Dual-enhanced Word Representations based on Knowledge Base

Fangyuan He[1], Yi Zhou[2,*], Haodi Zhang[3], Zhiyong Feng[4]

[1] School of Computer Science and Technology,Tianjin University
[2] School of Computing, Engineering and Mathematics, Western Sydney University
[3] College of Computer Science and Software Engineering, Shenzhen University
[4] School of Software,Tianjin University

Abstract. In this paper, we propose an approach for enhancing word representations twice based on large-scale knowledge bases. In the first layer of enhancement, we use the knowledge base as another contextual form corresponding to the corpus and add it to the training of distributed semantics including neural network based and matrix-based. In the second layer, we utilize local features of the knowledge base to enhance the word representations by mutual reinforcement between the keyword and the strongly associated words. We evaluate our approach not only on the well-known datasets but also on a brand-new dataset, IQ-Synonym-323. The results show that our approach compares favorably to other word representations.

1 Introduction

Word representations as the fundamental tool of NLP become increasingly important research. Currently, distributional semantics models that follow the distributional hypothesis represent the most popular approach of word representations. They commonly refer to the statistics derived from a large text corpus. Meanwhile, it is proved that the larger the corpus is, the better the model performs in most tasks [2, 4]. However, it also incurs obvious limitations. In [6] it's shown that using the specific-domain corpus has a definite advantage in addressing a specific task. With the expansion of corpus, it covers wider domains and concomitantly produces more mixed information in the context. Models relying on corpora as context will therefore be hindered in accuracy.

As another approach of word representations which attracts increasing attention, the knowledge-based approaches mainly rely on the external structured databases. The abundant and explicit lexical relationships between lexical items in databases can just make up for the blurring of contexts in the large corpus.

In this work, we propose an approach with double enhancement based on the large lexical database. Firstly, we take the related words in knowledge base as the additional accurate context in comparison with the large corpus. Afterwards, inspired from Kiela et al. [1], both contexts are added to the training process of representative distributional semantics for the first enhancement. In addition, we take advantage of the related words again to construct the second layer enhancement, which is a tuning process to highlight the strongly associated

words in extracted knowledge base. Our approach with double enhancement has outstanding performance on the benchmarks including SimLex-999 and the brand-new dataset IQ-Synonym-323 we build.

2 Approach

2.1 Knowledge Base as Accurate Context for Training

The knowledge base we use in our approach is composed of a large number of one-to-many relationship structures, i.e. given a keyword, the knowledge base will list its closest semantically related words. Therefore, in the first enhancement of our approach, we take the related words provided by knowledge base as the relative accurate context of keywords and inject them in the existing representative distributional semantics models. For comparison, we select skip-gram [2] and GloVe [4] representing neural network based and matrix based approach respectively of distributional semantics models.

Neural Network Based The original skip-gram is the neural network framework with a single hidden layer. Its basic idea is to predict the maximum probability of words appearing near the keyword. After the first step of the original training in large text corpus, our added step follows the formula (1). The objective of the second step is to maximize the following average log probability. w_1, w_2,... w_T is a sequence of training words. For keyword w_t, A_{w_t} is the set of its related words. And the length of the set is regarded as the context window size in the additional training step. We name this approach SG-KB-I.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{w^a \in A_{w_t}} \log p\left(w^a|w_t\right) \tag{1}$$

Matrix Based GloVe is an unsupervised learning approach which emphasizes the superiority of ratio in words' relevance and train log-bilinear regression model based on a global word-word co-occurrence matrix.

The cell of original matrix is the co-occurrence frequency of words in the fixed-length context window of the text corpora. In our approach to attaching knowledge base as accurate context, we add the co-occurrence frequency of keyword-related word in knowledge base to the original matrix. In this way, we use the modified cell values to adjust the degree of association between words. Then the original algorithm is applied to the new matrix to promote the word representations. We called this approach GloVe-KB-I.

2.2 Enhancement Based on Features of Knowledge Base

Within our extracted knowledge base, some pairs of words are mutual related. For instance, for the keyword "people", "human" is one of its related words in knowledge base. Meanwhile, "people" is also in the related word set of keyword "human". We consider "people" and "human" as a strongly associated words pair. For these word pairs, we attempt to tune their representations by mutual

reinforcement. In the formula (2), W_{sr}^{n} is the set of strongly associated words of keyword w, the v_{sr} is the vector of the elements in this set, n is the length of the set. W_{cr}^{m} is the set of commonly associated words of keyword w, v_{cr} is their vector, the number is m. W_{sr}^{n} and W_{cr}^{m} together form the related words set of keyword w in knowledge base. We set a weight value α to the strongly associated words, so that to pull keyword closer to the strongly associated words than the commonly ones.

$$v_w = \frac{1}{n * \alpha + m} \left(\alpha * \sum_{v_{sr} \in W_{sr}^{n}} v_{sr} + \sum_{v_{cr} \in W_{cr}^{m}} v_{cr} \right) \tag{2}$$

Afterwards, we use the SG-KB-I, GloVe-KB-I, as the initial vectors v_i. We tune each keyword's vector v_t by v_w and v_i. γ and β are the weight coefficients.

$$v_t = \gamma * v_w + \beta * v_i$$

2.3 Knowledge Base

Compared with raw corpus data, the knowledge base demonstrates more clarified relations between words. We choose two large lexical databases as our sources, namely WordNet and ConceptNet, which contain adequate concepts and a very broad range of word relationships. We extract more than 155 thousand keywords from WordNet, and 766 thousand from ConceptNet. After combining the two parts with mutual lexical items, we finally get 777 thousand keywords with related words, to constitute our lexical relation knowledge base.

3 Experimental Evaluation

3.1 Dataset

We evaluate our representations not only on the well-known dataset but also on the brand-new dataset we build. We construct a new dataset by collecting 323 synonym questions from related real IQ test books and websites for testing human intelligence, and name it IQ-Synonym-323. The questions we collected in our 323 synonym dataset have several types, like *"Choose the word most similar in meaning to X?"*, or *"Which word is closest to the X?"* , etc. But all these types can be included as the keyword and candidate words then we reorganized them. Our dataset will be available as an open source. Table 1 shows a sample.

3.2 Experimental Result

We choose a 11G dumps of English Wikipedia as text corpus. Table 2 shows the performances of all comparison approaches, including skip-gram and GloVe as the starting points, ConceptNet [5] and Counter-fitting [3] as state of the art models, SG-KB-I and GloVe-KB-I mentioned in the first layer of the approach, and SG-KB-II, GloVe-KB-II, two results trained by the second layer with different initial values. Comparing with the starting points, both layers of our approach improve the performance on the two benchmarks. SG-KB-II performs best on our dataset. Counter-fitting which takes the embedding tuned by SimLex-999 as start point has a particular advantage in this benchmark, however, it does not perform so well on our IQ-Synonym-323.

IQ question format	*"Which word is closest to the AUGUST?"* A.Common B. Ridiculous C. Dignified D. Petty **Answer: C**
Reorganized format	AUGUST::Common,Ridiculous,Dignified,Petty::Dignified

Table 1: A sample in IQ-Synonym-323

Approach	SimLex-999	IQ-Synonym-323
skip-gram	0.39	60.14%
GloVe	0.35	59.61%
GloVe-KB-I	0.44	72.34%
SG-KB-I	0.60	75.25%
GloVe-KB-II	0.55	80.85%
SG-KB-II	0.64	**84.08**%
ConceptNet	0.61	81.19%
Counter-fitting	**0.74**	64.75%

Table 2: Comparing the performances of the start points, our approaches and state of the art models on SimLex-999 and IQ-Synonym-323.

4 Conclusion

In this paper, we propose a double enhancement approach relying on knowledge base. Since knowledge base can specify more accurate related words of keywords as context information, we use it to compensate for the noises generated by multiple domains covered by the large corpus. Utilizing the features of knowledge base twice brings two significant improvements, as shown in Table 2. We evaluate our approach on the well-known SimLex-999, and the brand-new dataset, IQ-Synonym-323. The outstanding performance explains the advantage of our approach in embracing more accurate semantic similarity between similar vocabularies under large-scale corpora.

References

1. Kiela, D., Hill, F., Clark, S.: Specializing Word Embeddings for Similarity or Relatedness. In: EMNLP. pp. 2044–2048 (2015)
2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013)
3. Mrki, N., Séaghdha, D.Ó., Thomson, B., Gai, M., Rojas-Barahona, L.M., Su, P., Vandyke, D., Wen, T., Young, S.J.: Counter-fitting Word Vectors to Linguistic Constraints. CoRR abs/1603.00892 (2016), `http://arxiv.org/abs/1603.00892`
4. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
5. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: AAAI. pp. 4444–4451 (2017)
6. Stenetorp, P., Soyer, H., Pyysalo, S., Ananiadou, S., Chikayama, T.: Size (and Domain) Matters: Evaluating Semantic Word Space Representations for Biomedical Text. Proceedings of SMBM 12 (2012)