Multilingual Word Translation using Auxiliary Languages

Hagai Taitelbaum
Faculty of Engineering
Bar-Ilan University, Israel
hagait62@gmail.com

Gal Chechik
The Gonda Brain Research Center
Bar-Ilan University, NVIDIA
gal.chechik@biu.ac.il

Jacob Goldberger
Faculty of Engineering
Bar-Ilan University, Israel
jacob.goldberger@biu.ac.il

Abstract

Current multilingual word translation methods are focused on jointly learning mappings from each language to a shared space. The actual translation, however, is still performed as an isolated bilingual task. In this study we propose a multilingual translation procedure that uses all the learned mappings to translate a word from one language to another. For each source word, we first search for the most relevant languages. We then use the auxiliary translations to these languages to form an improved representation of the source word. Finally, this representation is used for the actual translation to the target language. Experiments on a standard multilingual word translation benchmark demonstrate that our model outperforms state of the art results.

1 Introduction

Monolingual continuous word embeddings are standard building blocks of many natural language tasks. The embedding spaces can exhibit similar structures across languages. Several studies (Mikolov et al., 2013; Klementiev et al., 2012) proposed to exploit this similarity by learning a linear mapping from a source to a target embedding space, and demonstrated this approach on a word translation task. Xing et al. (2015) showed that using orthogonality matrices can significantly improve performance.

Bilingual embedding can be extended to a multilingual setup by jointly learning mappings from each monolingual space to a shared word space. In recent years several studies have proposed aligning multiple languages simultaneously in a shared space by enforcing (or at least encouraging) transitive relations between the mappings (Chen and Cardie, 2018; Kementchedjhieva et al., 2018; Alaux et al., 2019; Jawanpuria et al., 2019; Taitelbaum et al., 2019). Dealing with multiple languages simultaneously has been shown to improve performance on some bilingual tasks by using knowledge learned from other languages (Ammar et al., 2016; Duong et al., 2017).

Once the multilingual mappings are learned, one can infer word correspondences for words that are not in the initial lexicons. Previous multilingual methods have focused on training procedures that benefit from the multilingual setting. The actual translation, however, is still done as an isolated bilingual task. It makes eminent sense however to go beyond this stage and utilize the relations between all the languages at the inference phase as it is done in the mapping learning phase.

In this study, we propose a new inference method for multilingual word translation that uses all the learned mappings to translate a word from one language to another. For each source word we first search for the most relevant languages that help translate the source word to the target language. We then use the auxiliary translations to form an improved representation of the source word. Finally, this multilingual-dependent representation is used for the actual translation to the target language.

Our main contributions is twofold: first, a new word translation inference method, which takes advantage of the multilingual setup not only in the train phase, but also in test phase. Second, evaluation on a recently-released multilingual word translation dataset on six languages (Lample et al., 2018) showing that our method outperforms state-of-the-art methods on this task.

2 Multilingual Word Mapping

In this section we review the concept of multilingual mapping to a shared space and in the next section we use it to form a multilingual translation method. Assume we are given a d-dimensional

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1330–1335,
Hong Kong, China, November 3–7, 2019.

word embedding data from a set of k languages. The task is to learn linear mappings between every pair of languages. In order to learn these mappings, we are also given a dictionary for each pair of languages, which contains pairs of corresponding words from the two languages. These dictionaries can be obtained in either a supervised or an unsupervised manner, and can also be created at each iteration of a dictionary refinement method (Artetxe et al., 2017; Lample et al., 2018). We can learn cross-language mappings independently from each source to each target language. This approach fails to benefit from the multilingual setup, and is very expensive because it requires to learn k^2 mappings, each with its own independent parameters. Another approach consists of choosing one language as a "pivot" and learning a mapping from each language to the pivot independently. This strategy, however, does not guarantee good indirect word translations between pairs of languages that do not include the pivot.

Recent studies proposed to jointly map all languages into a shared space. This approach was shown to outperform the above two approaches (Chen and Cardie, 2018). Let $T_1, \ldots, T_k$ be a set of mappings that correspond to the k different languages. The mapping T_i is used to translate the words from language i to a shared space that can be viewed as an embedding space of a universal language.

The translation matrices $T_1, ..., T_k$ can be found by minimizing the following mean-square error:

$$S(T_1, ..., T_k) = \sum_{i<j} \sum_{t} \|T_i x_{it} - T_j x_{jt}\|^2 \quad (1)$$

such that x_{it} and x_{jt} are embeddings of two corresponding words in languages i and j respectively. The optimal transformations map pairs of words with similar meanings to vectors in the shared space that are close to one another.

When more than two languages are involved, there is no closed-form solution for the global minimum of Eq. (1). Recently, several studies addressed this optimization challenge. (Kementchedjhieva et al., 2018; Chen and Cardie, 2018; Alaux et al., 2019; Taitelbaum et al., 2019).

3 Multilingual Translation Method

In this section we propose an inference procedure for multilingual translation that uses *all* the learned mapping $T_1, \ldots, T_k$ to translate from one language to another.

Before we describe our method, consider first a generic formulation of the inference procedure. The translation of a word embedding x from language i to a language j is obtained by:

$$\hat{y} = \arg\max_{y \in V_j} sim(x, y) \quad (2)$$

where V_j is the vocabulary of language j. $sim(x, y)$ can be, for example, the cosine similarity in the embedded space: $sim(x, y) = cos(T_i x, T_j y)$. It is commonly observed that inference using the nearest embedded neighbor suffers from the *hubness problem* (Dinu and Baroni, 2014). Hubs are words that appear too frequently in the neighborhoods of other words. To mitigate this effect, one can simply replace the cosine similarity by another criterion, such as *Inverted Softmax* (ISF) (Smith et al., 2017) or *Cross-domain Similarity Local Scaling* (CSLS) (Lample et al., 2018). Following recent bilingual and multilingual translation studies (e.g., Lample et al. (2018); Alaux et al. (2019); Chen and Cardie (2018); Joulin et al. (2018)), this study uses the CSLS metric, namely, for $x \in V_i, y \in V_j$:

$$sim(x, y) = \text{CSLS}(T_i x, T_j y). \quad (3)$$

The CSLS similarity is calculated as follows:

$$\text{CSLS}(z, w) = 2\cos(z, w)$$

$$-\frac{1}{n} \sum_{z' \in N_z(w)} \cos(z', w) - \frac{1}{n} \sum_{w' \in N_w(z)} \cos(z, w')$$

where cos is the cosine similarity, $N_z(w)$ is the set of n nearest neighbors of the point w in the first set of word vectors and $N_w(z)$ is similarly defined. In practice, we used $n = 10$.

To describe our multilingual word translation, note that current word translation methods take into account only the source and target languages at inference time. Since the translation is performed via the shared embedding space one can potentially design a better representation of the source word in the shared space to be translated into the target word. We can translate x to all other languages (except the target language j): $y_m = \arg\max_{y \in V_m} \text{CSLS}(T_i x, T_m y)$ and then compute the average word in the shared space:

$$z = \frac{1}{k-1} \sum_{m \neq j} T_m y_m \quad (4)$$

(for the source word we set $y_i = x$). Here, all languages except the target are used as auxiliary

sources of information about the correct embedding in the shared space. Then, we use z as a new multilingual representation of $T_i x$, and translate x according to:

$$\hat{y} = \arg\max_{y \in V_j} \text{CSLS}(z, T_j y) \quad (5)$$

Unfortunately, using the average across all auxiliary languages, may hurt performance for some language pairs. For example, a German translation of a Spanish word may not help with translating that Spanish word to Portuguese. More generally, translations to auxiliary languages can yield words that are far from the source word in the shared space and therefore may lead to incorrect translation. Here we describe an approach to select the relevant auxiliary languages for a given source word and target language. The main idea is to apply CSLS to select those languages that would be helpful for translating the desired source word. Specifically, a language m is selected as an auxiliary language only if the translated word y_m is closer than the target word y_j to the source word x in the shared space. A language m is thus included in the summation of Eq. (4) only if:

$$\text{CSLS}(T_i x, T_m y_m) > \text{CSLS}(T_i x, T_j y_j) \quad (6)$$

When averaging the auxiliary translations all these languages are equally weighted. Because the source word is more important than its auxiliary translations, we set the weight of the source $T_i x$ to be sum of all the weights of the auxiliary words. The proposed *Multilingual Word Translation* (MWT) procedure is depicted in Algo. 1

Algorithm 1 Multilingual Word Translation

Required: A set of mappings $T_1, .., T_k$.
Task: Translate the word $x \in V_i$ to language j.

for $m = 1, ..., k$ **do**
  $y_m = \arg\max_{y \in V_m} \text{CSLS}(T_i x, T_m y)$
  $c_m = \text{CSLS}(T_i x, T_m y_m)$
end
$z = T_i x$
$S = \{m \in \{1, ..., k\} \backslash \{i, j\} \,|\, c_m > c_j\}$
if $|S| \neq 0$ **then**
  $z \leftarrow \frac{1}{2} T_i x + \frac{1}{2|S|} \sum_{m \in S} T_m y_m$
end
$\hat{y} = \arg\max_{y \in V_j} \mathit{CSLS}(z, T_j y)$
return $\hat{y}$

Several studies have recently proposed using word vector averaging of the source and target embeddings as an improved shared word representation. This can be done when the monolingual word embeddings are mapped to the same space (Doval et al., 2018). Meta-embedding using several embeddings of the same language can be achieved even without mapping the embeddings to a shared space (Coates and Bollegala, 2018). In this study, we also built an improved source word representation by averaging. However, the averaging is done with translations of the source word to suitable auxiliary languages instead of the target word.

4 Experiments

To evaluate the proposed MWT algorithm, we used a recently released multilingual word translation dataset (MUSE) in six European languages: English, German, French, Spanish, Italian and Portuguese (Lample et al., 2018)[1]. In addition, we conducted another experiment mixing European (English, German, French, Spanish and Italian) together with Asian languages (Japanese , Chinese and Korean). This experiment demonstrates the power of MWT even for distant languages. The available dictionaries in MUSE dataset are between the European languages, and between English and each of the Asian languages. For any available pair of languages, a ground-truth bilingual dictionary is provided with a train-test split of 5000 and 1500 unique source words, respectively. All systems are tested on the 1500 test word pairs for each pair of languages.

Monolingual Embeddings. Pre-trained 300d fastText (monolingual) embeddings[2] (Bojanowski et al., 2017) trained on the Wikipedia corpus. These embeddings are popular among word translation systems (e.g., Lample et al. (2018); Alaux et al. (2019); Chen and Cardie (2018); Hoshen and Wolf (2018); Joulin et al. (2018); Kementchedjhieva et al. (2018).

Implementation details. For the six European languages experiment, the multilingual mapping set was trained using the state-of-the-art unsupervised *Multilingual Adversarial Training* (MAT) + *Multilingual Pseudo-Supervised Refinement* (MPSR) method (Chen and Cardie, 2018). We used their source code[3], with their default

[1] https://github.com/facebookresearch/ MUSE
[2] https://github.com/facebookresearch/ fastText
[3] https://github.com/ccsasuke/umwe

	en-de	en-fr	en-es	en-it	en-pt	de-en	de-fr	de-es	de-it	de-pt	fr-en	fr-de	fr-es	fr-it	fr-pt	
BI	75.1	82.7	82.6	79.1	81.8	73.4	76.9	68.4	72.2	63.3	82.0	71.0	84.1	83.6	80.3	
NT	74.7	82.8	82.7	80.5	81.5	72.7	77.3	**72.5**	76.5	69.7	81.6	71.3	84.5	84.2	82.3	
CNT	75.1	**83.4**	**83.2**	**81.1**	81.9	74.7	78.5	72.3	**77.1**	69.7	82.5	71.5	84.5	**84.6**	82.1	
CAT	**75.5**	83.1	83.1	80.9	**82.1**	**74.9**	**78.9**	72.2	76.3	**70.1**	**83.3**	**73.1**	**84.6**	84.4	**82.7**	
	es-en	es-de	es-fr	es-it	es-pt	it-en	it-de	it-fr	it-es	it-pt	pt-en	pt-de	pt-fr	pt-es	pt-it	avg
BI	**83.5**	68.7	86.9	85.0	**87.9**	77.7	68.9	88.1	88.5	82.7	79.7	65.2	86.3	**92.8**	82.3	79.4
NT	82.4	70.3	**88.7**	84.8	86.5	79.5	71.3	88.7	89.5	**84.7**	80.9	69.3	**89.2**	91.1	85.1	80.6
CNT	82.7	70.3	88.6	85.1	87.2	79.6	71.7	89.1	**89.9**	**84.7**	81.1	69.2	89.1	92.3	**85.2**	80.9
CAT	83.0	**71.6**	**88.7**	**86.1**	87.1	**79.7**	**72.5**	**89.2**	89.8	**84.7**	**81.5**	**69.7**	88.2	92.3	**85.2**	**81.1**

Table 1: Multilingual word translation results for English, German, French, Spanish, Italian and Portuguese. The reported numbers are precision@1 in percentage.

source word	**direct (BI)**	**multilingual (CAT)**	**auxiliary words**
pace (en)	veloz (es)	ritmo (es)	ritmo (pt), tempo (de)
Cornell (en)	Harvard (it)	Cornell (it)	Cornell (de), Cornell (fr), Cornell (es)
lens (en)	pentaprisma (it)	lente (it)	lente (es), lente (pt), linse (de)

Table 2: Examples of erroneous translation of BI corrected by CAT, and its relevant auxiliary words.

hyper-parameters, and got similar results to the reported results (+0.1% in average). For the European-Asian experiment, MAT failed to converge for some language pairs, so the multilingual mapping set was trained using supervised MPSR, where the supervision was obtained by pairs of words with identical string matching. For each experiment, we used the same mappings for all the methods we compare. Our code and mapping matrices will be publicly available.

Compared methods. All methods retrieve word translations using their CSLS similarity in the learned embedding space.

(1) BI (Bilingual Inference). A standard inference process which does not take the multilingual setup into account, as in Chen and Cardie (2018).

We implemented 3 translation variants using auxiliary languages:

(2) NT (Nearest Translation). Average the source word with the closest auxiliary translation.

(3) CNT (Conditional Nearest Translation). Only average the source word with the closest auxiliary translation, if it is closer than the translation to the target language. In fact, CNT chooses for each source word one of {BI, NT} depends on whether the closest auxiliary translation is closer than the target translation (NT) or not (BI).

(4) CAT (Conditional All Translations). Weighted average of the source word and all auxiliary translations, that are closer than the target language translation. CAT is formally described in Algorithm box 1.

Results. Table 1 presents detailed results for all 30 language pairs and the average results. It shows that using all relevant auxiliary languages (CAT) increases performance significantly (+1.7% on average, top method in 19/30 tasks, $p < 0.001$ [4]). The largest performance boost of CAT over BI was in languages pairs involving *German* (+3.17% on average), which is the most distant language in this set of languages, thus, gains a lot from using other languages. This was found in particular for the translations between German and Portuguese (*de-pt*: +6.8%, *pt-de*: +4.5%), which are the most distant languages in this language set. This suggests that using MWT for distant languages may help. However, for close languages pairs the best way is still to translate directly (BI), as can be seen for Spanish and Portuguese. Table 2 presents three examples of erroneous bilingual translations that were corrected using auxiliary languages.

A similar behaviour is shown in the European-Asian languages (Table 3). CAT seems to improve word translation, especially over BI (+1.6% on average, top method in 18/26 tasks, $p < 0.001$). Moreover, CAT especially improves word translations between English and the Asian languages

[4] We performed a statistical significance test as in (Glavas et al., 2019) in addition to a statistical significance test over the accuracies obtained from each language pair.

	en-de	en-fr	en-es	en-it	en-ja	en-zh	en-ko	de-en	de-fr	de-es	de-it	fr-en	fr-de	
BI	75.1	82.3	82.2	78.5	27.1	30.7	28.8	73.5	76.0	68.4	72.2	82.3	70.3	
NT	74.7	81.9	81.6	79.6	28.4	32.6	30.0	72.7	76.9	71.9	75.4	81.5	71.1	
CNT	**75.3**	82.4	82.3	80.0	28.4	32.7	30.0	**74.4**	77.9	72.1	**76.3**	82.1	71.4	
CAT	74.8	**82.5**	**82.5**	**80.3**	**28.9**	**33.3**	**30.2**	74.0	**78.1**	**73.3**	75.4	**82.5**	**73.1**	
	fr-es	fr-it	es-en	es-de	es-fr	es-it	it-en	it-de	it-fr	it-es	ja-en	zh-en	ko-en	avg
BI	84.1	82.9	**83.1**	68.1	86.7	84.5	77.7	69.0	88.3	88.4	17.6	24.0	31.1	66.7
NT	84.3	83.1	81.6	70.9	87.6	83.9	**79.3**	70.9	88.1	88.0	14.1	25.8	35.1	67.3
CNT	84.7	**83.7**	82.2	**71.0**	**88.1**	84.7	**79.3**	71.0	**88.7**	88.9	16.7	**27.1**	35.9	68.0
CAT	**84.9**	**83.7**	81.9	70.9	87.9	**85.3**	79.1	**71.6**	**88.7**	**89.2**	**18.5**	26.9	**37.5**	**68.3**

Table 3: Multilingual word translation results for English, German, French, Spanish, Italian, Japanese, Chinese and Korean. The reported numbers are precision@1 in percentage.

→	en	de	fr	es	it	pt
en		es (29%)	– (29%)	– (29%)	fr (28%)	fr (29%)
de	– (36%)		en (37%)	en (34%)	en (37%)	en (37%)
fr	it (29%)	it (27%)		– (29%)	– (32%)	it (30%)
es	pt (44%)	pt (40%)	pt (39%)		pt (43%)	– (47%)
it	fr (33%)	es (36%)	es (35%)	– (35%)		es (34%)
pt	es (59%)	es (55%)	es (58%)	– (62%)	es (57%)	

Table 4: Auxiliary languages that are most frequently selected by the CNT method ('–' is none), for every pair of source and target languages. Value in parenthesis denotes how often that auxiliary language was selected.

(+1.93% in average from English and +3.4% to English).

We next show more detailed analysis for CNT, when using at most one auxiliary language for the six European languages experiment. Table 4 shows auxiliary language that is most commonly selected, for each pair of source-target languages. Interestingly, Spanish and Portuguese often help each other. Also, German often uses English as an auxiliary language for translating better into all other languages.

For CAT, each source word may use a different number of auxiliary languages. We can see the number of auxiliary languages as a mean to qualitatively measure closeness of languages, by looking on the average number of auxiliary languages used for each source-target pair. We found Portuguese and Spanish to be the closest (*pt-es* 0.6), and German the farthest from them (*pt-de* 3.1, *es-de* 3.2).

5 Conclusion

We presented a general concept to improve the quality of bilingual word translation by using the translation of the source word to auxiliary languages. We discussed several variants for deciding which and how many languages should be used as suitable auxiliary languages. The same translation principle can be used in the dictionary refinement step of the mapping training process.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *International Conference on Learning Representations*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Annual Meeting of the Association for Computational Linguistics*, pages 451–462. Association for Computational Linguistics.

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.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embeddingcomputing metaembeddings by averaging source word embeddings. In *Conference of the North American Chapter of the Association for Computational - Linguistics Human Language Technologies*.

Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

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

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Coh. 2017. Multilingual training of crosslingual word embeddings. In *The Conference of the European Chapter of the Association for Computational Linguistics*.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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, Herve Jegou, , and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *CONLL*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Int. Conference on Learning Representations.*

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. 2019. A multi-pairwise extension of procrustes analysis for multilingual word translation. In *Conference on Empirical Methods in Natural Language Processing*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.