1 How to Generate a Good Word Embedding?

Paper by Lai et al.[1]

- Word embedding, also known as distributed word representation, can capture both the semantic and syntactic information of words from a large unlabeled corpus and has attracted considerable attention from many researchers. In recent years, several models have been proposed, and they have yielded state-of-the-art results in many natural language processing (NLP) tasks.
- We observe that almost all methods for training word embeddings are based on the same distributional hypothesis: words that occur in similar contexts tend to have similar meanings.
- Training on a large corpus generally improves the quality of word embeddings, and training on an in-domain corpus can significantly improve the quality of word embeddings for a specific task.
- Previous works have shown that models that predict the target word capture the paradigmatic relations between words
- we can conclude that using a larger corpus can yield a better embedding, when the corpora are in the same domain
- In most of the tasks, the influence of the corpus domain is dominant. In different tasks, it impacts performance in the different ways
- The corpus domain is more important than the corpus size. Using an indomain corpus significantly improves the performance for a given task, whereas using a corpus in an unsuitable domain may decrease performance.

References

[1] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016.