Improving Bilingual Lexicon Induction on Distant Language Pairs

Wenhao Zhu[1], Zhihao Zhou[1], Shujian Huang[1(✉)], Zhenya Lin[2],
Xiangsheng Zhou[2], Yaofeng Tu[2], and Jiajun Chen[1]

[1] Nanjing University, Nanjing 210023, China
{whzhu,zhouzh}@smail.nju.edu.cn,
{huangsj,chenjj}@nju.edu.cn
[2] ZTE Corporation, Shenzhen, China
{lin.zhenya,zhou.xiangsheng,tu.yaofeng}@zte.com.cn

Abstract. Aligning the representation spaces of two languages to induce a bilingual lexicon achieves attractive results on European language pairs. Unfortunately, current solutions perform terribly on distant language pairs. To address this problem, we analyze existing models for the lexicon induction task of distant language pairs, such as English-Chinese. We propose an framework for the task with improved preprocessing, mapping and inference accordingly. Experimental results show that our proposed approach enhances the accuracy of bilingual lexicons substantially on English-Chinese, as well as some other distant language pairs.

Keywords: Natural language processing · Machine translation · Bilingual lexicon induction

1 Introduction

The Lexical translation table (or bilingual lexicon) is an essential part of machine translation (MT). Traditionally, dictionaries for bilingual lexicons are composed manually, which involves massive expert knowledge and expense. Since the research showing that the representation spaces of two languages can be aligned through a simple linear mapping [9], bilingual lexicon induction (BLI) has achieved great success on English-Italy, English-German language pairs and is drawing increasingly attention recently [7,13,14].

However, existing BLI models perform much worse on distant language pairs [11]. Intuitively, larger distance between two languages does bring more difficulty in aligning the two representation spaces. But previous researches do not pay enough attention on why the accuracy of these methods degrades substantially on distant language pairs.

In this paper, we make deep analysis of typical BLI models, which consist of three steps: preprocessing, mapping and inference [1,3,12]. We discuss the obstacles for applying current model directly to distant language pairs, and try

S. Huang and K. Knight (Eds.): CCMT 2019, CCIS 1104, pp. 1–10, 2019.
https://doi.org/10.1007/978-981-15-1721-1_1

to improve the induction performance on distant language pairs by improving each step correspondingly.

More specifically, in the preprocessing step, we verify that "center" [1] is the key operation which can bring great gain for performance; in mapping, we propose to use multiple local mappings instead of a single one; in inference, we propose an approximated searching algorithm to determine the hyper parameter K in the *CSLS* method [8], so that "topic words" could be successfully distinguished from "hub words".

To demonstrate effectiveness of our method, quantitative experiments are conducted on English-Chinese fasttext dataset [6]. Experimental results show that our methods could tackle observed weaknesses and the improved framework outperforms existing methods. Furthermore, we demonstrate that our approach can be applied to other distant language pairs as well.

2 Background

Given the word embedding of two languages as input, the task of bilingual lexicon induction is to align the two embedding spaces and retrieve word pairs (bilingual lexicons) as output for downstreaming tasks. There are two popular branches in researches of BLI. One is the supervised methods, which require aligned word pairs as a seed dictionary [1,3,12]. Another branch of research is unsupervised methods, such as self-learning [2,4] and GAN-based models [5,8,15]. Because unsupervised methods are extremely unstable on distant language pairs, we mainly discuss the supervised methods in this paper.

For convenience, we will use the following definitions throughout this paper. We denote source word embedding as $\hat{X} \in \mathbb{R}^{n \times d}$ and target word embedding as $\hat{Y} \in \mathbb{R}^{m \times d}$, each row of which represents a single word vector. We use $X \in \mathbb{R}^{t \times d}$ and $Y \in \mathbb{R}^{t \times d}$ to denote the word vectors of aligned word pairs. So the i^{th} rows of X and Y represent words that are translation of each other.

Following Artetxe et al. [3], typical supervised BLI models consist of three main steps: preprocessing, mapping and inference, where the embedding of both languages are transformed; the mapping function is learned; and finally, the bilingual lexicon is inferred. We will briefly introduce these steps in the following subsections.

2.1 Preprocessing

In preprocessing, some simple operations are applied to transform the representation space before mapping. These operations aim at making embeddings in the two representation spaces distribute as similarly as possible. Taking source language embedding X as an example, Xing et al. [14] proposed the "unit" operation to ensure word vector X_{i*} is of unit length. Later, Artetxe et al. [1] proposed the "center" operation, which let the mean of each column vector X_{*i} to be 0. Besides, Artetxe et al. [3] presents several other operations, such as "whiten", "re-weight", "de-whiten", "reduction". Please refer to their original paper for details.

Previous research has demonstrated that all of them contribute to the improvement of model performance on close-related language pairs. However, there is no guidance on using these transformations for distant language pairs.

2.2 Mapping

After getting two transformed representation spaces, a mapping function could be learned to build the mapping between the two, so that the embedding vectors of aligned pairs stay as close as possible.

The function is usually a linear transformation matrix W. Mikolov et al. [9] treat it as a linear regression problem. The training objective function is to minimize the sum of squared Euclidean distances:

$$\arg\underset{W}{min} \sum_{i} ||X_{i*}W - Y_{i*}||^2 \tag{1}$$

More generally, it can be rewritten into the matrix form of Frobenius norm:

$$\arg\underset{W}{min} \; ||XW - Y||_F^2 \tag{2}$$

Xing et al. [14] propose to add an orthogonal constrain ($W^TW = I$) into the process, which keeps the monolingual invariance after mapping. The neural mapping with a hidden state [11] has also been tried but it suffers the severe overfitting problem. Up to now, orthogonal mapping has become a standard way to project language space.

With the mapping function, e.g. W, source embedding $\hat{X}$ and target embedding $\hat{Y}$ are expected to be projected into the same space.

2.3 Inference

For inference, retrieval methods are used to obtain translation pairs from the mapped space. For a given word x, its induction translation y is

$$\arg\underset{y}{min} \; f(xW, y) \tag{3}$$

where f is the retrieval function.

Mikolov et al. [9] apply nearest neighbour (NN) to get the corresponding target word, where $\cos(\cdot, \cdot)$ is used as measure. Dinu et al. [7] find that NN approach will suffer severe "hubness problem". More specifically, hub is some meaningless target words which appear as the nearest neighbour of many source words. As a result, methods such as *invnn* [7], *invsoftmax* [12], and *CSLS* [8] are proposed to alleviate this problem.

Taking speed and accuracy into consideration, *CSLS* is recognized as the best way to induce bilingual lexicons. It considers the mean similarity of a source word x to its target neighbour as:

$$r_T(xW) = \frac{1}{K} \sum_{y \in \mathcal{N}_T(xW)} \cos(xW, y) \tag{4}$$

where $\mathcal{N}_T(xW)$ is the K nearest target neighbours of source word x; K is a hyper-parameter, which is usually set as 10. $r_S(y)$ can be denoted in the same way. Thus the whole retrieval function of *CSLS* is:

$$CSLS(xW, y) = 2\cos(xW, y) - r_T(xW) - r_S(y) \tag{5}$$

3 Improved Framework

Here we present our contributions to the three steps of the BLI tasks.

3.1 Preprocessing

Current preprocessing operations are weakly explainable. Simply stacking them can't ensure the same effect on distant language pairs. We provide an empirical analysis of the transformations with English-Chinese language pair as an example. We find that "unit" and "center" are the most important transformation, while other transformations do not bring significant improvement. Details of the empirical analysis are provided in the experiment section (Sect. 4.2).

3.2 Multiple Local Mappings

Previously all research papers use a single matrix W as transformation function based on the assumption that vector spaces have similar geometric arrangement [9]. However we doubt it's not held for distant language pairs and that's also the main reason why the model performance degrades under such settings. Experimental results show that a single mapping learns poorly on the training set, let alone the test set. Similar geometric distribution may only happens locally. A set of multiple local mappings $\{W_i\}_{i=1}^m$ rather a single mapping W better model BLI on distant pairs. The objective function of the local area centered at x_c is:

$$\arg\min_{W_i} \sum_{x_j \in \mathcal{N}_S(x_c)} ||x_j W_i - y_j||^2 \tag{6}$$

Following the objective and method described in Sect. 2.2, multiple local mappings $\{W_i\}_{i=1}^m$ can be obtained. Then given a source word x as a test case, the local mapping whose center is the closest to x will be applied to project it.

The remaining problem is how to produce multiple local mappings. In this paper, we propose to organize words by their topics. Assume topic word x_c are chosen as the center of a sub seed dictionary, such as "animals" or "politics", which summarizes a bunch of words. Analogous to *CSLS*, we define $\mathcal{N}_S(x_c)$ as K nearest source neighbour of x_c. For each word pair in the seed dictionary, word pairs surrounding x_c will be put into the sub seed dictionary $\{(x_i, y_i)$, where $x_i \in \mathcal{N}_S(x_c)\}$. In this way, multiple sub seed dictionaries centered at different topic words can be built for training multiple local mappings.

3.3 Approximated Searching

Though *CSLS* enjoys success in its efficiency and low computation expense, it still faces some problems in practice. We find that *CSLS* always confuses "topic words" with "hub words", as both have great similarity with neighbour words which always makes "topic words" punished wrongly as "hub words".

In Table 1. we list some wrong translation cases. For example, 液体 (liquid) is the so-called topic word. It is always mistaken as "hub words" by *CSLS* so that it won't be chosen as candidate translation.

Table 1. Some representative wrong translation cases in which the *CSLS* method punish "topic words" as "hub words" incorrectly.

Original Word	Translation Word	Ground Truth
液体 (liquid)	pressurizing	liquid
二手 (secondhand)	buyers	secondhand
反正 (anyway)	surprising	anyway

However, we find this phenomena can be changed by setting K value correctly. This is easy to explain when considering the difference between "topic words" and "hub words". When the parameter value is small, both topic word and hub words have great similarity with neighbour word which makes them hard to distinguish. As the value raises, it reaches the balance to translate both type of words correctly. Since the similarity between "topic words" and its neighbour word declines while it is not the case for "hub words". But if K gets too large, the accuracy will decline because hub words no more stay closed to its K-NN words.

In original paper, K is recommended to be set as 10. We observe that induction accuracy keeps raising if we increase K and then declines when K gets too large. Therefore we propose an approximated searching algorithm to choose K in *CSLS* formula:

- increase K in step of 10 and compute model accuracy on the training set;
- once induction performance declines, we choose K in the last step as optimal value.

4 Experiments

4.1 Setup

All of the analysis are conducted on the fasttext dataset [6]. It provides word vectors of various languages in dimension 300 that are pretrained on Wikipedia corpus by skip-gram model [10] described in the paper of Bojanowski et al. [6]. The dataset also contains seed dictionary for different language pairs. According to source word frequency, the top 5000 words and their matched pairs make up for the training set. The top 5000 to 6500 words and their translation make up for the test set. The results are evaluated by the final accuracy of the retrieved bilingual lexicons on the test set.

We present detailed analysis about different steps (Sects. 4.2, 4.3 and 4.4) of the BLI models, with English-Chinese as an example language. Experiments of the whole improved framework are then presented, with a comparison to related studies, across multiple language pairs (Sect. 4.5). Further analysis are provided in Sect. 4.6.

4.2 Empirical Study of Transformations

We first compare the different transformations used in the preprocessing step. Following previous work [1], we take an orthogonal matrix as the mapping function and nearest neighbour as the retrieval method. The results are shown in Table 2.

Table 2. Accuracy of BLI models that take different combinations of preprocessing on English-Chinese.

unit	center	whiten	de-whiten	re-weight	reduction	Acc.
						27.33%
✓						27.13%
✓	✓					42.47%
✓	✓	✓				42.47%
✓	✓	✓	✓			42.47%
✓	✓	✓	✓	✓		42.47%
✓	✓	✓	✓	✓	✓	42.47%

The results show that “center” brings most performance gain and “unit” plus “center” is the optimal combination for distant language pairs. Additional transformation doesn’t help enhancing accuracy but increases computational burden.

The possible explanation is that, for distant language pairs, two representation space are far from similar. “unit” and “center” are the simplest but effective way to normalize the two spaces, which enables the model to learn a high quality mapping more easily.

4.3 Employing Multiple Mapping Function

We then study the effect of mapping functions. We doubt whether a single mapping is suitable for distant language pairs since the results made by it is not satisfying. While two distributions differ significantly as a whole, but in the partial aspect the difference is smaller in our observation. Multiple mappings maybe a better solution, which project vector space part by part. We keep the setting of using "unit" and "center" in preprocessing and *CSLS* as the retrieval method. We manually choose 10 topic words and divided the seed dictionary into 10 sub groups. Different local mappings are learnt for different groups.

Table 3. Train set accuracy (ACC_{tr}) and test set accuracy (ACC_{te}) of high quality local mappings on English-Chinese datasets. The last line is the accuracy of baseline. The next-to-last line is the average accuracy of representative groups.

topic word	train dict size	ACC_{tr}	test dict size	ACC_{te}
"animal"	1230	94.74	471	51.15
"culture"	1331	92.95	342	52.34
"education"	1315	92.60	351	51.24
Average		93.43		51.58
Single mapping		45.14		32.47

The results are listed in Table 3. For simplicity, we list the accuracy and related information of multiple mappings for three representative groups, with topic words "animal", "culture", "education", respectively. Both the representative groups and the average results show that the accuracy of using multiple local mappings is substantially better than a single global map for different groups. Besides, we find that the baseline model acts poorly on training set which indicates that a single mapping is far from perfect.

However, although multiple local mappings demonstrate their ability by considerable improvements, we do notice that automatically choosing the number of local mappings and selecting reasonable topic words for each mapping are difficult. At the current stage, this method is not integrated into our final system. We leave this as an important future work.

4.4 Inference with Approximated Searching

CSLS usually fails to distinguish "topic words" from "hub words". But we find that it can be overcome by tuning K in the formula. To show the effect of different K, we take two language pairs (English-Chinese and English-German) as examples, and draw the accuracy curves as K changes in Fig. 1.

As we can see in Fig. 1, the curve keeps raising at the beginning and declines when K gets too large. To conclude, a medium K suits the case most. Our proposed approximated searching algorithm can quickly determine a medium K which ensure it achieves best performance in inference part.

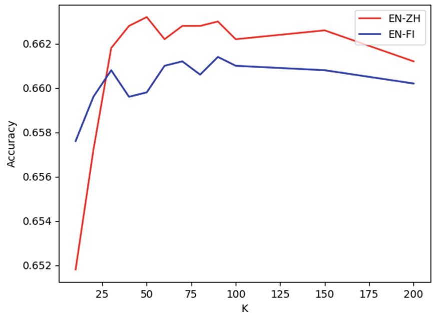

Fig. 1. Accuracy curve of the model when K in *CSLS* formula changes. (“EN-ZH” is English-Chinese, “EN-FI” is English-Finnish)

Table 4. Precision for BLI task compared with previous work. The baseline model employs an orthogonal mapping as mapping function, *CSLS* as retrieval metric and no preprocessing. (“EN” is English, “ZH” is Chinese, “JA” is Japanese, “KO” is Korean, “FI” is Finnish, “DE” is German)

	Distant pairs				Closed pairs
	EN-ZH	EN-JA	EN-KO	EN-FI	EN-DE
Mikolov et al. [9,10]	13.27	14.16	16.11	32.47	61.20
Xing et al. [14]	27.13	2.54	24.64	38.67	68.13
Dinu et al. [7]	27.00	32.49	25.32	43.33	66.33
Artetxe et al. [1]	42.47	45.65	27.03	42.93	70.30
Smith et al. [12]	12.47	1.10	25.05	44.60	71.40
Nakashole et al. [11]	43.27	-	-	-	68.50
Baseline	32.47	1.71	31.47	47.60	73.37
uc + CSLS	45.33	51.68	31.54	65.76	79.02
Improved	**45.80**	**51.68**	**32.29**	**66.08**	**79.34**

4.5 The Improved Framework

Here we present the results of our final framework, which is a combination of following two improvements: the preprocessing with “unit” and “center” and *CSLS* with our searching for K.

We conduct experiments on both distant and close language pairs and present results in Table 4. The last two line show performance gain brought by improved preprocessing and inference respectively. It’s obvious that both parts contribute to the improvement of accuracy. On top of that, results show that the modified framework outperforms existing models on distant language pairs in particular. For distant language pairs, improved framework achieved more than ten percentage points on average above the baseline expect on English-Korean. For closed language pairs, the improvement is much smaller.

4.6 Further Analysis

Though improved lexicon quality has been achieved by our model, we still want to figure out what prevents the model inducing perfect lexicon. Therefore we contrast the error bilingual lexicons with the ground truth and find that the bad cases are mostly due to synonyms. Some representative mistakes are listed below in Table 5. We find that the BLI model is so smart that it predicts 舌头 (tongue) as ear's translation where they are already very closed. However the model is not smart enough to close the gap between 舌头 (tongue) and the true translation 耳朵 (ear).

Table 5. Some representative wrong translation pairs made by our improved framework on English-Chinese where predicted words have great similarity with correct translations.

Source Word	Predicted Word	Ground Truth
ear	舌头 (tongue)	耳朵 (ear)
myanmar	泰国 (thailand)	缅甸 (myanmar)
honey	柚子 (Pomelo)	蜂蜜 (honey)
plural	单数 (singular)	复数 (plural)

Therefore in future work, we want to close the gap and predict translation more precisely instead of choosing synonyms as the target translation. If this problem is alleviated, the performance of BLI model will boost.

5 Conclusion

In this paper, we make deep analysis on the English-Chinese word translation task where both languages are familiar to us. Based on comparison and analysis, we propose three methods to address observed problems. We present an improved framework with proposed methods for bilingual lexicon induction on distant language pairs. Experimental results demonstrate that our framework behaves excellently on distant language pairs and outperforms other existing models. Furthermore, we analyze wrong translations made by our framework and point out the gap that blocks model to perform perfectly on distant language pairs. In the future, we want to complete the algorithm of multiple local mappings and eliminate the effect brought by synonyms to predict translation more precisely.

Acknowledgement. We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by the National Science Foundation of China (No. 61772261), the Jiangsu Provincial Research Foundation for Basic Research (No. BK20170074), “13th Five-Year” All-Army Common Information System Equipment Pre-Research Project (No. 31510040201). This work is also partially supported by the research funding from ZTE Corporation.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 2289–2294 (2016)
2. Artetxe, M., Labaka, G., Agirre, E.: 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), pp. 451–462 (2017)
3. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: AAAI Conference on Artificial Intelligence, pp. 5012–5019 (2018)
4. Artetxe, M., Labaka, G., Agirre, E.: 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, vol. 1, pp. 789–798 (2018)
5. Barone, A.: Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In: Meeting of the Association for Computational Linguistics, pp. 121–126 (2016)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**(1), 135–146 (2017)
7. Dinu, G., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. In: International Conference on Learning Representations (2014)
8. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jegou, H.: Word translation without parallel data. In: International Conference on Learning Representations (2018)
9. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
11. Nakashole, N.: NORMA: neighborhood sensitive maps for multilingual word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 512–522. Association for Computational Linguistics, Brussels (2018)
12. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: International Conference on Learning Representations (2017)
13. Vulic, I., Korhonen, A.: On the role of seed lexicons in learning bilingual word embeddings, vol. 1, pp. 247–257 (2016)
14. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation, pp. 1006–1011 (2015)
15. Zhang, M., Liu, Y., Luan, H., Sun, M.: Adversarial training for unsupervised bilingual lexicon induction, vol. 1, pp. 1959–1970 (2017)