Embedding Projection for Query Understanding

Yan Song
Microsoft
One Microsoft Way
Redmond, WA, USA, 98052
yansong@microsoft.com

Chia-Jung Lee
Microsoft
One Microsoft Way
Redmond, WA, USA, 98052
cjlee@microsoft.com

ABSTRACT

Learning reliable embeddings requires large data and is not trivial to be adapted to specific tasks due to certain constraints. For queries, the lack of sufficient context can prohibit learning quality representations and thus effective query understanding. We propose to project embeddings from a source space learned with natural language into a target space on queries. The projection function is learned via an overlap vocabulary set shared by both source and target spaces. Experimental results show that both linear and nonlinear embedding projections can help query intent classification and query slot tagging, even when the amount of data used for learning the projection is limited.

Keywords

Word embeddings, projection, query understanding

1. INTRODUCTION

Conventionally, word embeddings are trained via neural networks in an unsupervised manner on large-scale free text corpora, such as Wikipedia, news articles, or web text, which are often expressed in natural language. Although word embedding has demonstrated to be useful in NLP tasks, its effectiveness is highly dependent on the size of data and the availability of sufficient context. As a result, learning reliable embeddings from queries may be hindered since the average length of a web query is very short, providing limited amount of context. Directly using embeddings trained from free text for queries may also result in impotence since query language is very different from natural language.

In this paper, we propose to project embeddings from a source space learned with large, natural language corpora, into a target space on queries where the context is much more limited. Unlike existing embedding projection approaches that requires mapping among languages [5] or knowledge concepts [4], our embedding projection focuses on providing an inexpensive way to finding reliable representations for a target task where building embeddings has certain challenges (e.g., queries lack context). We propose to learn the projection function by minimizing the distance between the word embeddings from source and target spaces, which are connected by a set of overlapped vocabulary. In practice the vocabulary size can be very small. This design, when starting a new task, has potential to address cold-start problems considering that the amount of initial data in the target domain is often very limited. Therefore it provides an effective means for goal-specific tasks to leverage existing generic embeddings without having to re-train over large corpora.

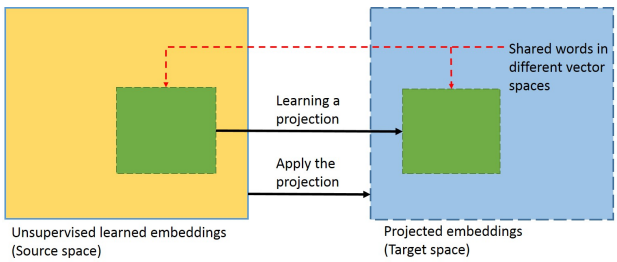

Figure 1: The concept of learning and applying embedding projection between source and target spaces.

We evaluate our approach using two common query understanding tasks, namely intent classification and slot tagging. Experimental results on public and commercial query logs show that our embedding projections can help these two tasks, where the result of using projected embbedings outperforms using either space or target embeddings.

2. EMBEDDING PROJECTION

Our framework of embedding projection is shown in Figure 1. The left big box represents the source embedding space $\mathbb{S}$, where the embeddings of words inside this space are learned in a conventional unsupervised manner. The right big box stands for the projected embedding space $\mathbb{T}$. Embedding projection from $\mathbb{S}$ to $\mathbb{T}$ is learned via the shared words ($w \in O$), represented by the small green boxes in both spaces. Based on this shared vocabulary, our goal is to find an f such that

$$\forall w \in O : E(w_{\mathbb{T}}) = f(E(w_{\mathbb{S}})) \quad (1)$$

where $E(w)$ refers to the embedding for w. Once f is found, projected embeddings for any arbitrary word $w \in \mathbb{S}$ can be constructed by simply applying f to its source $E(w_{\mathbb{S}})$.

We consider two approaches to building f. For **linear projection**, f is intrinsically a matrix P that aligns the

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/pp049

embeddings from $\mathbb{S}$ to $\mathbb{T}$ such that

$$\sum_{w \in O} \|E_{\mathbb{S}}(w)P - E_{\mathbb{T}}(w)\|_2 \leq \varepsilon \quad (2)$$

This formula can be solved by least square fitting. The assumption is that there exists a transformation between the two spaces where one can be mapped to the other through linear operations.

Non-linear projection presents a more generic way to projecting embeddings from $\mathbb{S}$ to $\mathbb{T}$. In practice , the projection is implemented with a two-layer perceptron, which can be formulated as

$$f(E_{\mathbb{S}}(w))_{w \in O} = G^{(2)}(b^{(2)} + W^{(2)}(G^{(1)}(b^{(1)} + W^{(1)}E_{\mathbb{S}}(w)))) \quad (3)$$

Here W and b are weight matrix and bias vector, respectively. G is the activation function, which is tanh in this paper. The superscripts (1) and (2) denote the first and second layer. The parameters are updated through stochastic gradient descent based on back-propagation algorithms.

3. EXPERIMENTS

3.1 Experiment Settings

We evaluate our projection approaches on two labeled datasets. The first is a benchmark dataset, ATIS [1], which is widely used in the community for intent classification and slot tagging. It contains 26 different intents and 153 slot types. In total there are 4978 utterances for training and 893 for testing. The second dataset is from annotated search logs, which contains 100,000 queries with 104 intents and 36 slot types. We use 80% of the dataset for training and the rest for testing. Labeled datasets form the target space $\mathbb{T}$, and the embeddings trained from $\mathbb{T}$ are regarded as target space embeddings $E(w_{\mathbb{T}})$. For training embeddings from source space $\mathbb{S}$, we use the Wikipedia dataset from its latest dump of articles[1], containing over 2 billion words. All labeled and unlabeled data are tokenized and normalized in our experiments. The embeddings from source and target spaces are trained by `word2vec` [3] with the continuous bag-of-word structure. In this task we use a bidirectional LSTM (bLSTM) with one hidden layer of 256 units for learning slots and intents jointly. The bLSTM is trained by minimizing the categorical cross entropy error over each training set using Adam optimizer [2]. In our experiments, we keep the hyper-parameters of the bLSTM fixed and test with different input embeddings.

3.2 Results

We have three sets of experiments: (1) directly use $\mathbb{S}$ embeddings trained from Wikipedia; (2) directly use $\mathbb{T}$ embeddings from ATIS or commercial queries; (3) apply the projected embeddings, $\mathbb{S} \rightarrow \mathbb{T}$. Another key factor that affects the projection performance is the choice of the shared word set O. To construct O, we rank the words in the target space according to their frequencies, and then select top n frequent term to form O. We test n = 10, 100 and 1000 for learning f. Results are shown in Table 1, where intent classification is measured by accuracy and slot tagging by F1. The best result in each column is marked bold.

Table 1 shows that both linear and nonlinear projections effectively adapt the embeddings from $\mathbb{S}$ to $\mathbb{T}$. All projected embeddings yield better performance than source embeddings, even when using only very limited data for learning the projection (i.e., small n). This property is useful to cold-start scenario in real applications. Second, the size of O affects the projection performance. In general, linear projection works better than nonlinear projection when the size of O is very small (e.g., $O(10)$). By enlarging O, however, the performance of linear projection tends to drop, especially on the evaluation of queries. This may be caused by that using more data increases the difficulty of solving the linear fitting between two matrices. For real traffic web queries, the assumption that $\mathbb{S}$ and $\mathbb{T}$ should be linearly correlated is less realistic in practice. For nonlinear projection, its performance improves with the increase of the size of O. Third, our results suggest that the projected embeddings can perform better than the embeddings directly learned from target space. The reason might be that the source embeddings trained from very large dataset can bring more reliable semantic relationship over to the target space by projection.

Table 1: Performance of query understanding on ATIS and Queries with using different embeddings.

Embeddings		ATIS Intent	ATIS Slot	Queries Intent	Queries Slot
$\mathbb{S}$		93.17	93.08	84.14	82.70
$\mathbb{T}$		**94.74**	94.26	85.06	88.20
$\mathbb{S} \rightarrow \mathbb{T} : O(10)$	linear	93.51	94.29	85.97	87.89
	non-linear	93.28	94.08	84.91	84.56
$\mathbb{S} \rightarrow \mathbb{T} : O(100)$	linear	93.88	**94.44**	85.56	86.38
	non-linear	93.71	93.98	86.18	87.38
$\mathbb{S} \rightarrow \mathbb{T} : O(1000)$	linear	94.40	93.96	85.44	84.97
	non-linear	94.39	94.10	**86.45**	**88.61**

4. CONCLUSIONS

We proposed an approach to embedding projection for improving query understanding, in terms of intent classification and slot tagging. Experiments on two datasets showed that this technique can improve results of using either source or target embeddings only. Our approach combines the benefits of leveraging more reliable embeddings learned from large data and the adaption to task-specific languages. Moving forward, we are interested in studying the effect in cold start problems and different ways to creating projections.

5. REFERENCES

[1] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pages 96–101, 1990.

[2] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, abs/1412.6980, 2014.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*, abs/1301.3781, 2013.

[4] F. Tian, B. Gao, E. Chen, and T. Liu. Learning Better Word Embedding by Asymmetric Low-Rank Projection of Knowledge Graph. *J. Comp. Sci. Tech.*, 31(3):624–634, 2016.

[5] Y. Zhang, D. Gaddy, R. Barzilay, and T. Jaakkola. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *Proceedings of the NAACL-HLT 2016*, pages 1307–1317, June 2016.

[1] https://dumps.wikimedia.org/enwiki/latest/