Bilingual Word Embedding: A Survey from Supervised to Unsupervised

Haipeng Sun[1*] , Rui Wang[2], Kehai Chen[2], Masao Utiyama[2], Eiichiro Sumita[2], and Tiejun Zhao[1]
1 Harbin Institute of Technology
2 National Institute of Information and Communications Technology

hpsun@hit-mtlab.net and tjzhao@hit.edu.cn
{wangrui, khchen, mutiyama, eiichiro.sumita}@nict.go.jp

1 Introduction

Word embedding is the basis of neural network based nature language processing (NLP) tasks. As monolingual word embedding has been widely used in many languages (Mikolov et al., 2013a), similarities among different languages were exploited by Mikolov et al. (2013b). The similarity between words in two or more languages can be represented by the correlation between word embedding. Bilingual word embedding (BWE) is to mapp the relationship between different monolingual word embeddings. Recently, several works show that BWE can be learn without bilingual supervision (Zhang et al., 2017a; Artetxe et al., 2018b). In this paper, we survey the BWE from supervised methods to unsupervised methods: supervised BWE methods that need a large training dictionary, semi-supervised BWE methods that need a small seed dictionary or only parallel corpus, unsupervised BWE methods.

2 Supervised BWE

The supervised BWE proposed by Mikolov et al. (2013b) exploits similarities between the source language and the target language by a linear transformation matrix. Supervised BWE methods usually need a large dictionary, and the linear matrix would be trained between language pairs in this dictionary. A set of word pairs $\{X, Z\}$ and their word vector representations $\{x_i, z_i\}_{i=1}^{n}$ are given, where n is the size of the word pairs, d is the dimension of word embedding. The training objective function of BWE is as follows:

$$\min_W \sum_i \|Wx_i - z_i\|^2 = \min_W \|WX - Z\|^2, \quad (1)$$

where W is the transformation matrix to be learned such that WX approximates Z. X, Z are the word embedding matrices of size $d \times n$, respectively. Dinu et al. (2015) used a L2-regularized least-squares error objective instead of previous objective function:

$$\min_W \|WX - Z\|^2 + \lambda \|W\|. \quad (2)$$

They also used a globally corrected neighbour retrieval method to mitigate the hubness problem. Xing et al. (2015) improved the performance of bilingual word embedding by enforcing word embedding normalized and an orthogonality constraint on W to preserve the length normalization. Then the bilingual word embedding problem has become the Orthogonal Procrustes problem and $W = UV^T$ in the equation (2) can be acquired by the singular value decomposition$ZX^T = U\Sigma V^T$. After word embedding could be normalized, the equation (2) that is equivalent to maximize the sum of cosine similarities can be reformulated as follows:

$$\max_W \sum_i cos(Wx_i, z_i). \quad (3)$$

Artetxe et al. (2016) considered the monolingual invariance through orthogonality to prevent the degradation in monolingual tasks. Length normalization and mean centering could be taken into consideration, so the equation (2) is also equivalent to maximizing the sum of dimension-wise covariance as

$$\max_W \sum_i cov(Wx_i, z_i). \quad (4)$$

*This work was conducted while Haipeng was an intern in NICT.

Smith et al. (2017) showed that the linear transformation should be orthogonal to preserve self-consistent. The inverted softmax retrieval replaced nearest neighbors method to achieve better translation pairs. Artetxe et al. (2018a) generalized and improved BWE mappings with a multi-step framework of linear transformations such as normalization, whitening, orthogonal transformation, reweighting, and dimensionality reduction behavior.

Compared with previous methods that learn transformation matrix from source language to target language, Faruqui and Dyer (2014) mapped the word embedding to a shared space where their similarity was maximized through canonical correlation analysis that is a statistical analysis method for measuring the linear correlation between two multidimensional variables as

$$\begin{aligned} W, G = CCA(X, Z) &= \arg\max_{W,G} \rho(WX, GZ) \\ &= \arg\max_{W,G} \rho(\frac{E[(WX)(GZ)]}{\sqrt{E[(WX)^2]E[(GZ)^2]}}), \end{aligned} \tag{5}$$

where $\rho(\cdot)$ means the correlation between the projected vectors, and W, G are projection matrices which map X, Z to the shared space in the equation (5). Lu et al. (2015) extended previous work with deep canonical correlation analysis to learn non-linear mapping.

3 Semi-supervised BWE

As BWE was developing, some researchers found that the supervision is not always necessary. Therefore, some works tried to reduce the supervision in BWE. Specifically, the aligned sentence pairs and small lexicon seeds were used as supervision of BWE.

3.1 Aligned Sentences as Supervision

BilBOWA (Bilingual Bag-of-Words without Alignments) (Gouws et al., 2015) is a simple and computationally-efficient model for learning bilingual distributed representations of words which can scale to large monolingual datasets and does not require word-aligned parallel training data. Instead it trains directly on monolingual data and extracts a bilingual signal from a smaller set of raw-text sentence-aligned data. The objective was optimized as

$$\min_{\theta^e, \theta^f} \sum_{l \in \{e,f\}} \sum_{w,h \in \mathbb{D}^l} \mathbb{L}^l(w, h; \theta^l) + \lambda \Omega(\theta^e, \theta^f), \tag{6}$$

where $\Omega(\cdot)$ is cross-lingual regularization term that was used to constrain monolingual models over the context h and target word w training pairs in the dataset $\mathbb{D}$ during jointly training.

Wang et al. used bilingual maximum complete subgraphs (cliques), which play the role of a minimal unit for bilingual sense representation (Wang et al., 2016, 2018). Cliques are dynamically extracted according to the contextual information. Consequently, correspondence analysis, principal component analyses, and neural networks are used to summarize the clique-word matrix into lower dimensions to build the embedding model.

3.2 Small Lexicon as Supervision

Artetxe et al. (2017) proposed self-learning framework to learn BWE with almost 25 word dictionary. During the self-learning, the squared Euclidean distance could be minimized:

$$\min_{W} \sum_{i} \sum_{j} D_{ij} \left\| W x_i - z_j \right\|^2, \tag{7}$$

where D is a binary matrix, and $D_{ij} = 1$ if the ith source word is aligned with the jth target word in the dictionary. The source and target embedding would be the length normalized and mean centered, and W would be constrained to be an orthogonal matrix. So the equation (7) is equivalent to maximizing the dot product as

$$\max_{W} Tr(W^T Z D^T X^T), \tag{8}$$

where $Tr(\cdot)$ means the sum of all elements in the main diagonal of matrix, $W = UV^T$ in the equation (8) can be acquired by the singular value decomposition $ZD^TX^T = U\Sigma V^T$.

4 Unsupervised BWE

However, the lack of large word pair dictionary poses a major practical problem for many language pairs. The unsupervised BWE has attracted much attention. Zhang

et al. (2017a) matched the distributions of the transformed source language embedding $x \sim \mathbb{P}_x$ and target language embedding $z \sim \mathbb{P}_z$ via generative adversarial network training. Two sets of monolingual word embedding $\{x_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^m$ with dimensionality d are trained separately on two languages. The discriminator D loss function is given by

$$L_D = -\mathbb{E}_{z\sim\mathbb{P}_z} logD(z) - \mathbb{E}_{x\sim\mathbb{P}_x} log(1 - D(Wx)). \tag{9}$$

The generator loss function is as follows:

$$\begin{aligned} L_G &= -\mathbb{E}_{x\sim\mathbb{P}_x} logD(Wx) \\ &= -\mathbb{E}_{x\sim\mathbb{P}_x} logD_1(Wx) - \mathbb{E}_{z\sim\mathbb{P}_z} logD_2(W^T z) \\ &= -\mathbb{E}_{x\sim\mathbb{P}_x} logD(Wx) - \lambda\mathbb{E}_{x\sim\mathbb{P}_x} cos(x, W^T Wx), \end{aligned} \tag{10}$$

where $W \in \mathbb{R}^{d\times d}$ is the transformation matrix to be learned such that Wx approximates z, its transpose W^T which can transform the target language z back to the source language x. D_1, D_2 are two separate discriminators. λ is a hyperparameter. Zhang et al. (2017b) used the Wasserstein distance instead of cross-entropy in the loss function. The Wasserstein distance is

$$\begin{aligned} &\mathcal{W}(\mathbb{P}_{Wx}, \mathbb{P}_z) \\ =&\frac{1}{K} \sup_{\|f\|_L \le K} \mathbb{E}_{z\sim\mathbb{P}_z}[f(z)] - \mathbb{E}_{x\sim\mathbb{P}_x}[f(Wx)], \end{aligned} \tag{11}$$

where f that is a K-Lipschitz function can be approximated with a neural network. The objective function for the discriminator D can be formulated as

$$L_D = \mathbb{E}_{x\sim\mathbb{P}_x}[f_D(Wx)] - \mathbb{E}_{z\sim\mathbb{P}_z}[f_D(z)]. \tag{12}$$

The generator loss function is given by

$$L_D = -\mathbb{E}_{x\sim\mathbb{P}_x}[f_D(Wx)]. \tag{13}$$

Conneau et al. (2018) used the same discriminator loss function as the equation (9). For the the generator G, the objective is to minimize

$$L_G = -\mathbb{E}_{x\sim\mathbb{P}_x} logD(Wx) - \mathbb{E}_{z\sim\mathbb{P}_z} log(1 - D(z)). \tag{14}$$

They also built a synthetic parallel vocabulary to refine the mapping W, and used cross-domain similarity local scaling (CSLS) instead of nearest neighbors to measure the similarity between mapped source words and target words, as

$$CSLS(Wx, z) = 2cos(Wx, z) - r(Wx) - r(z). \tag{15}$$

$$r(Wx) = \frac{1}{K} \sum_{z\in\mathcal{N}(Wx)} cos(Wx, z). \tag{16}$$

$$r(z) = \frac{1}{K} \sum_{Wx\in\mathcal{N}(z)} cos(Wx, z). \tag{17}$$

where $z \in \mathcal{N}(Wx)$ means the K nearest neighborhood of the mapped source embedding Wx and $Wx \in \mathcal{N}(z)$ means the K nearest neighborhood of the target embedding z. Dou et al. (2018) added variational autoencoder into the generative adversarial network in order to learn latent variables that can capture semantic meaning of words. Therefore the discriminator D loss function is

$$\begin{aligned} L_D = &\mathbb{E}_{v_z\sim q_{G_z}(v|z)}[logD(v_z)] \\ &+ \mathbb{E}_{v_x\sim q_{G_x}(v|x)}[log(1 - D(v_x))]. \end{aligned} \tag{18}$$

The objective function of generator G_x is to minimize

$$\begin{aligned} L_{G_x} = &\mathbb{E}_{v_x\sim q_{G_x}(v|x)}[logP_{x'}(x|v_x)] \\ &- \mathbb{E}_{v_x\sim q_{G_x}(v|x)}[logD(v_x)], \end{aligned} \tag{19}$$

where $q_{G_x}(v|x)$ and $q_{G_z}(v|z)$ is the posterior distribution of the latent variables, $P_{x'}(x|v_x)$ is the reconstruction distribution. The generator G_z function is similar. Sinkhorn distances and back-translation losses were used in the (Xu et al., 2018). The objective function is

$$L(W, G) = d_{sh}(W) + d_{sh}(G) + \beta d_{bt}(W, G), \tag{20}$$

$$\begin{aligned} d_{bt}(W, G) = &\sum_i 1 - cos(x_i, G(W(x_i))) + \\ &\sum_j 1 - cos(z_i, W(G(z_i))), \end{aligned} \tag{21}$$

where G is the transformation matrix to be learned such that Gz approximates x, $d_{sh}(W)$ is the Sinkhorn distance between P_{Wx}, P_z, $d_{sh}(G)$ is the Sinkhorn distance between P_{Gz}, and P_x, β is the hyperparameter.

Artetxe et al. (2018b) proposed a self-learning method like Artetxe et al. (2017) without any word dictionary. The objective function is similar as the equation (7), which considered two directional transformation as

$$\max_{W,G} \sum_i \sum_j D_{ij}(Wx \cdot Gz) \tag{22}$$

$W = V^T$ and $G = U^T$ in the equation (22) can be acquired by the singular value decomposition $ZD^T X^T = U\Sigma V^T$.

5 Conclusion and Future Work

In this paper, we surveyed the developing of bilingual word embedding methods, from supervised to unsupervised. We focused on the algorithms and methods instead of the corresponding performances. In the future, we will present the performances of different methods. In addition, we will survey the applications of bilingual word embedding in NLP, especially unsupervised neural machine translation.

参考文献

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*, Austin, Texas.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, Vancouver, Canada.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*, New Orleans, Louisiana.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, Melbourne, Australia.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*, Vancouver, Canada.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *ICLR*, San Diego, California.

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. Unsupervised bilingual lexicon induction via latent variable models. In *EMNLP*, Brussels, Belgium.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*, Gothenburg, Sweden.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, Lille, France.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL*, Denver, Colorado.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/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. In *ICLR*, Toulon, France.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A bilingual graph-based semantic model for statistical machine translation. In *IJCAI*, New York, USA.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2018. Graph-based bilingual word embedding for statistical machine translation. *TALLIP*, 17(4).

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, Denver, Colorado.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *EMNLP*, Brussels, Belgium.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, Vancouver, Canada.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *EMNLP*, Copenhagen, Denmark.