
	DUONG2016	25.70	56.62	62.39
	Proposed Method	32.77	61.04	68.17
Average	VECMAP	30.82	46.79	51.87
	DUONG2016	22.12	45.71	52.08
	Proposed Method	30.91	56.00	62.24

Table 4.8: Precision@ N for BLI for in-vocabulary words. The best precision for each dataset and evaluation measure is shown in boldface.

Language	Number of Pairs	Number of Unique Pairs
Afrikaans	4170	2229
Albanian	1452	908
Azerbaijani	1176	806
Bengali	7632	3314
Bosnian	1404	942
Croatian	23826	9809
Estonian	14811	10543
Greek	19773	8260
Hebrew	1306	978
Hindi	13667	8746
Hungarian	21073	10687
Turkish	9572	5102

Table 4.9: The number of translation pairs and unique pairs (i.e., the source language word is unique) in each test set, for each language.

map it to the shared space using the learned mapping matrix and then find its nearest neighbor in the target language. This is the approach that was considered in Chapter 3. Nonetheless, such an approach is unlikely to perform as well as the proposed approach because sub-words were not part of the learning process of forming the shared space.

I design a test similar to the one introduced in Chapter 3. Test-sets are extracted from Panlex, in which the source language words are OOVs and the target language words are in-vocabulary. Then, to eliminate the noisy translation pairs from Panlex, the source words are intersected with the source language training corpus and only those words that are in the corpus but are not in the embedding matrix are kept. The details of the test sets are shown in Table 4.9.

Language	Method	% Precision English Target		
		@1	@5	@10
Afrikaans	VECMAP	2.15	5.11	7.49
	Copy	10.68	-	-
	Proposed Method	9.42	21.80	27.41
	Proposed Method+Copy	19.16	30.15	35.17
Albanian	VECMAP	2.53	6.28	8.37
	Copy	5.62	-	-
	Proposed Method	7.93	15.20	18.61
	Proposed Method+Copy	13.11	19.49	22.58
Azerbaijani	VECMAP	0.99	1.74	2.48
	Copy	5.96	-	-
	Proposed Method	10.17	16.00	17.25
	Proposed Method+Copy	10.92	19.35	21.96
Bengali	VECMAP	1.18	2.99	3.92
	Copy	0.27	-	-
	Proposed Method	5.31	10.95	13.85
	Proposed Method+Copy	5.28	10.86	13.76
Bosnian	VECMAP	1.06	2.44	3.40
	Copy	21.23	-	-
	Proposed Method	8.17	15.71	18.58
	Proposed Method+Copy	29.19	35.88	38.11
Croatian	VECMAP	2.84	6.92	9.35
	Copy	4.35	-	-
	Proposed Method	11.86	24.70	30.13
	Proposed Method+Copy	15.65	28.02	33.21
Estonian	VECMAP	1.52	4.28	5.97
	Copy	7.56	-	-
	Proposed Method	8.15	18.79	23.66
	Proposed Method+Copy	14.93	24.65	29.15
Greek	VECMAP	3.44	8.20	11.07
	Copy	1.90	-	-
	Proposed Method	11.65	23.55	28.05
	Proposed Method+Copy	13.50	25.15	29.58
Hebrew	VECMAP	1.02	3.27	4.91
	Copy	11.15	-	-
	Proposed Method	8.18	17.08	20.55
	Proposed Method+Copy	19.02	26.89	29.75
Hindi	VECMAP	0.93	2.78	4.04
	Copy	0.06	-	-
	Proposed Method	4.57	11.64	15.39
	Proposed Method+Copy	4.60	11.66	15.41
Hungarian	VECMAP	1.50	4.45	6.25
	Copy	4.60	-	-
	Proposed Method	7.82	17.42	21.66
	Proposed Method+Copy	11.62	20.56	24.53
Turkish	VECMAP	1.10	3.57	5.23
	Copy	8.15	-	-
	Proposed Method	7.13	15.43	19.38
	Proposed Method+Copy	14.27	21.31	24.77
Average	VECMAP	1.69	4.34	6.04
	Copy	6.70	-	-
	Proposed Method	8.36	17.36	21.21
	Proposed Method+Copy	14.27	22.83	26.50

Table 4.10: Precision@ N for bilingual lexicon induction for the dataset of translation pairs with OOV source language words. The best precision@ N , for each language and methodology, is shown in boldface.

I compare the proposed method against two other methods that each handle OOV words differently. DUONG2016 is an extension of word2vec (Mikolov et al., 2013a), thus, it only works at the word level and cannot handle OOVs. Since DUONG2016 cannot form any representation for an OOV word, I employ a simple mechanism to handle OOVs. I copy the source language word exactly to the target language. This approach is referred to as the “Copy” baseline. This method works well especially when the OOVs are named entities and borrowings (Sennrich et al., 2016). For VECMAP, since I employ fastText to learn word representations, it also has sub-word embeddings, and I form an embedding for OOVs in the source language space and then transfer them to the cross-lingual shared space using the same approach proposed in Chapter 3. In the case of the proposed method, a representation of an OOV is formed from its sub-word representations in the target language. In this case, however, it not only has sub-word information, but this information is incorporated in the procedure of learning the shared space.

Table 4.10 shows the results. I find the same trend as for Chapter 3 in the results achieved by VECMAP. The precision is overall quite low, but this nevertheless shows that sub-word information can be transferred in a cross-lingual setting because this method outperforms a random baseline which achieves precision of zero. However, since the sub-words are not part of the training process to form the shared space, this method cannot provide very precise representations for OOV words in the shared space. In the case of the proposed model, since the sub-words are already in the shared

space and have been incorporated in the process of learning the shared space, the proposed method is able to better represent OOVs and the results are substantially better. Although there are some exceptions in which the copy baseline performs substantially better than the proposed method, which could be due to the test data containing many named entities, on average the proposed model not only outperforms the copy baseline but also outperforms VECMAP by a great margin.

Since in some cases the Copy method performed better than the proposed model, like in the case of *Afrikaans* and *Bosnian*, I also combine the proposed model with the Copy baseline to enhance its performance. To combine these two methods, I look into the target language matrix and if I find the source language word, I assume this word is a named entity or a borrowed word and use its exact form as its translation into the target language; otherwise, I return its translation as obtained by the proposed method.⁸ I refer to this method as “Proposed Method+Copy” in Table 4.10. Table 4.10 shows this mechanism improved the results substantially, except in the case of *Bengali*, in which there is a small reduction relative to the proposed method. Overall, the results demonstrate that if we employ methods that incorporate sub-words during training of cross-lingual embeddings, the learned shared space is better suited for low-resource and morphologically-rich languages and handling unseen words.

⁸This assumption can be incorrect, e.g., Afrikaans *kits* is in-vocabulary for English, but translates to English *moment*.

4.3.3.4 Summary

I evaluated the proposed method for learning cross-lingual embeddings that incorporates sub-word information during training, which could be well-suited to lower-resource and morphologically-rich languages because it can be trained on modest amounts of monolingual data and can represent OOVs. In two BLI tasks for twelve lower-resource languages focused on in-vocabulary words and OOVs, I found that since this method includes sub-word knowledge during training it is well-suited for languages that lack large amounts of training data. The results on the OOV test sets also indicate how vital it is to include sub-word information in the training phase to form representations for OOVs. In this chapter, I showed that incorporating sub-word information in the training process of learning cross-lingual word embeddings leads to forming high-quality cross-lingual embeddings. The experiments throughout this chapter confirmed the advantage of cross-lingual embeddings trained using our proposed method over baselines and state-of-the-art methods which do not use sub-word information for learning cross-lingual embeddings. The results specifically showed the importance of having sub-word information in the case of low-resource and morphologically-rich languages and how the proposed method can effectively represent OOVs, which is an important problem for low-resource and morphologically-rich languages, in the shared cross-lingual space for these languages.

Chapter 5

Conclusion

In this thesis, I investigated the importance of sub-word information in learning cross-lingual word embeddings especially in the case of low-resource and morphologically-rich languages. In this chapter, first, I summarize the contributions of this thesis, and then I revisit the research questions posed in Chapter 1 and explain how this thesis answered each of those questions. Finally, I discuss possible directions for future work.

5.1 Contributions

In this thesis, I have studied the effectiveness of sub-word knowledge in cross-lingual embeddings, and how it can be leveraged to overcome the out-of-vocabulary (OOV) word problem. What comes in the following is a breakdown of the contributions in this thesis.

- In Chapter 3, the effectiveness of sub-word embeddings in cross-lingual embeddings is shown by evaluating two types of sub-word representations, character n -grams and BPE, in three cross-lingual word representation methods with different degrees of supervision. In this evaluation, the case of a truly low-resource language, Cherokee, is considered in addition to five other languages from several language families. The results demonstrate that sub-word embeddings indeed provide information in the cross-lingual space, but that current methods are not able to fully use this information to overcome the OOV problem.
- A novel bilingual lexicon induction task is introduced in Chapter 3 to evaluate the OOV representations in the cross-lingual space. The ability of a model to represent OOVs is very crucial especially when the language is low-resource or morphologically-rich. Therefore, an evaluation method is proposed to measure the performance of the cross-lingual word embedding method to form embeddings for OOV words in the shared cross-lingual space.
- In Chapter 4, a novel method for learning cross-lingual word embeddings is proposed that incorporates sub-word information in the training process. This model only requires modest size monolingual corpora in the source and the target language and needs a bilingual dictionary as the cross-lingual signal. The resource requirements and the fact that the model learns sub-word embeddings while training the cross-lingual

embeddings makes it a good candidate to be employed for low-resource and morphologically-rich resource languages.

- In Chapter 4, three evaluation tasks are conducted to measure the performance of the proposed method, bilingual lexicon induction, monolingual word similarity, and zero-shot document classification. In all three tasks, the proposed method outperformed baselines and benchmark approaches.
- In order to demonstrate the impact of incorporating sub-word information during training on low-resource and morphologically-rich languages, in Chapter 4, two tasks are considered to evaluate the proposed method against several strong baselines on twelve low-resource languages. First, bilingual lexicon induction is used for in-vocabulary words and it is shown that incorporating sub-word information can indeed increase the performance. Second, the novel bilingual lexicon induction task for OOV words is considered to show the effectiveness of the proposed method on low-resource and morphologically-rich languages where OOVs are expected to be particularly frequent. The results clearly show the advantage of the proposed model over the established benchmarks which do not use sub-word information in their training.

5.2 Research questions revisited

Question 1: Can we leverage sub-word embeddings in cross-lingual models to address the OOV problem in the cross-lingual domain?

- In Chapter 3, it is shown how we can employ sub-word information with conventional cross-lingual word embeddings to overcome the OOV problem. It is demonstrated that even though the current methods do not use sub-word information while learning a cross-lingual shared space, it is possible to leverage sub-word information to represent OOVs in the shared space.
- It is also shown that not only can sub-word information be employed to form representations for OOV words in simulated low-resource scenarios, but also sub-word information is helpful when the target language is a truly low-resource language.

Question 2: Can cross-lingual word representations be improved by incorporating sub-word information in the process of training cross-lingual word representations and does it impact the performance in downstream tasks?

- In Chapter 4, a novel cross-lingual word embedding method is introduced that incorporates sub-word information during training. It is demonstrated that incorporating sub-word information during training leads to learning cross-lingual word embeddings of higher quality.
- A zero-shot document classification task is selected in order to show

the impact of incorporating sub-word information in downstream tasks. It is shown in Chapter 4 that the proposed model, which has knowledge of sub-words, outperforms a word-level model (Duong et al., 2016) and two strong contextualized language models, LASER and mBERT.

Question 3: Does the proposed method for learning cross-lingual embeddings by incorporating knowledge of sub-word information during training improve over benchmark approaches for learning cross-lingual embeddings on truly low-resource and morphologically-rich languages?

- I argued in Chapter 1 that it is essential to introduce a model that has sub-word information when the language of choice is a low-resource or morphologically-rich language. In Chapter 4, I compared the performance of the proposed model with two strong baselines on two bilingual lexicon induction tasks, one for in-vocabulary words and the other for OOV words, for twelve low-resource languages. The proposed model performs on par with the current-state-of-the-art on in-vocabulary words, and outperforms the baselines and a benchmark approach significantly on OOV words.

5.3 Future Work

First, I would like to employ the cross-lingual word embeddings learned by the proposed method in other downstream tasks, focusing on low-resource languages. In this thesis, I demonstrated the impact of the proposed model

on document classification. It would be interesting to investigate its impact on other NLP applications, such as part-of-speech tagging and named entity recognition. These two tasks have been addressed in the literature in the cross-lingual space (e.g., Duong et al., 2013; Garrette et al., 2013; Wu and Dredze, 2019), including for some low-resource languages (e.g., Duong et al., 2014; Cotterell and Duh, 2017). One particular challenge related to low-resource languages is the absence of a gold-standard dataset to evaluate the model. Thus, before performing any evaluation, it is first needed to build a gold-standard dataset that itself is a very valuable resource. The universal dependencies project (Nivre et al., 2020) contains POS tagged data for many languages including some low-resource language. It would be possible to extend it further and include more low-resource languages.

Second, the proposed method in this thesis does not consider the position of the words in the context vector. Another direction for the future would be to also consider the position of each word in the context vector while learning the cross-lingual embeddings. In this way, the words closer to the target will contribute more to the meaning of the target word, and the words at the far ends of the context window will have little impact on the target word meaning. This addition is shown to be effective in the monolingual space (Grave et al., 2018), and it possibly could also be useful in the cross-lingual space.

Third, I would like to consider other methods for providing sub-word information. In the proposed method, I employ character n -grams; however,

there are other methods that break a word into pieces differently, with byte-pair-encoding (BPE) being one such method. BPE breaks a word into the most frequent consecutive sequences of characters. Although it is shown in Chapter 3 that BPE is not a good choice for low-resource scenarios and representation of OOVs, it should be taken into account that BPE was not incorporated during the training and it could be possible to achieve better performance by considering BPE while learning the cross-lingual shared space. One interesting evaluation would be to evaluate these two variants, BPE and character n -grams, on morphologically-rich languages and observe which form of sub-word information is more suitable to learn cross-lingual word embeddings when working with a morphologically-rich language. Another interesting line of research is to use existing finite-state morphological analyzers (eg., Kessikbayeva and Cicekli, 2014; Seiss, 2012) instead of employing unsupervised methods for providing sub-word information. A finite-state morphological analyzer is a rule-based analyzer which can split a word into its meaningful units, e.g., a root with one or more affixes (prefix, suffix, infix), in contrast to character n -grams or BPEs that do not have any linguistic knowledge of morphemes and break a word into character sequences of a pre-determined length, in the case of character n -grams, or most frequent sequences of characters, in the case of BPE. This could potentially lead to improvements since we would learn embeddings for morphemes and we can simply add them together to form a vector representation for a word consisting of multiple morphemes.

Fourth, with the advancements in contextualized language models (e.g., Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019), they are replacing the classical word embeddings methods, e.g., word2vec, in many NLP tasks such as POS tagging (Wu and Dredze, 2019), document classification (Conneau and Lample, 2019) and named entity recognition (Wu and Dredze, 2019). However, these models require a large training corpus and are very computationally expensive to train. Due to these reasons and since there is a lack of training resources for low-resource languages, they are still not applicable to low-resource languages. Even though there has been some research on transferring knowledge for contextualized language models to other languages (e.g., Conneau et al., 2020; Artetxe et al., 2020a), there is still room for further work. Research on this topic could range from transfer learning, e.g., making the current pre-trained language models suitable for another language, to discovering a new training procedure, so these contextualized language models can be trained with less training data.

Bibliography

- Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.
- Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.
- Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

- Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.
- Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.
- Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.
- Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework

- of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, New Orleans, Louisiana.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020a. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020b. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Coling*

- 2010: *Demonstrations*, pages 37–40, Beijing, China. Coling 2010 Organizing Committee.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134, Mumbai, India. The COLING 2012 Organizing Committee.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Fabienne Braune, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2018. Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193, New Orleans, Louisiana. Association for Computational Linguistics.
- Statistics Canada. 2018. Linguistic characteristics of Canadians.

<https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm> [Accessed: 2019-03-25].

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Gobinda G Chowdhury. 2010. *Introduction to modern information retrieval*. Facet publishing.

Christopher Cieri, 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'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning, ICML '08*, pages 160–167, Helsinki, Finland. ACM.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model

pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhaloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics.

- Hope Dawson, Michael Phelan, et al. 2016. *Language files: Materials for an introduction to language and linguistics*. The Ohio State University Press.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Long Duong. 2017. *Natural language processing for resource-poor languages*. Ph.D. thesis, The University of Melbourne.
- Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China. Association for Computational Linguistics.

- Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. A neural network model for low-resource Universal Dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348, Lisbon, Portugal. Association for Computational Linguistics.
- Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Doha, Qatar. Association for Computational Linguistics.
- Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 634–639, Sofia, Bulgaria. Association for Computational Linguistics.
- Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.
- Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceed-*

ings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 894–904, Valencia, Spain. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

Jakob Elming, Barbara Plank, and Dirk Hovy. 2014. Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, Baltimore, Maryland. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Manaal Faruqui 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*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of POS-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592, Sofia, Bulgaria. Association for Computational Linguistics.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.
- Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.
- Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

- Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia. Association for Computational Linguistics.
- Ali Hakimi Parizi and Paul Cook. 2020a. Evaluating sub-word embeddings in cross-lingual models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2712–2719, Marseille, France. European Language Resources Association.
- Ali Hakimi Parizi and Paul Cook. 2020b. Joint training for learning cross-lingual embeddings with sub-word information without parallel corpora. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM 2020)*, Barcelona, Spain. Association for Computational Linguistics.
- Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 619–624, Valencia, Spain. Association for Computational Linguistics.
- Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

- Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland. Association for Computational Linguistics.
- Thai Hoang and Phuong Vu. 2020. Not-NUTs at WNUT-2020 task 2: A BERT-based system in identifying informative COVID-19 English tweets. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 466–470, Online. Association for Computational Linguistics.
- Yuting Hu and Suzan Verberne. 2020. Named entity recognition for Chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.
- Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Eval-*

uation Conference, pages 2562–2572, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Gulshat Kessikbayeva and Ilyas Cicekli. 2014. Rule based morphological analyzer of Kazakh language. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 46–54, Baltimore, Maryland. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for Computational Linguistics.

- Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.
- Young-Bum Kim, Benjamin Snyder, and Ruhi Sarikaya. 2015. Part-of-speech taggers for low-resource languages using CCA features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Lisbon, Portugal. Association for Computational Linguistics.
- Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.
- Guokun Lai, Barlas Oguz, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. *arXiv preprint arXiv:1909.07009*.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

- Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Beijing, China. PMLR.
- David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.
- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.
- Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceed-*

ings of the Seventeenth Conference on Computational Natural Language Learning, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Low-resource parsing with crosslingual contextualized representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 304–315, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and

- Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.
- Mayur Patidar, Surabhi Kumari, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, Lovekesh Vig, and Gautam Shroff. 2019. From monolingual to multilingual FAQ assistant using multilingual co-training. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 115–123, Hong Kong, China. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized

- word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94, Denver, Colorado. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Multilingual NER transfer for low-resource languages. *CoRR*, abs/1902.00193.

- Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Sofia, Bulgaria. Association for Computational Linguistics.
- Sebastian Ruder, Ivan Vulniundefined, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.
- Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Michael Schlichtkrull and Anders Søgaard. 2017. Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 220–229, Valencia, Spain. Association for Computational Linguistics.

- Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.
- Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Melanie Seiss. 2012. A rule-based morphological analyzer for murrinh-patha. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 751–758, Istanbul, Turkey. European Language Resources Association (ELRA).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Anders Søgaard, 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), pages 1713–1722, Beijing, China. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22.

Michael Ustaszewski. 2019. Optimising the europarl corpus for translation studies with the europarlextract toolkit. *Perspectives*, 27(1):107–123.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

- Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.
- Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.
- Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia. Association for Computational Linguistics.
- Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Beijing, China. Association for Computational Linguistics.
- Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-

- lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.
- Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, Michigan. Association for Computational Linguistics.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber. 2020. Why overfitting isn’t always bad: Retrofitting cross-lingual word

- embeddings to dictionaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Online. Association for Computational Linguistics.
- Yi Zhu, Benjamin Heinzerling, Ivan Vulić, Michael Strube, Roi Reichart, and Anna Korhonen. 2019a. On the importance of subword information for morphological tasks in truly low-resource languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.
- Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019b. A systematic study of leveraging subword information for learning word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 912–932, Minneapolis, Minnesota. Association for Computational Linguistics.
- Will Y. Zou, 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*, pages 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics.

Vita

Candidate's full name: Ali Hakimi Parizi

University attended (with dates and degrees obtained):

- Sept. 2017 - Present Ph.D. candidate, University of New Brunswick
- 2013 - 2015 MSc in Computer Engineering, Razi University
- 2006 - 2012 BSc in Computer Engineering, University of Sistan and Baluchestan

Selected Publications:

- **Ali Hakimi Parizi** and Paul Cook. “Joint Training for learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora”. 2020, In Proceedings of the 9th Joint Conference on Lexical and Computational Semantics, Barcelona, Spain
- **Ali Hakimi Parizi** and Paul Cook. “Evaluating sub-word embeddings in cross-lingual models”. 2020, In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France
- **Ali Hakimi Parizi**, Milton King and Paul Cook. “Using Language Models to Detect Hate Speech and Offensive Language”. 2019, In Proceedings of the 13th International Workshop on Semantic Evaluation
- **Ali Hakimi Parizi** and Paul Cook. “Do Character-Level Neural Network Language Models Capture Knowledge of Multiword Expression Compositionality?”. 2018, In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions
- Milton King, **Ali Hakimi Parizi** and Paul Cook. “Evaluating unsupervised approaches to capturing discriminative attributes”. 2018, In Proceedings of the 12th International Workshop on Semantic Evaluation

Selected Conference Presentations:

- **Ali Hakimi Parizi** and Paul Cook. “Joint Training for learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora”. 2020, In Proceedings of the 9th Joint Conference on Lexical and Computational Semantics, Barcelona, Spain
- **Ali Hakimi Parizi** and Paul Cook. (2019) “Evaluation of Cross-Lingual Sub-word Embeddings”. 43rd Annual Meeting of the Atlantic Provinces Linguistic Association, Fredericton, Canada