Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2712–2719
Marseille, 11–16 May 2020

Evaluating Sub-word embeddings in cross-lingual models

Ali Hakimi Parizi, Paul Cook
University of New Brunswick
Fredericton, New Brunswick, Canada
{ahakimi, paul.cook,}@unb.ca

Abstract

Cross-lingual word embeddings create a shared space for embeddings in two languages, and enable knowledge to be transferred between languages for tasks such as bilingual lexicon induction. One problem, however, is out-of-vocabulary (OOV) words, for which no embeddings are available. This is particularly problematic for low-resource and morphologically-rich languages, which often have relatively high OOV rates. Approaches to learning sub-word embeddings have been proposed to address the problem of OOV words, but most prior work has not considered sub-word embeddings in cross-lingual models. In this paper, we consider whether sub-word embeddings can be leveraged to form cross-lingual embeddings for OOV words. Specifically, we consider a novel bilingual lexicon induction task focused on OOV words, for language pairs covering several language families. Our results indicate that cross-lingual representations for OOV words can indeed be formed from sub-word embeddings, including in the case of a truly low-resource morphologically-rich language.

Keywords: Cross-lingual Word Embeddings, Low-resource Languages, Morphologically-rich Languages

1. Introduction

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., 2015). 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. Sub-word level embeddings (Bojanowski et al., 2017, e.g.) — 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) — has considered sub-word embeddings in cross-lingual models. In this paper we evaluate whether sub-word embeddings can be leveraged in cross-lingual models. Specifically, we 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. Our findings indicate that sub-word embeddings do carry information that can be leveraged in cross-lingual models.

2. Related Work

A variety of approaches have been proposed to find cross-lingual embedding spaces. Some of these methods require very expensive signals for training, such as sentence alignments (Zou et al., 2013, e.g.) or comparable documents (Vulić and Moens, 2016, e.g.). Large amounts of such data might not be available for many languages, particularly low-resource languages. We therefore focus on methods that require only monolingual corpora and a bilingual dictionary.

Mikolov et al. (2013) 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 the following optimization problem, we get a transformation matrix, W:

$$\min_{\mathrm{W}} ||\mathrm{AW} - \mathrm{B}||_F \quad (1)$$

where F is the Frobenius norm. The transformation matrix maps the vectors of language A to the vector space of language B. Moreover, most methods which employ a dictionary as their cross-lingual training signal need only a relatively small number of training seeds to find the mapping between two languages (Mikolov et al., 2013; Hauer et al., 2017; Duong et al., 2016), so the bilingual dictionary need not be very large. Xing et al. (2015) show that enforcing an orthogonality constraint on W gives improvements. Artetxe et al. (2017) propose a method that can work with a small seed lexicon, as low as 25 pairs. They solve the same optimization problem as Mikolov et al. (2013), and in a process of self-learning and in several rounds of bootstrapping add more translation pairs to the bilingual dictionary. In another work, Artetxe et al. (2018) 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.

Braune et al. (2018) consider an English–German lexicon induction task focused on low frequency, in-

vocabulary words. They show that sub-word embeddings and orthographic similarity can be incorporated to give improvements. In contrast to Braune et al. (2018), we consider OOVs, not low frequency words, and a greater number of languages.

3. Corpora

In this paper, we consider the same languages that Adams et al. (2017) used in their experiments on applying cross-lingual word embeddings for low-resource language modelling; specifically, we 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 The corpus for each language is a Wikipedia dump from 20 September 2018, except for Japanese, where we 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 we use WP2TXT.[1] Details of the corpora for each language are provided in Table 1, in the "Full corpus" columns.

In preliminary experiments we observed that for the full Wikipedia corpora, relatively few words in the evaluation dataset (discussed in Section 4.) were OOVs, yet OOVs are required for our experimental setup. Therefore, following Adams et al. (2017) we carried out experiments in which we learned cross-lingual embeddings, but down-sized the size of the corpora. We found that bilingual lexicon induction for in-vocabulary items performed with reasonably high accuracy down to corpora of roughly 100M tokens. We 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 1, in the "Sample" columns.

In the bilingual lexicon induction experiments in this paper we attempt to find an in-vocabulary target language translation for an OOV source language word. We 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 substantial number of gold-standard translations will be OOV in the source language, and to simulate a lower-resource source language.

4. 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 1B translations. We use Panlex to build gold-standard evaluation data.

In our experiments we 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). We begin by extracting all single-word translations from Panlex for these language pairs. For each language pair, we then create a gold-standard evaluation dataset by keeping only those translations for which the source language word is not in 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). We observed that some translations in Panlex appear to be noisy.[2] We therefore further eliminated any translation for which the source language word does not appear in the aspell dictionary for that language.[3] Details of the evaluation datasets are shown in Table 2.

5. Results

In this work we use three approaches, a supervised, a semi-supervised and an unsupervised method, to learn a transformation matrix from monolingual embeddings to have a comprehensive comparison between methods with various degrees of supervision. For the supervised method we use a publicly-available implementation of Conneau et al. (2018).[4] For the semi-supervised and fully-unsupervised methods, we use publicly available implementations of the approaches of (Artetxe et al., 2017) and (Artetxe et al., 2018), respectively.[5] For all three models, we use their default settings to learn the transformation matrix.

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, we use fastText (Bojanowski et al., 2017) to create an embedding matrix for each corpus. We use the default fastText parameters, except for the number of dimensions for the embeddings, which is set to 300.

For the second approach, we use the method introduced by Zhu et al. (2019), 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. We use byte pair encoding (BPE) (Sennrich et al., 2016) as the method which provides sub-word information. To train word embeddings using the Zhu et al. (2019) framework, we 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. We refer to this approach — which is based on Zhu et al. (2019) and incorporates BPE — as BPE.

Results are presented in the following subsections. In subsection 5.1., the results for bilingual lexicon induction for OOVs are presented using the various approaches to representing OOVs and learning the transformation matrix. Subsection 5.2. describes a further experiment, in which the test data consists of both in-vocabulary and OOV

[1] https://github.com/yohasebe/wp2txt

[2] For example, some English entries consist of no Latin letters and appear to be non-English words, while other entries consist entirely of non-alphabetic symbols.

[3] http://aspell.net/

[4] https://github.com/facebookresearch/MUSE

[5] https://github.com/artetxem/vecmap

		Full corpus (target language)			Sample (source language)		
Language	Family	#Tokens	#Types	#Embeddings	#Tokens	#Types	#Embeddings
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 1: The size of the full corpus, and sample, for each language, in terms of the number of tokens, types, and resulting embeddings. Language families are also shown.

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

Table 2: The number of pairs in each test set, for each language, with English as both the source and target language.

source language words (as opposed to only OOVs). In subsection 5.3. we 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 a truly low-resource source language, specifically Cherokee.

5.1. OOV Bilingual Lexicon Induction

In the case of the supervised method, given source and target language embeddings, we require a set of translations to learn the transformation matrix W in Equation 1. Following previous work (Conneau et al., 2018; Joulin et al., 2018; Jawanpuria et al., 2019, e.g.), we use the training pairs provided by Conneau et al. (2018).[6] For training the semi-supervised method, we take a random sample of 25 pairs from these training pairs. Given a gold-standard evaluation pair, we construct a representation for its (OOV) source language word by averaging its sub-word embeddings using fastText or BPE. We 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. We report accuracy@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.

We compare against two baselines. We consider a random baseline, which randomly ranks the target language words for a given source language word. We 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. We 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, we only calculate accuracy@1 for this method.[7]

Results are shown in Table 3. For all languages and translation directions, when the source of sub-word information is fastText, accuracy@1 is higher using the supervised, semi-supervised and unsupervised methods than using the copy baseline, except for accuracy@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 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 finding 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 accuracy@N, with the only difference being the source of sub-word information, fastText always outperforms BPE. Zhu et al. (2019) noted that BPE is not effective for dealing with OOVs, and we observe the same here.

Focusing on approaches using fastText, and considering the differing levels of supervision, we 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. We return to consider the level of supervision in Section 5.4. when we

[6] https://github.com/facebookresearch/MUSE

[7] Razmara 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 paper do not require a parallel corpus. Because of the substantially higher resource requirements of the method of Razmara et al. (2013) we do not compare against this approach.

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

Language	Method	% Accuracy					
		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: % accuracy@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 accuracy@N, for each language and translation direction, are shown in boldface.

consider the case of a truly low-resource language.

We 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 we always use the full corpus for the target language, but a sample for the source language. Therefore, we 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. We 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 accuracy@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

POS tagging or dependency parsing.

5.2. Combined In-vocabulary and OOV Test Set

In this subsection we consider an evaluation that considers both in-vocabulary and OOV source language words. We show that, although the accuracies reported in Table 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.

We build a new test dataset consisting of both in-vocabulary and OOV source language words. For each language, we select 1500 test pairs that are in-vocabulary in both the source and target languages from the data of Conneau et al. (2018),[9] and an equal number of test pairs from our 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, we compose the representations of OOV source language words from their sub-word embeddings. In the case of in-vocabulary source language words, we use their embeddings as their representation. We then use the transformation matrix to find their target language translations. We refer to this approach — which uses sub-word information for OOVs — as "Sub-Word". We compare this against a method that does not use sub-word embeddings. For this latter method, we 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, we apply the copy baseline. We refer to this method — that has no knowledge of sub-words, and relies on the copy baseline to translate OOVs — as "Copy".

Results are shown in Table 4. Because the results for BPE in Table 3 were relatively poor, in these experiments we only report results for fastText as the source of sub-word information.[10] For all languages, with English as either the source or target language, and for every level of supervision, the Sub-Word approach outperforms the Copy approach, except in a small number of cases, specifically Finnish as the source language for accuracy@1 using the semi-supervised and unsupervised approaches, German as the target language for accuracy@1 using the supervised approach, and Russian as the target language for accuracy@10 using the semi-supervised approach. 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.

We applied a Mcnemar's test with continuity correction to determine whether the results using the sub-word method were significantly different from those using the Copy method. To avoid carrying out an overly-large number of tests, we only conduct tests for the supervised method — which based on the findings in Table 4 often gives the best performance — and for accuracy@10 — where the accuracies are highest — although we 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 Sub-Word is significant in the case of the supervised method for accuracy@10.

5.3. Interpolation

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. We therefore also considered incorporating edit distance into our approach for bilingual lexicon induction for OOVs. In these experiments, we only considered the supervised method to form cross-lingual word embeddings, and fastText embeddings as the source of sub-word information. We considered only these approaches because of the previous findings that fastText outperforms BPE, and that the supervised method often performs best.

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

$$\lambda \text{sim}(s,t) + (1-\lambda)\text{NMED}(s,t) \quad (2)$$

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 we randomly sampled 1000 test pairs, and 1000 development pairs from each dataset in Table 2.[11] In these experiments, as for subsection 5.1., all the source language words are OOVs. The development data was used to tune λ by grid search. We did not consider Japanese and Russian because they do not use the Latin alphabet. For each language, $\lambda = 0.7$ gave the best results on the development data.

Results are shown in Table 5. Focusing on accuracy@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.[12] In terms of accuracy@5 and @10 we see the same trend, although there are some exceptions for Finnish. Nevertheless, we do not see the massive gains in absolute accuracy from incorporating edit distance reported by Braune et al. (2018), suggesting that finding translations for OOVs might be particularly challenging compared to finding translations for low frequency words attested in a corpus.

[9] We found very few of the words in these translation pairs to be OOV in our corpora, motivating the construction of the dataset described in Section 4 .

[10] We carried out the experiments described in this sub-section using BPE, and the results were indeed relatively poor. For each language, and level of supervision, the accuracy using BPE was less than half the accuracy than when using fastText.

Language	Evaluation		% Accuracy					
			English source			English target		
	Supervision	Method	@1	@5	@10	@1	@5	@10
Finnish	Supervised	Copy	12.01	23.30	27.43	25.92	37.04	39.76
		Sub-Word	12.18	24.28	28.71	25.92	37.42	40.53
	Semi-supervised	Copy	11.54	21.00	24.57	21.96	29.98	33.22
		Sub-Word	11.62	21.88	25.80	21.94	30.39	33.96
	Unsupervised	Copy	11.71	20.97	25.17	22.05	30.84	33.41
		Sub-Word	11.92	21.92	26.39	21.99	31.20	34.16
German	Supervised	Copy	15.33	25.18	27.90	18.72	26.32	28.46
		Sub-Word	15.29	26.31	29.63	19.02	27.69	30.56
	Semi-supervised	Copy	16.07	24.79	27.60	19.62	26.37	28.72
		Sub-Word	16.15	25.78	29.15	19.86	27.72	30.88
	Unsupervised	Copy	16.33	24.97	27.26	16.45	23.42	25.51
		Sub-Word	16.45	26.12	28.76	16.79	24.73	27.64
Japanese	Supervised	Copy	20.34	29.98	33.08	13.66	22.29	24.23
		Sub-Word	20.55	30.57	33.93	13.70	22.60	24.66
	Semi-supervised	Copy	19.79	28.75	30.79	10.14	16.22	18.28
		Sub-Word	19.87	29.21	31.46	10.25	16.74	18.88
	Unsupervised	Copy	20.38	28.70	31.38	9.98	15.58	17.53
		Sub-Word	20.47	29.13	32.19	10.10	15.99	18.06
Russian	Supervised	Copy	14.73	25.52	28.22	22.25	29.71	31.57
		Sub-Word	14.90	25.87	28.88	22.57	30.74	33.44
	Semi-supervised	Copy	13.05	24.02	27.16	19.32	26.89	29.04
		Sub-Word	13.13	24.36	25.57	19.83	27.72	30.88
	Unsupervised	Copy	12.91	24.28	27.33	19.99	27.37	29.59
		Sub-Word	12.95	24.62	27.93	20.50	28.59	31.25
Spanish	Supervised	Copy	24.06	34.47	36.42	27.39	34.86	37.27
		Sub-Word	24.06	35.62	38.33	27.52	35.72	38.57
	Semi-supervised	Copy	25.17	34.20	36.11	27.35	34.30	36.97
		Sub-Word	25.24	35.47	37.95	27.57	35.16	38.27
	Unsupervised	Copy	25.03	34.07	35.93	25.97	32.44	34.30
		Sub-Word	25.16	35.21	37.73	26.17	33.13	35.38

Table 4: % accuracy@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 "Sub-Word" indicates employing sub-word embeddings to find translations for OOVs.

Language	λ	% Accuracy					
		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	-	-	0	-	-

Table 5: % accuracy@N incorporating edit distance for OOV source language words.

Language	Supervision	% Accuracy English target		
		@1	@5	@10
Cherokee	supervised	0.75	2.12	2.49
	semi-supervised	0	0	0
	unsupervised	0	0	0

Table 6: % accuracy@N for Cherokee, using English as the target language, for each level of supervision.

5.4. Low-resource Language Experiments

So far we have simulated lower-resource languages by down-sampling the source language corpora. In this section we consider the case of Cherokee, a low-resource language. Cherokee is an endangered Iroquoian language, spoken in the United States, with approximately $12k$ speakers. It is a polysynthetic language written using a syllabary.

In these experiments, we use Cherokee as the source language, and English as the target language. Because of the previous findings that fastText embeddings outperform BPE, we only consider fastText embeddings in these experiments. Specifically, we use fastText embeddings pre-trained on Cherokee Wikipedia.[13] The embedding table for Cherokee contains only 7033 words. For English embeddings, we use the embeddings trained on the full English Wikipedia from our previous experiments.

We build training and test datasets from the English–Cherokee translations in Panlex. For the supervised method, we 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, we 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.[14]

Results are shown in Table 6. The accuracy@1 for the copy baseline is 0% (results not shown). 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 abut translations for OOV words. However, the accuracy@N for the semi-supervised and unsupervised methods is 0%, which suggests that in the case of a truly low-resource language, these methods might not be capable of handling OOVs.

[11]We downsampled the dataset due to the large number of edit distance calculations required in this evaluation.

[12]The results for $\lambda = 1$ differ from those in Table 3 because they are for a sample of the full dataset.

[13]`https://fasttext.cc/docs/en/pretrained-vectors.html`

[14]We do not consider the case of a test set consisting of both in-vocabulary and OOV source language words because we only have 1143 translation pairs that are in-vocabulary for both languages, all of which are used for training the supervised approach.

6. Conclusions

In this paper we considered whether sub-word embeddings can be leveraged in cross-lingual word embedding models. Specifically we evaluated sub-word embeddings in a novel bilingual lexicon induction task in which we identify target language translations for OOV source language words. Our 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. We 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. Our findings further indicate that bilingual lexicon induction for OOVs can be improved by incorporating orthographic similarity. In future work we plan to consider alternative approaches to learning cross-lingual embeddings that incorporate knowledge of sub-words during training.

7. Bibliographical References

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria.

Artetxe, M., Labaka, G., and Agirre, E. (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.

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

Baldwin, T., Pool, J., and Colowick, S. (2010). Panlex and lextract: Translating all words of all languages of the world. In *In 23rd International Conference on Computational Linguistics*, pages 37–40, Beijing, China.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Braune, F., Hangya, V., Eder, T., and Fraser, A. (2018). Evaluating bilingual word embeddings on the long tail. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 188–193, New Orleans, Louisiana.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel

data. In *6th International Conference on Learning Representations*, Vancouver, Canada.

Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). 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.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas.

Fang, M. and Cohn, T. (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.

Hauer, B., Nicolai, G., and Kondrak, G. (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*, volume 2, pages 619–624, Valencia, Spain.

Jawanpuria, P., Balgovind, A., Kunchukuttan, A., and Mishra, B. (2019). Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, E. (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.

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

Razmara, M., Siahbani, M., Haffari, R., and Sarkar, A. (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.

Sennrich, R., Haddow, B., and Birch, A. (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.

Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Xing, C., Wang, D., Liu, C., and Lin, Y. (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.

Zhu, Y., Vulić, I., and Korhonen, A. (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.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (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, USA.