Evaluating a Joint Training Approach for Learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora on Lower-resource Languages

Ali Hakimi Parizi and **Paul Cook**
Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3 Canada
{ahakimi,paul.cook}@unb.ca

Abstract

Cross-lingual word embeddings provide a way for information to be transferred between languages. In this paper we evaluate an extension of a joint training approach to learning cross-lingual embeddings that incorporates sub-word information during training. This method could be particularly well-suited to lower-resource and morphologically-rich languages because it can be trained on modest size monolingual corpora, and is able to represent out-of-vocabulary words (OOVs). We consider bilingual lexicon induction, including an evaluation focused on OOVs. We find that this method achieves improvements over previous approaches, particularly for OOVs.

1 Introduction

Word embeddings are an essential component in systems for many natural language processing 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). Cross-lingual word representations provide a shared space for word embeddings of two languages, and make it possible to transfer information between languages (Ruder et al., 2019). A common approach to learn cross-lingual embeddings is to learn a matrix to map the embeddings of one language to another using supervised (e.g., Mikolov et al., 2013b), semi-supervised (Artetxe et al., 2017), or unsupervised (e.g., Lample et al., 2018) methods. These methods rely on the assumption that the geometric arrangement of embeddings in different languages is the same. However, it has been shown that this assumption does not always hold, and that methods which instead jointly train embeddings for two languages produce embeddings that are more isomorphic and achieve stronger results for bilingual lexicon induction (BLI, Ormazabal et al., 2019), a well-known intrinsic evaluation for cross-lingual word representations (Ruder et al., 2019; Anastasopoulos and Neubig, 2020). The approach of Ormazabal et al. uses a parallel corpus as a cross-lingual signal. Parallel corpora are, however, unavailable for many language pairs, particularly low-resource languages.

Duong et al. (2016) introduce a joint training approach that extends CBOW (Mikolov et al., 2013a) to learn cross-lingual word embeddings from modest size monolingual corpora, using a bilingual dictionary as the cross-lingual signal. Bilingual 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.

Hakimi Parizi and Cook (2020b) propose an extension of Duong et al. (2016) that incorporates sub-word information during training and therefore can generate representations for OOVs in the shared cross-lingual space. This method also does not require parallel corpora for training, and could therefore be particularly well-suited to lower-resource, and morphologically-rich, languages. However, Hakimi Parizi and Cook only evaluate on synthetic low-resource languages. We refer to the methods of Duong et al. and Hakimi Parizi and Cook as DUONG2016 and HAKIMI2020, respectively.

Most prior work on BLI focuses on in-vocabulary (IV) 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 OOVs (Hakimi Parizi and Cook, 2020a) and low-resource languages (Anastasopoulos and Neubig, 2020). In this paper, we evaluate HAKIMI2020 on BLI for twelve lower-resource

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 302–307
August 5–6, 2021, Bangkok, Thailand (online) ©2021 Association for Computational Linguistics

languages, and also consider an evaluation focused on OOVs. Our results indicate that HAKIMI2020 gives improvements over DUONG2016 and several strong baselines, particularly for OOVs.

2 Joint Training Incorporating Sub-word Information

Equation 1 shows the cost function for DUONG2016, which jointly learns embeddings for a word w_i and its translation $\bar{w}_i$, where h_i is a vector encoding the context, α is a weight parameter, and D_s and D_t are the source and target language vocabularies, respectively.

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

Following Bojanowski et al. (2017), HAKIMI2020 modifies Equation 1 by including sub-word information during the joint training process as follows:

$$\begin{aligned} O = \sum_{i \in D_s \cup D_t} \Big(& \alpha \log S(w_i, h_i) \\ & + (1-\alpha) \log S(\bar{w}_i, h_i) \\ & + \sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log -S(w_j, h_i) \Big) \end{aligned} \tag{2}$$

$$S(w, h) = \frac{1}{|G_w|} \sum_{g \in G_w} z_g^T h \tag{3}$$

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

HAKIMI2020 use character n-grams as sub-words. Specifically, each word is augmented with special beginning and end of word markers, and then represented as a bag of character n-grams, using n-grams of length 3–6 characters. The entire word itself (with beginning and end of word markers) is also included among the sub-words.

Language	Family	# Tokens	# Dict. entries
Afrikaans	Germanic	25M	70k
Albanian	Albanian	21M	17k
Azerbaijani	Turkic	36M	25k
Bengali	Indic	26M	114k
Bosnian	Slavic	18M	23k
Croatian	Slavic	54M	388k
Estonian	Uralic	38M	201k
Greek	Greek	78M	253k
Hebrew	Semitic	143M	79k
Hindi	Indic	34M	296k
Hungarian	Uralic	133M	460k
Turkish	Turkic	79M	319k

Table 1: The language family, size of corpus, and size of Panlex dictionary, for each source language.

3 Experimental Setup

We consider BLI from twelve lower-resource source languages to English. The languages (shown in Table 1) were selected to cover a variety of language families, while having small to medium size Wikipedias and BLI evaluation datasets available. We compare HAKIMI2020 with DUONG2016, VECMAP (Artetxe et al., 2018), and MEEMI (Doval et al., 2018). In each case, we use cosine similarity to find the closest target language translations for a source language word. We evaluate using precision@N (Ruder et al., 2019) for $N = 1, 5, 10$.

3.1 Training Corpora and Dictionaries

The corpus for each language is a Wikipedia dump from 27 July 2020, cleaned using tools from Bojanowski et al. (2017), and tokenized using EuropalExtract (Ustaszewski, 2019), except for Bengali and Hindi, which are tokenized using NLTK (Bird et al., 2009). Because DUONG2016 and HAKIMI2020 can learn high quality cross-lingual embeddings from monolingual corpora of only 5M sentences each, we down-sample the English corpus for these two methods to 5M sentences.

DUONG2016 benefits from a relatively large training dictionary (Duong et al., 2016), therefore, for DUONG2016 and HAKIMI2020 we follow Duong et al. to create large training dictionaries by extracting translation pairs from Panlex. Details of the training corpora and Panlex dictionaries are shown in Table 1.

3.2 Baselines

We compare against two baselines: VECMAP (Artetxe et al., 2018), a supervised mapping-based method, and MEEMI (Doval et al., 2018), a post processing method. We consider various training corpora and dictionaries to create strong baselines.

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, except for Azerbaijani, which is not included in MUSE, where training pairs from Anastasopoulos and Neubig (2020) are used. We first train VecMap using these MUSE pairs, and embeddings learned from the full English corpus, to give this baseline access to as much training data as is available. We then consider this approach, but using the down-sampled English corpus. We found that the smaller English corpus gave higher precision@N (for N = 1, 5, and 10) for both the IV and OOV evaluations in Section 4. This could be due to the smaller corpus having a smaller vocabulary. We then also consider VecMap trained using Panlex training pairs and embeddings learned from the down-sampled English corpus.

We next consider Meemi applied to each of the three sets of cross-lingual embeddings obtained from VecMap. In each case we train Meemi using the same training pairs (MUSE or Panlex) that were used to train VecMap. In Section 4 we report results for the baseline that performs best.

3.3 Hyper-Parameter Settings

Hakimi Parizi and Cook (2020b) show that duong2016 performs best using its default parameters, i.e., an embedding size of 200 and window size of 48, but that hakimi2020 performs better using an embedding size of 300 and window size of 20. We use these parameter settings here.

fastText is used to train monolingual embeddings for VecMap and Meemi. We use skipgram with its default settings, except the dimension of the embeddings is set to 300 (Bojanowski et al., 2017).

4 Experimental Results

In this section, we present results for BLI for IV words, and then OOV source language words.

4.1 BLI for In-Vocabulary Words

For these experiments we use MUSE test data for all languages except Azerbaijani, where we use test data from Anastasopoulos and Neubig (2020). Because our focus here is on IV words, we only consider translation pairs that are IV with respect to the embedding matrices learned from our corpora. We compare hakimi2020 with duong2016 and Meemi trained using the down-sampled English corpus and MUSE training pairs, which performed best of the baselines considered for each evaluation measure. Results are shown in Table 2.[1]

Method	% Precision		
	@1	@5	@10
Meemi	**38.64**	55.42	60.45
duong2016	22.12	45.71	52.08
hakimi2020	30.91	**56.00**	**62.24**

Table 2: Precision@N for BLI for IV words, averaged over the twelve languages. The best precision for each evaluation measure is shown in boldface.

hakimi2020 improves over duong2016, indicating that duong2016 can indeed be improved by incorporating sub-word information during training. Comparing hakimi2020 and Meemi, the results are more mixed. In terms of precision@1, Meemi substantially outperforms hakimi2020, although for precision@10 hakimi2020 outperforms Meemi.

4.2 BLI for OOVs

Following Hakimi Parizi and Cook (2020a) we use Panlex to construct a test dataset of translation pairs in which the source language words are OOV and the target language words are IV. However, Hakimi Parizi and Cook observe that some translations in Panlex are noise. To avoid noisy translations, we use all translation pairs for which the source language word is OOV with respect to the embedding matrix, i.e., the embedding models have no direct knowledge of these words, but is attested in the source language corpus, i.e., there is evidence that this is indeed a word in the source language.[2] The resulting test datasets consist of between 806 translation pairs in the case of Azerbaijani to roughly 11k pairs for Hungarian.

Here we compare against the VecMap baseline using the down-sampled English corpus and Panlex training pairs, which performed best of the baselines considered for each evaluation measure. For VecMap, we follow Hakimi Parizi and Cook (2020a) by forming a representation

[1] Results for each of the twelve languages are available in the appendix.

[2] For each embedding method, we set the minimum frequency for words in the embedding matrix to 5; as such, all methods have the same source language vocabulary.

Language	Method	% Precision		
		@1	@5	@10
Afrikaans	VecMap	5.65	11.84	14.89
	copy	10.68	-	-
	hakimi2020	9.42	21.80	27.41
	hakimi2020+copy	**19.16**	**30.15**	**35.17**
Albanian	VecMap	6.28	12.00	15.75
	copy	5.62	-	-
	hakimi2020	7.93	15.20	18.61
	hakimi2020+copy	**13.11**	**19.49**	**22.58**
Azerbaijani	VecMap	3.60	8.93	11.41
	copy	5.96	-	-
	hakimi2020	10.17	16.00	17.25
	hakimi2020+copy	**10.92**	**19.35**	**21.96**
Bengali	VecMap	1.60	4.50	6.00
	copy	0.27	-	-
	hakimi2020	**5.31**	**10.95**	**13.85**
	hakimi2020+copy	5.28	10.86	13.76
Bosnian	VecMap	3.82	8.28	10.83
	copy	21.23	-	-
	hakimi2020	8.17	15.71	18.58
	hakimi2020+copy	**29.19**	**35.88**	**38.11**
Croatian	VecMap	6.41	13.29	17.03
	copy	4.35	-	-
	hakimi2020	11.86	24.70	30.13
	hakimi2020+copy	**15.65**	**28.02**	**33.21**
Estonian	VecMap	5.29	10.61	13.79
	copy	7.56	-	-
	hakimi2020	8.15	18.79	23.66
	hakimi2020+copy	**14.93**	**24.65**	**29.15**
Greek	VecMap	6.66	14.30	17.91
	copy	1.90	-	-
	hakimi2020	11.65	23.55	28.05
	hakimi2020+copy	**13.50**	**25.15**	**29.58**
Hebrew	VecMap	3.07	8.38	10.53
	copy	11.15	-	-
	hakimi2020	8.18	17.08	20.55
	hakimi2020+copy	**19.02**	**26.89**	**29.75**
Hindi	VecMap	2.09	5.16	6.98
	copy	0.06	-	-
	hakimi2020	4.57	11.64	15.39
	hakimi2020+copy	**4.60**	**11.66**	**15.41**
Hungarian	VecMap	4.30	9.49	12.50
	copy	4.60	-	-
	hakimi2020	7.82	17.42	21.66
	hakimi2020+copy	**11.62**	**20.56**	**24.53**
Turkish	VecMap	3.39	7.23	9.62
	copy	8.15	-	-
	hakimi2020	7.13	15.43	19.38
	hakimi2020+copy	**14.27**	**21.31**	**24.77**
Average	VecMap	4.35	9.50	12.27
	copy	6.70	-	-
	hakimi2020	8.36	17.36	21.21
	hakimi2020+copy	**14.27**	**22.83**	**26.50**

Table 3: Precision@N for BLI for OOV source language words. The best precision for each dataset and evaluation measure is shown in boldface.

for the OOV source language word from its sub-word embeddings, and then mapping it into the shared space. We cannot, however, compare directly against duong2016 because it is a word-level approach that cannot represent OOVs. We therefore instead compare against a baseline in which the OOV source language word is copied into the target language. This approach, referred to as copy, could work well in the case of borrowings and named entities.[3]

Table 3 shows the results. hakimi2020 outperforms VecMap for all languages and evaluation measures. This finding suggests that sub-word information can be more effectively transferred in a cross-lingual setting when sub-words are incorporated into the training process — as is the case for hakimi2020 — than when they are not — as for VecMap here. Comparing hakimi2020 to copy, although there are several languages for which copy outperforms hakimi2020, on average, hakimi2020 performs better. In the cases that copy outperforms hakimi2020, it appears to be largely related to the presence of English abbreviations in the source language Wikipedia dump.

Because of the relatively strong performance of copy on several languages, we propose an approach that combines copy and hakimi2020, referred to as hakimi2020+copy. Given a source language word, we first check whether it is in the target language embedding matrix. If so, we assume it is a word that does not require translation (e.g., a named entity) and copy it into the target language.[4] If the source language word is not in the target language embedding matrix, we apply hakimi2020 to find the target language translation under this model. This approach improves over both copy and hakimi2020 for all languages, except Bengali, and gives substantial improvements on average.[5] Although copy is a very simple approach, it is complementary to hakimi2020, and the two approaches can be effectively combined to improve BLI for OOVs.

5 Conclusions

We evaluated an extension of a joint training approach to 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

[3] copy only produces one target language candidate for a given source word, and as such we only compute precision@1 for this method.

[4] This assumption can be incorrect, e.g., Afrikaans *kits* is IV for English, but translates to English *moment*.

[5] We also observe that there is little improvement for hakimi2020+copy over hakimi2020 on Hindi. For both Hindi and Bengali copy achieves very low precision, and so little or no improvement can be obtained over hakimi2020 by combining copy with hakimi2020.

trained on modest amounts of monolingual data and can represent OOVs. In two BLI tasks for twelve lower-resource languages focused on IV words and OOVs, we found that this method improved over previous approaches, particularly for OOVs. Evaluation data and code for learning the cross-lingual embeddings is available.[6]

In future work we plan to explore the impact of the target language on the quality of the cross-lingual embeddings, and in particular consider source and target languages from the same family. We further intend to evaluate these cross-lingual embeddings in down-stream tasks for low-resource languages, such as language modelling (Adams et al., 2017) and part-of-speech tagging (Fang and Cohn, 2017), and to compare against approaches based on contextualized language models.

References

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.

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.

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. 2018. 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.

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.

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.

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.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304, Brussels, Belgium. 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.

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.

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 subword information without parallel corpora. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 39–49, Barcelona, Spain (Online). Association for Computational Linguistics.

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,*

[6] `https://github.com/Cons13411/XLing_Subword`

ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA.

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.

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.

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

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.

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.