Extensions on Distributed Representation of Words and Phrases

Ruikang Zhao

1. Introduction

In natural language processing, grouping similar words in a vector space, which is also called distributed representation of words, could also achieve a better performance. One of the methods of learning vector representations of words from huge sets of text data is the Skip-gram model, introduced by Mikolov et al. [2]. The major advantage of Skip-gram model is its high efficiency of training because it does not implement matrix multiplication. Its patterns are also interesting in that they can be represented as linear translations: vec(“Madrid”) – vec(“Spain”) + vec(“France”) is very close to vec(“Paris”). [2, 3]

Even though Skip-gram model is good enough, there are still several existing extensions of the original Skip-gram model to improve both the training speed and the quality of vectors, including subsampling, simplified variant of Noise Contrastive Estimation(NCE), transformation from word-based vector to phrase-based vector representation, and basic mathematical operations on vectors.

2. Existing extensions on the Skip-gram model

- Subsampling of frequent words:

In large set of text data, the most frequent words usually carry less important information than are words. For example, existence of words like “in”, “the”, “a” increases meaningless co-occurrence with other words. Since the effectiveness of Skip-gram model benefit a lot from analyzing the co-occurrence of different words, the high frequency of meaningless words could harm the performance of the model. By applying subsampling formula to words with frequency larger than a threshold, we can improve the accuracy of obtained vectors of rare words.

- Simplified variant of Noice Contrastive Estimation(NCE)

In the training of the Skip-gram model, many softmax functioncan be used to define p(w_t+j|w_t) (the relative probability of word t and j), like hierarchical softmax, that use a binary tree representing relative probability in child nodes, and Noise Contrastive Estimation(NCE) that applies logistic regression to distinguish data from noise. Since the Skip-gram model only learns high-quality vector representations, softmax terms can be replaced by negative sampling [1].

- Transformation from word-based vector to phrase-based vector representation

A huge limitation of word-based vector representation is the inability of representing idiomatic phrases. For example, “Rolling Stone” is a magazine, so it is not a simple semantic combination of “Rolling” and “Stone”. Therefore, we can implement an extension to a phrase-based model by initially identify a large set of tokens that contains words and phrases.

- Vector arithmetic

An interesting result from the Skip-gram model is that it performs precise analogical reasoning when applying simple vector arithmetic. For example, vec("Germany") + vec("capital") is close to vec("Berlin"), and vec("Russia") + vec("river") is close to vec("Volga River") [1].

Since the Skip-gram model analyzes co-occurrences trained in such a way, the word vectors could predict the surrounding words in the context. Another way to explain the arithmetic from the perspective of semantic graph is that "Berlin" is a subcategory under both "Germany" and "capital", so it should be on the path between them.

3. Conclusion

Existing extensions on the Skip-gram model have several important impacts: simpler way to train the model, better precision by applying subsampling and transformation to phrase-based model. In addition, applying the methodology of semantic graph would better explain the arithmetic patterns on vector representations.

References

[1] Tomas Mikolov, Ilya SutsKever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and Phrases and their Compositionality. NeurIPS, 2013.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

[3] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.