Super awesome embeddings

Grzegorz Beringer, Mateusz Jabłoński, Piotr Januszewski, and Julian Szymański

Faculty of Electronic Telecommunications and Informatics
Gdańsk University of Technology, Gdańsk, Poland

Abstract. Abstract

Keywords: word sense disambiguation

1 Introduction

Word Sense Disambiguation (WSD) is an open problem of natural language processing (NLP) and ontology. WSD is identifying which sense of a word (i.e. meaning) is used in a sentence based on the word context. Difficulty is when the word has multiple meanings (e.g. a decision tree, a tree data structure, a tree in a forest). The problem requires two inputs: a dictionary to specify the senses which are to be disambiguated and a corpus of language data to be disambiguated. WSD task has two variants: "lexical sample" and "all words" task. The former aim to disambiguate the occurrences of a small sample of selected target words, while in the latter all the words in a piece of running text need to be disambiguated. Our solution target the former one, but could be extended to the latter variant. The solution to WSD would be useful in many NLP related problems as: relevance of search engines, anaphora resolution, coherence, inference, etc.

Word embeddings are product of feature learning techniques in NLP where words from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a dimensionality reduction from a space with one dimension per word to a continuous vector space with a much lower dimension. Methods to generate this mapping include artificial neural networks[1][2][3], dimensionality reduction on the word co-occurrence matrix[4] and probabilistic models[5]. Word embeddings are commonly used as the input representation. They have been shown to boost the performance in NLP tasks such as syntactic parsing[6] and sentiment analysis[7].

Word embeddings can't distinguish between different meaning of words. There is only one embedding for each word e.g. for word "tree" there is only one real valued vector. It means that word embeddings doesn't distinguish between a decision tree, a tree data structure or a tree in a forest. What we can do is try to distinguish the meaning base on the context in which the word was used. Then, we treat each meaning as a keyword, which has its own embedding. We propose simple method to infer the word meaning: an average of the context and the word and number of improvements to this approach in the chapter "Method". Experiments with our approach are presented in the chapter "Experiments". To conduct those experiments we've created the dataset composed of 6 keywords with 4 to 7 meanings to disambiguate from and numerous texts with tagged words to be disambiguated. We describe our dataset in the chapter "Dataset". In the chapter "Related work" we present other approaches to WSD.

2 Related work

3 Dataset

References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. arXiv e-prints (2013) arXiv:1310.4546
2. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: In EMNLP. (2014)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14, Cambridge, MA, USA, MIT Press (2014) 2177–2185

5. Globerson, A., Chechik, G., Pereira, F., Tishby, N.: Euclidean embedding of co-occurrence data. J. Mach. Learn. Res. **8** (2007) 2265–2295
6. Socher, R., Bauer, J., Manning, C.D., Andrew Y., N.: Parsing with compositional vector grammars. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2013) 455–465
7. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2013) 1631–1642