Bilingual Lexicon Learning With Feed Forward Network

Halil İbrahim Öztürk
Hacettepe University Computer Science Department
halil_ozturk@hacettepe.edu.tr

Abstract

In this study, we proposed two feed forward neural network models which predict similarity between meanings of source language word and target language word. Models are designed by influence of Siamese [2] network. The models accept monolingual word embeddings of input words. There is not publicly available bilingual lexicon learning dataset. Because of that we created our dataset which contains English to Italian translation pairs. Performances of our models are sufficient. Code and our dataset are available at https://github.com/hibrahimozturk/feed_forward_bll.

1. Introduction

The goal of the bilingual lexicon learning is that creating a lexicon which acts as a bridge between two languages. Bilingual lexicon is used to find target language word which share similar meaning with source language word. Learned bilingual lexicon forms the basis for cross language natural language processing applications (e.g. statistical machine translation, cross-language information retrieval).

After advent of monolingual word embeddings, they have become essential part of natural language processing applications. Although word embeddings of different languages are trained on texts isolated from other ones, the words which share similar meaning across different languages have similar geometric arrangements in both spaces as in figure 1. This similarity of word embedding spaces has provided a progress in cross language NLP applications, e.g. [13].

We consider Bilingual Lexicon Induction as a one shot learning problem. Because, trained model search a word from target language which is translation of a source language word which is not included to training dataset. [18] proposes method for Person Re-Identification which is one shot learning problem in computer vision era. The study [18] uses siamese [2] network to find person which is most similar to given quary person. Siamese [2] network decides similarity of given two signatures. We suggest that a model which shares similar architecture with siamese network can measure similarity between source and target language words.

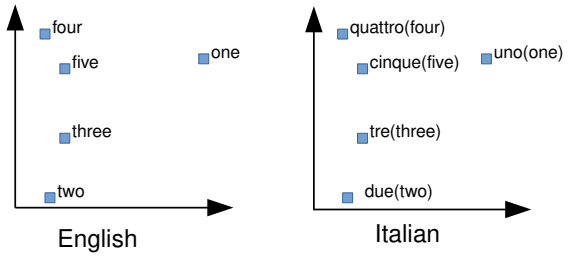


Figure 1: Illustrative monolingual word embeddings of English and Spanish, adapted from [13]

We proposed two methods to find words which share same meaning from across languages. Feed forward neural networks are employed to learn a nonlinear mapping from monolingual word embeddings to similarity score. First method fuses word embeddings early, second method fuses at mid layers, we call them early-fusion network and mid-fusion network, respectively. The results show that our models generate accurate results.

2. Related Works

Deselaers et al. [3] and Silberer et al. [17] show that images contain complementary information to the information extracted from text. Kiela et al. [9] and Bergsma et al. [1] studied vision-based bilingual lexicon learning by using features extracted from Convolutional Neural Networks. Vulic et al [19] combines the proposed visual-based approaches to get more successful results.

Supervised bilingual lexicon learning techniques requires a lot of annotated data. Vulic et al. [20] found that at least hundreds of paired word translations are needed to provide sufficient performance. Zhang et al. [21] uses

adversarial learning to generate word embeddings which used to calculate similarity between source and target language's words by using general adversarial networks which proposed at Goodfellow at al. [5].

Mikolov et al. [13], 2013; Dinu et al. [4], 2015; Lazaridou et al., 2015 [11] learn linear mapping which acts as bilingual lexicon.

The first time Heyman et al. [6] used character-level cross-lingual embeddings to bilingual lexicon induction. Irvine et al. [8] studies including raw word frequencies, temporal word variation and word burstiness to BLI process.

3. Method

The models accept monolingual word embedding vectors from source vocabulary V^S and target vocabulary V^T . Feed forward network weights are optimized by true and wrong translation pairs. Vectors $w_{ling} = [f_1, ..., f_dl]$ where d_l is size of the vector comprise real values. Our network measures similarity between source word and target word, $FFN(w_S, v_T) = sim$. The networks consist several layers which followed by activation functions in order to provide nonlinearity. In this respect, representation capability of our networks are increased.

Monolingual word embeddings are generated by pretrained models. The networks accepts word2vec vectors obtained from pretrained source and target language word2vec models.

If network outputs a similarity score that is higher than threshold for candidate translation pair, w word from source vocabulary V^S and v from target vocabulary V^T, the candidate translation pair is accepted as true.

$$g(w, v) = \begin{cases} +1 & \text{if } FFN(w_S, v_T) > t \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

Mid-Fusion model has two branches for each words in candidate translation pair, as in figure 2b. The outputs of dense branches are concatenated after second dense layers in branches. Similarity score is generated after three fully connected layers.

Early-Fusion model concatenates monolingual word embedding vectors at input layer, as in figure 2a. This model requires more parameters than mid-fusion model.

4. Implementation Details

4.1. Dataset Preparation

We extracted words from Wikipedia English Italian comparable corpora [1]. The corpora consists aligned contents (i.e. sentences). Bilingual lexicon learning algorithms require translated word pairs. Because of that, we translated English words to Italian words by Google Cloud Translation API. We thank to Google for 300$ free credit. Some translations contain more than one word, we eliminated them in order to use word2vec model properly. Also if a translated word which is not in word2vec vocabulary has been eliminated.

Dataset Splits	Number of Pairs
Training	20977
Validation	5322
Testing	1381

Table 1: Dataset distrubition

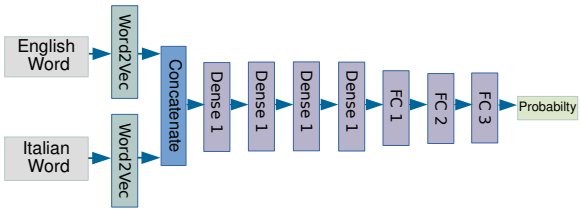
Vulic et. al. [19] translated BNC most frequent word list [2] to three language pairs. The word list comprises only headwords (e.g.'help', 'helps', 'helping', 'helped' are mapped to 'help'). Since we only lemmatized words extracted from Wikipedia comparable corpora, our dataset consists other forms of words (e.g. 'help', 'helping', 'helped').

Hence the training process requires positive and negative pairs, we generated wrong pairs for each English words in our dataset. We split dataset to training, validation and test sets. The wrong translations are not from other sets in order to make fair evaluation.

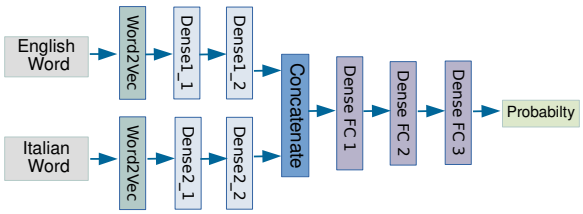
4.2. Network Architecture

After several experiments we have chosen best network architecture. The experiments are in commit history of project github repo. Rescaling word2vec vectors to [-1,+1] range has not improved accuracy. Therefore the word2vec vectors are not preprocessed before passing them to the network. 1D Batch normalization [7] follows each dense layer except last layer in branch in some models in table 2. In the paper [7] of batch normalization suggests that positioning batch normalization layer before activation function. But we found positioning batch normalization layer after activation function provides better performance. We used ReLU [14] and leaky ReLU [12] as activation

[1] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/
[2] http://www.kilgarriff.co.uk/bnc-readme.html



(a) Early-Fusion Siamese-BLL model



(b) Mid-Fusion Siamese-BLL model

	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
Mid Fusion BN ReLU	0.84	0.81	0.90	0.85
Mid Fusion BN LReLU	0.83	0.82	0.86	0.84
Mid Fusion ReLU	0.74	0.70	0.83	0.76
Mid Fusion LReLU	0.80	0.76	0.87	0.81
Early Fusion BN ReLU	0.86	0.86	0.88	0.87
Early Fusion BN LReLU	0.88	0.86	**0.91**	0.88
Early Fusion ReLU	0.90	0.89	0.90	0.90
Early Fusion LReLU	**0.91**	**0.92**	0.89	**0.90**

Table 2: Quantitative results

function after each layer except last layer. Sigmoid activation function follows last fully connected layer, because generated output should be in [0,1] range.

Dense(n, k) represents dense layer which takes an input vector which is n dimensional and outputs a vector which is k dimensional. In mid-fusion model, both of the branches accepts 300 dimensional word2vec vectors, the input follows throughout Dense1(300,256) $\rightarrow$ Dense2(256,128) layers until concatenation in both of the branches . DenseFC1(256,128) $\rightarrow$ DenseFC2(128,32) $\rightarrow$ DenseFC3(32,1) layers come after concatenation.

Early-fusion model concatenates word2vec vectors at first layer. Concatenated vector follows throughout Dense1(600,512) $\rightarrow$ Dense2(512,256) $\rightarrow$ Dense3(256,256) $\rightarrow$ Dense4(256,128) $\rightarrow$ FC1(128,64) $\rightarrow$ FC2(64,32) $\rightarrow$ FC3(32,1).

Early-fusion network comprises 258817 parameters, mid-fusion network comprises 550593 parameters.

We trained our model by Adam [10] optimizer with 0.0001 learning rate. Objective function is binary cross entropy. We employed SGD optimizer before switching to Adam, unfortunately loss value did not decrease in training. After each 50 step, learning rate is divided by 2.

We employed Gensim [16] framework to retrieve monolingual word embeddings and PyTorch [15] framework to train and evaluate our models. Training and testing are made on GPU.

5. Experiments

We employed accuracy, precision, recall and F1 metrics to measure performance of our method. As we mentioned before in our dataset a word exists in different forms, we did not select headword in order to increase dataset size. Word2Vec vectors of different forms of a word are not distinct. For example, true Italian translation of 'sketched' word is 'abbozzato'. But 'disegnato' and 'sketched' translation pair generates highest probability in pairs of all Italian test words and 'sketched' pairs where true translation of 'disegnato' is 'sketch'. The most similar target words can be translation of other form of source word, it causes that decreasing in *Top-1* accuracy. Because of that we do not employ *Top-1* accuracy metric to evaluate our model.

5.1. Results

Performance results are in table 2 for English to Italian translation pair dataset. We experimented on two different models with two different activation functions and with and without batch normalization. Batch normalization improves performance in mid-fusion networks. Since the input monolingual word embedding vectors follow two different branch until concatenation layer, ranges of layers outputs in branches can be different. We think batch normalization keeps outputs of layers at the branches in similar range. Batch normalization layers has not increased performance of early-fusion models unlike mid-fusion networks. This results supports our idea about batch

normalization in mid-fusion networks.

Leaky ReLU activation function provides better performance than ReLU activation function in all setups except mid-fusion network with batch normalization. In ReLU and leaky ReLU comparison respect, best improvement is observed at mid-fusion model without batch normalization.

6. Conclusion

We proposed two feed forward neural networks to predict similarity between words from two different languages. The models accept monolingual word embeddings of input words. We have provided performance results of the models. We could not employed *Top-1* accuracy metric, because there are different forms of words (e.g. follow, followed, following). Unfortunately we can not compare our results with previous works, because there is not publicly available benchmark set. Results of the models are better than random baseline. Early-fusion model outperforms mid-fusion model, but early fusion model involves more parameter, there is a trade-off. Leaky ReLU works mostly better than ReLU activation function. Batch normalization contributes to mid-fusion model, but it decreases performance of early-fusion model.

References

[1] Shane Bergsma and Benjamin Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[3] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *CVPR 2011*, pages 1777–1784. IEEE, 2011.

[4] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Geert Heyman, Ivan Vulić, and Marie-Francine Moens. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1085–1095, 2017.

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] Ann Irvine and Chris Callison-Burch. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310, 2017.

[9] Douwe Kiela, Ivan Vulić, and Stephen Clark. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, 2015.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 270–280, 2015.

[12] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

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

[14] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[16] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[17] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 721–732, 2014.

[18] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European conference on computer vision*, pages 135–153. Springer, 2016.

[19] Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 188–194, 2016.

[20] Ivan Vulić and Anna Korhonen. 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)*, volume 1, pages 247–257, 2016.

[21] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1959–1970, 2017.