Joint Training for Learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora

Ali Hakimi Parizi
Faculty of Computer Science
University of New Brunswick
ahakimi@unb.ca

Paul Cook
Faculty of Computer Science
University of New Brunswick
paul.cook@unb.ca

Abstract

In this paper, we propose a novel method for learning cross-lingual word embeddings, that incorporates sub-word information during training, and is able to learn high-quality embeddings from modest amounts of monolingual data and a bilingual lexicon. This method could be particularly well-suited to learning cross-lingual embeddings for lower-resource, morphologically-rich languages, enabling knowledge to be transferred from rich- to lower-resource languages. We evaluate our proposed approach simulating lower-resource languages for bilingual lexicon induction, monolingual word similarity, and document classification. Our results indicate that incorporating sub-word information indeed leads to improvements, and in the case of document classification, performance better than, or on par with, strong benchmark approaches.

1 Introduction

State-of-the-art approaches in natural language processing (NLP) typically require a substantial amount of human-annotated data (i.e, for supervised approaches to tasks such as part-of-speech tagging or dependency parsing) or they need a very large amount of unannotated text for training (e.g., methods for learning word embeddings). This poses a particular problem for building NLP systems for low-resource languages. There are thousands of human languages, and creating annotated datasets for all of them would be very expensive. Furthermore, many languages have a relatively small number of speakers, and in many cases large amounts of text are not readily-available for building corpora for these languages. A further related challenge is posed by morphologically-rich languages, because many word-forms would not be expected to be observed in a training corpus. One way to address these problems is to transfer knowledge from a rich-resource language to a lower-resource language.

Word Embeddings are a key feature in approaches for a wide range of NLP tasks, such as part-of-speech tagging (Al-Rfou' et al., 2013), dependency parsing (Chen and Manning, 2014), and named entity recognition (Pennington et al., 2014). If we are able to transfer the knowledge captured in word embeddings for a rich-resource language to another low-resource language, then developing NLP tools could become more feasible for the low-resource language. There has therefore been a wealth of research on cross-lingual word embeddings (e.g., Mikolov et al., 2013b; Vulić and Moens, 2016; Lample et al., 2018), in which embeddings for multiple languages are learned in a shared space, and which can be used to transfer knowledge between languages, such as from a rich-resource language to a low-resource one (Ruder et al., 2019).

Despite the wide range of research on learning cross-lingual embeddings, there are some limitations of these methods that have not been addressed. In the case of a low-resource language, due to the relatively small size of available corpora, a relatively small number of embeddings would be learned. Moreover, in the case of a morphologically-rich language, many wordforms would not be observed in the corpus on which embeddings are trained. As a result, given a subsequent text to process, many words would be expected to be out-of-vocabulary (OOV) with respect to the embedding model. This is a very important issue, because in the case of OOVs, we do not have an embedding for these words, and models for downstream NLP tasks that use embeddings would therefore lack information for these words. Where

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 39–49
Barcelona, Spain (Online), December 12–13, 2020

the number of OOVs is relatively high, such as for low-resource and morphologically-rich languages, this could lead to particularly poor performance in down-stream tasks. This problem has been addressed in monolingual settings by learning embeddings for sub-word units, and then composing representations for OOVs based on their sub-words (Bojanowski et al., 2017; Zhu et al., 2019).

Recently, with advances in language modelling (Artetxe and Schwenk, 2019) and contextualized language models (Devlin et al., 2019; Conneau and Lample, 2019), transfer learning has become feasible between languages by using a byte pair encoding (BPE, Sennrich et al., 2016) shared vocabulary, and fine-tuning the models for specific tasks (Wu and Dredze, 2019). Nevertheless, these models require a substantial amount of training data (Conneau and Lample, 2019), and in some cases parallel corpora (Artetxe and Schwenk, 2019; Conneau and Lample, 2019), and are very computationally expensive to train. There is therefore a need for methods that can be trained from a more-limited amount of data and require less computational resources for training, but that nevertheless show comparable performance.

In this paper, we propose a model that can learn cross-lingual word embeddings from a modest amount of monolingual data and a bilingual dictionary. We 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. Our 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. We 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, we form a shared embedding space 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.

To evaluate our proposed model, we use modest amounts of data for relatively well-resourced languages. We first consider two intrinsic evaluations: (1) the widely-considered task of bilingual lexicon induction (BLI), and (2) a monolingual word similarity task to show the effectiveness of our proposed approach when the embeddings are used in a monolingual setting. Our results on these tasks demonstrate that incorporating sub-word information leads to improvements for both cross-lingual and monolingual representations. For extrinsic evaluation, to show the impact of having sub-word knowledge in a down-stream NLP task, we consider cross-lingual document classification. Again our results indicate that incorporating sub-word information leads to improvements, and furthermore we find our proposed model to perform on par with, or better than, strong benchmark approaches.

2 Related Work

A variety of methods have been proposed for learning cross-lingual word embeddings. These methods vary with respect to the level of supervision, and the cross-lingual signals used, such as parallel corpora and bilingual dictionaries.

Klementiev et al. (2012) propose a method to learn cross-lingual representations by training a language model on the source and target language and optimizing their objective function jointly. This method, however, requires a parallel corpus, which is not available for many languages, especially low-resource ones. More recently, 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. Aside from requiring a parallel corpus, it is also computationally expensive to train.

Mikolov et al. (2013b) argue that the geometric arrangement of word embeddings in two different languages is the same. They therefore propose a method to learn a linear transformation to align the vector spaces of two languages by using a seed lexicon of known translation pairs. Xing et al. (2015) show that normalizing all word vectors to be unit length, and applying an orthogonality constraint on the transformation matrix, improves the approach of Mikolov et al. Artetxe et al. (2017) introduce an alignment-based method which relaxes the requirement of having a bilingual seed lexicon. Their approach begins with a very small seed lexicon — as few as 25 pairs — and in a process of self-learning and through several rounds of bootstrapping, increases the size of the bilingual dictionary. Artetxe et al. (2018b) further relax the need for a bilingual

dictionary, and propose a fully unsupervised approach. Their method solves the same mapping problem as Artetxe et al. (2017), but creates the initial seed lexicon in an unsupervised manner by exploiting the similarity distribution of words in the source and target language.

All of these mapping-based methods require pre-trained monolingual word embeddings, the quality of which the final cross-lingual word embeddings are greatly dependent upon. This is problematic in the case that we do not have access to enough training data to learn high quality monolingual embeddings, as would be the case for many 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). Furthermore, 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. They show that methods which jointly learn the embedding space for the source and target language from a parallel corpus are superior to mapping-based methods. However, parallel corpora are a very expensive cross-lingual signal.

In an alternative approach to learning cross-lingual word embeddings, a pesudo-bilingual corpus is first constructed using a bilingual dictionary, and embeddings for the source and target language are then learned from this corpus. Gouws and Søgaard (2015) concatenate and shuffle the source and target language corpora, and then randomly replace words in this corpus using a bilingual dictionary. They then run CBOW on the constructed corpus to learn word embeddings for both the source and target language. 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. Compared to mapping-based methods, this approach does not require as large of a corpus for training, because for each word, the context in not only the source language, but also the target language, is used. However, these pseudo-bilingual corpus methods are more expensive to train than mapping-based methods, because the embeddings are learned from scratch, in contrast to mapping-based methods which use pre-trained embeddings and only need to learn the mapping function.

Recent approaches to learning cross-lingual embeddings have been trained on concatenated monolingual corpora. Multilingual BERT (mBERT) is a BERT model (Devlin et al., 2019) 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, but is computationally expensive to train. Chaudhary et al. (2018) present a method that uses sub-word information, such as lemmas, morpheme tags, and phoneme n-grams, to transfer knowledge from rich-resource languages to low-resource ones. They train skip-gram on concatenated monolingual corpora of two related languages and learn representations in a shared space by relying on similar sub-words to map related words close to each other in the shared space. They also consider an approach which first trains a model on the rich-resource language and then uses the trained sub-word embeddings to initialize the model for the low-resource language.

The approach for learning cross-lingual embeddings proposed in this paper incorporates sub-word information — similar to Chaudhary et al. (2018) and mBert — but in contrast to Chaudhary et al. does not require language-specific tools such as morphological analyzers which might not be available for low-resource languages, and in contrast to mBert is less computationally-expensive to train. The proposed approach is an extension of Duong et al. (2016) that incorporates sub-word information, and requires only modest size monolingual corpora and a bilingual lexicon for training.

3 Methodology

In this section we first describe the approach of Duong et al. (2016) to learning cross-lingual word embeddings, and then present our proposed model, which is an extension of this approach.

3.1 Learning Cross-lingual Word Embeddings 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 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. The CBOW objective function, shown below, is only capable of capturing monolingual similarities:

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

Equation 2 is therefore proposed to adapt it to cross-lingual settings:

$$O = \sum_{i \in D_s \cup D_t} (\alpha \log \sigma(u_{w_i}^T h_i) + (1-\alpha) \log \sigma(u_{\bar{w}_i}^T h_i) + \sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i)) \quad (2)$$

where h_i encodes the context vector, $\bar{w}_i$ is the translation of w_i, α is a weight parameter, and D_s and D_t are the source and target language vocabularies, respectively.

Duong et al. (2016) also propose an approach to find the best translation for polysemous words using the expectation maximization algorithm and cosine similarity between the context vector — the average of the embeddings for the words in the context — and possible translations. Thus the translation for a word is selected based on its context.

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 context words are pushed closer together in V space, while the target word and its translation become closer in U space and further away from the negative samples. Duong et al. 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 as shown in Equation 3.

$$O' = O + \delta \sum_{w \in V_s \cup V_t} \| u_w - v_w \|_2^2 \quad (3)$$

For the remainder of the paper we refer to this approach as DUONG2016.

3.2 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. In this paper, we extend DUONG2016 by incorporating sub-word information during training.

To incorporate sub-word information, we follow a similar approach to Bojanowski et al. (2017). 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 our experiments we consider sequences of length 3–6 characters. We 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:

$$O = \sum_{i \in D_s \cup D_t} (\alpha \log S(w_i, c) + (1-\alpha) \log S(\bar{w}_i, c) + \sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log -S(w_j, c)) \quad (4)$$

where c is the context. S, shown in Equation 5, measures the similarity between a word and context, taking into account sub-words:

$$S(w, c) = \frac{1}{|G_w|} \sum_{g \in G_w} z_g^T v_c \quad (5)$$

Language	Family	# Tokens	# Types	# Embeddings	# Dict. entries
Chinese	Sino-Tibetan	30M (64%)	0.2M (20%)	86K (43%)	1983K
Dutch	Germanic	84M (64%)	1.3M (8%)	303K (28%)	406K
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 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, is also shown for each language. The parenthetical numbers indicate coverage in the dictionary.

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, we average representations for each word appearing in c, where each word is represented by the average of its sub-word embeddings.[1]

4 Resources

For evaluation, we simulate lower-resource languages using 9 well-resourced languages: Chinese, Dutch, English, French, German, Italian, Japanese, Russian, and Spanish. These languages include those considered by Duong et al. (2016), as well as those in the MLDoc dataset (Schwenk and Li, 2018, which we use for evaluation in Section 5.3). Following previous work (e.g., Duong et al., 2016; Lample et al., 2018), we only consider pairs of languages with English as either the source or target language, and one of the remaining 8 languages as the other language.

To train word embeddings for each language, we 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), we randomly select 5 million sentences for each language from their Wikipedia dump. Table 1 shows the number of tokens and types in each corpus.

We use a bilingual dictionary as the cross-lingual signal in our proposed approach. Our study builds on the work of Duong et al. (2016), and so for languages that they consider — Dutch, German, Italian, Japanese, and Spanish — we use the same dictionaries that they did, which were extracted from Panlex.[2] For Chinese, French, and Russian we extract dictionaries from Panlex using a similar approach to Duong et al.

Table 1 also shows the size of each dictionary, with English as the source language, and the other language as the target language.[3] 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.

5 Experimental Results

We present experimental results for two intrinsic evaluations, bilingual lexicon induction and monolingual word similarity, and an extrinsic evaluation on cross-lingual document classification.

5.1 Bilingual Lexicon Induction

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, we try to find the translation of a source language word in the target language by looking at its nearest neighbours. Ideally, a word and its translation would be located close to each other in the shared cross-lingual word embedding space. Here we focus on comparing our proposed method with DUONG2016 and so consider the same four languages as Duong et al. (2016): English, Dutch, Italian, and Spanish. In all cases, English is the target language and the other languages are treated as the source language.

[1] This differs from fastText which sums the sub-word embeddings.

[2] https://github.com/longdt219/XlingualEmb

[3] The dictionary size for English is therefore not shown.

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
Our Model ($c = 48$, $d = 200$)	60.15	79.84	84.26	54.62	73.83	78.92	42.25	67.39	72.80
Our Model ($c = 5$, $d = 200$)	41.39	78.63	85.06	36.21	72.42	79.45	36.54	69.15	76.25
Our Model ($c = 20$, $d = 200$)	59.14	83.12	87.27	54.02	77.64	82.00	47.56	73.00	78.69
Our 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 2: Precision@N for bilingual lexicon induction. The best performance, for each dataset and evaluation measure, is shown in boldface.

Following previous work (e.g. Lample et al., 2018; Joulin et al., 2018; Jawanpuria et al., 2019), we consider MUSE test sets for evaluation. Word pairs occurring in both the MUSE test sets and our training dictionaries are removed from the training data before training the embeddings. We 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 top-N most similar target language words (Ruder et al., 2019). We use cosine as the similarity measure.

Results are shown in Table 2. We begin by considering DUONG2016 and our 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 embedding size (d) of 200, sub-sampling of $1e^{-4}$, α of 0.5, and δ set to 0.01.[4] In terms of precision@1, our model outperforms 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 our 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. We 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 our model For both models, a window size of 5 performs relatively poorly. For DUONG2016, the original window size of 48 performs best in terms of precision@1 for Spanish and Italian, but not Dutch. For our 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 we consider increasing the embedding size to 300, which is commonly used for fastText (Bojanowski et al., 2017). We consider this for the best window size for each model, i.e., 48 for DUONG2016 and 20 for our model.[5] Our 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 our 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, we 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. Our model, on the other hand, is trained on substantially smaller corpora because we focus on approaches that could be applied to lower-resource languages. mBERT and Chaudhary et al. (2018) are further points of comparison that we do not include because of the resource requirements, and reliance

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

[5] We also considered a window size of 20 and embedding size of 300 for DUONG2016, but this did not give improvements.

Model	WS-en	WS-de	RW-en	RW-en+OOV
DUONG2016	74.46	69.72	44.06	37.68
Our Model	**75.67**	**70.49**	**51.57**	**49.51**
trained on 5 million sentences				
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
trained on 10 million sentences				
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
trained on full Wikipedia corpora				
fastText (skipgram, $c = 5$, $d = 300$)	73.77	66.63	48.61	48.09

Table 3: Spearman's correlation for monolingual similarity on each dataset, for each method considered. The best performance on each dataset is shown in boldface.

on language-specific tools, respectively, of these methods.

For the rest of the paper, "our 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, we continue to use the best parameter settings from Duong et al. (2016) for this method.

5.2 Monolingual Word Similarity

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

Following Duong et al. (2016), we consider English and German for these experiments. We 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). We use cosine as the similarity score. 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, we therefore report results only for in-vocabulary items. For RW-en, however, roughly 25% of the test pairs include an OOV. For this dataset we 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 we assign these test pairs the average cosine similarity score over test pairs that are in-vocabulary.

Table 3 shows the results. For each dataset, our 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.

5.3 Document Classification

Here we consider an extrinsic evaluation which uses cross-lingual word embeddings in a downstream task, specifically cross-lingual document

Model	Target language							
	Chinese	French	German	Italian	Japanese	Russian	Spanish	Average
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
Our 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: 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.

classification. This task is motivated by the situation where sufficient labelled training data is not available for a low-resource language. We consider zero-shot classification, i.e., we 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.

Following previous work (e.g., Artetxe and Schwenk, 2019; Wu and Dredze, 2019), we 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 (2019), we use English as the source language, and the other languages as target languages.

To build corpora to train embeddings, again following previous work (Duong et al., 2016; Klementiev et al., 2012), we first randomly sample $400k$ sentences for each of the source and target language from RCV1/RCV2,[6] and then combine these in-domain corpora with larger Wikipedia corpora. We use the Wikipedia corpora described in Section 4.

We represent documents as the average of their words' embeddings, where the embeddings are learned by our proposed approach from the corpora described above. We 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.

We compare our approach against several benchmarks. First we consider the same approach described above, but using embeddings from DUONG2016 instead of our proposed approach. In this case, embeddings for OOVs are not available, and so OOVs are simply ignored in forming document representations. We further consider two strong benchmark approaches — LASER (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] We use these improved results for comparison. We use mBERT results reported by Wu and Dredze (2019).

Results are shown in Table 4. None of the approaches considered performs best for all languages. However, in terms of the average accuracy over all target languages, our proposed model performs better than DUONG2016, LASER and mBERT. It is worth noting that our model is trained on only 5.4 million sentences in each language, and does not require a parallel corpus. Artetxe and Schwenk (2019), the next best method in terms of average accuracy, on the other hand, is trained on 225 million parallel sentences. Furthermore, our 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 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 our proposed model on the other

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

[7] https://github.com/facebookresearch/LASER/tree/master/tasks/mldoc

languages, but also requires a large parallel corpus for training.

6 Conclusions

In this paper we 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. This approach could be particularly well-suited to lower-resource, morphologically-rich languages, for which large parallel corpora are not available.

We evaluated our proposed approach, on a variety of simulated lower-resource languages, for the tasks of BLI, monolingual word similarity, and document classification. Our 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. For zero-shot cross-lingual document classification, incorporating sub-word information again led to improvements, and our proposed model outperformed benchmark approaches that have substantially higher resource requirements for training. Code and data to reproduce these results has been made available.[8]

In future work, we plan to evaluate our proposed approach on truly lower-resource languages to determine the impact of smaller training corpora and bilingual dictionaries on the performance of cross-lingual word embeddings. It would also be interesting to consider the morphological richness of languages in this analysis. We further intend to investigate using alternative approaches to forming sub-word representations, such as byte-pair encoding, as well as incorporating positional embeddings into our model (e.g., Grave et al., 2018), to determine their impact on the quality of the resulting cross-lingual embeddings. Finally we plan to evaluate our proposed approach on further extrinsic tasks, such as POS tagging and named entity recognition, focusing on lower-resource languages.

Acknowledgments

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the New Brunswick Innovation Foundation, and the University of New Brunswick. This research was enabled in part by support provided by ACENET (https://www.ace-net.ca/) and Compute Canada (www.computecanada.ca).

[8] https://github.com/Cons13411/XLing_Subword

References

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.

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.

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 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.

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.

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.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

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.

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.

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

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).

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.

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.

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.

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*.

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.

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 *Proceedings 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.

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.

Sebastian Ruder, Ivan Vuliundefined, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

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).

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, 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.

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 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. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

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.

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.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. 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.