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.

2 Better Word Representations with Recursive Neural Networks for Morphology

Paper by Luong et al.[2]

- The use of word representations or word clusters pretrained in an unsupervised fashion from lots of text has become a key "secret sauce" for the success of many NLP systems in recent years, across tasks including named entity recognition, part-ofspeech tagging, parsing, and semantic role labeling.
- The main advantage of having such a distributed representation over word classes is that it can capture various dimensions of both semantic and syntactic information in a vector where each dimension corresponds to a latent feature of the word. As a result, a distributed representation is compact, less susceptible to data sparsity, and can implicitly represent an exponential number of word clusters.
- The Rare-word dataset introduced by Luong et al.

3 Evaluation methods for unsupervised word embeddings

Paper by Schnabel et al. [3]

- Neural word embeddings represent meaning via geometry. A good embedding provides vector representations of words such that the relationship between two vectors mirrors the linguistic relationship between the two words.

- Existing schemes fall into two major categories: extrinsic and intrinsic evaluation. In extrinsic evaluation, we use word embeddings as input features to a downstream task and measure changes in performance metrics specific to that task. Examples include part-of-speech tagging and named-entity recognition (Pennington et al., 2014). Extrinsic evaluation only provides one way to specify the goodness of an embedding, and it is not clear how it connects to other measures. Intrinsic evaluations directly test for syntactic or semantic relationships between words (Mikolov et al., 2013a; Baroni et al., 2014).

- This is the first paper to conduct a comprehensive study covering a wide range of evaluation criteria and popular embedding techniques. In particular, we study how outcomes from three different evaluation criteria are connected: word relatedness, coherence, downstream performance.

- Embedding methods should be compared in the context of a specific task, e.g., linguistic insight or good downstream performance.

- We observe that word embeddings encode a surprising degree of information about word frequency. We found this was true even in models that explicitly reserve parameters to compensate for frequency effects. This finding may explain some of the variability across embeddings and across evaluation methods. It also casts doubt on the common practice of using the vanilla cosine similarity as a similarity measure in the embedding space.

- Extrinsic evaluations measure the contribution of a word embedding model to a specific task. There is an implicit assumption in the use of such evaluations that there is a consistent, global ranking of word embedding quality, and that higher quality embeddings will necessarily improve results on any downstream task. We find that this assumption does not hold: different tasks favor different embeddings

- Performance on downstream tasks is not consistent across tasks, and may not be consistent with intrinsic evaluations. Comparing performance across tasks may provide insight into the information encoded by an embedding, but we should not expect any specific task to act as a proxy for abstract quality.

- Word frequency information in the embedding space also affects cosine similarity

- Also, the above results mean that the commonly-used cosine similarity in the embedding space for the intrinsic tasks gets polluted by frequency-based effects
- Factors such as word frequency play a significant and previously unacknowledged role. Word frequency also interferes with the commonly-used cosine similarity measure.

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.

[2] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, 2013.

[3] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, 2015.