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, word embeddings

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 targets 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 a 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 cannot distinguish between different meanings of ambiguous words by themselves. By definition, there is only one embedding for each word e.g. for word “tree” there is a single real-valued vector. What can be done, is to try to distinguish the meaning based on the context, in which the word was used. Then, we treat each meaning as a separate keyword, which has its own embedding. We propose a simple method to infer the word meaning: an average of the context and the word embeddings and number of improvements to this approach in the chapter “Our method”. Experiments with our solution are presented in the chapter “Experiments”. To conduct those experiments we have created the dataset composed of 6 ambiguous words, with 4 to 7 meanings each, and collected real-world usage examples of those meanings, 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

Typically, there are two kinds of approach for WSD: supervised, which make use of sense-annotated training data, and knowledge-based, which make use of the properties of lexical resources. In supervised approach, the most widely used training corpus used is SemCor[8], with 226,036 sense annotations from 352 manually annotated documents. Knowledge-based systems usually exploit WordNet[9] or BabelNet[10] as semantic network. Our solution join two approaches. We use the knowledge-based word embeddings like Word2vec[1] and use them to create the meaning embeddings in supervised fashion.

The most usual baseline is the Most Frequent Sense[11] (MFS) heuristic, which selects for each target word the most frequent sense in the training data. Recent growth of sequence learning techniques using artificial neural networks contributed to WSD research: Raganato et al.[12] propose a series of end-to-end neural architectures directly tailored to the task, from bidirectional Long

Short-Term Memory (LSTM) to encoder-decoder models. Melamud et al.[13] also use bidirectional LSTM it their work. They use large plain text corpora to learn a neural model that embeds entire sentential contexts and target words in the same low-dimensional space, which is optimized to reflect inter-dependencies between targets and their entire sentential context as a whole.

Iacobacci et al.[14] were first to try to use word embeddings for WSD. They consider four different strategies for integrating a pre-trained word embeddings as context representation in a supervised WSD system: concatenation, average, fractional and exponential decay of the vectors of the words surrounding a target word. Peters et al.[15] create word representations that differ from traditional word embeddings in that each token is assigned a representation that is a function of the entire input sentence. They use vectors derived from a bidirectional LSTM that is trained with a coupled language model objective on a large text corpus.

Most of the previous neural networks applications to WSD ignore lexical resources like glosses (sense definitions) and relay solely on word's context. In Luo et al.[16] paper, they integrate the context and glosses of the target word into a unified framework in order to make full use of both labeled data and lexical knowledge.

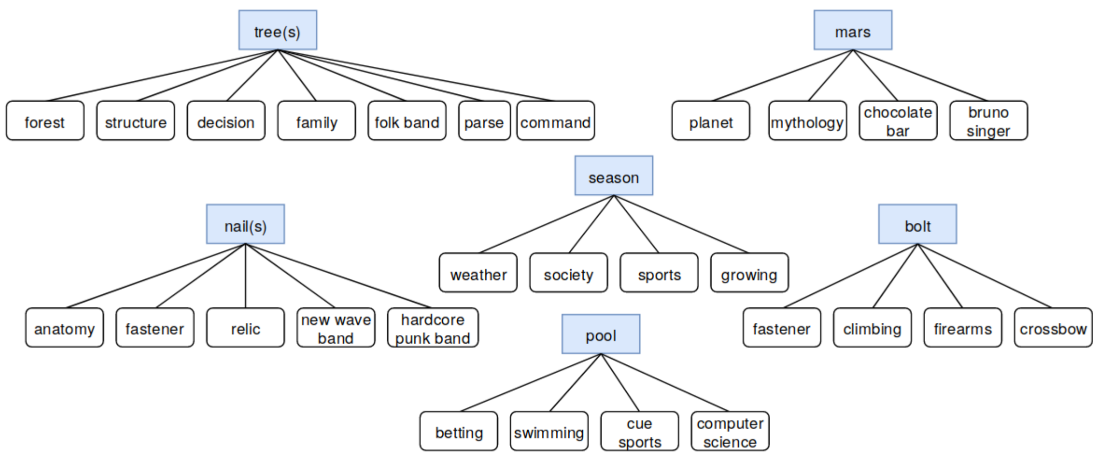
Entity Linking (EL) and WSD both address the lexical ambiguity of language. The aim of EL is to discover mentions of entities within a text and to link them to the most suitable entry in a reference knowledge base. The two tasks are pretty similar, but they differ fundamentally: in EL the textual mention can be linked to a named entity which may or may not contain the exact mention, while in WSD there is a perfect match between the word form and a suitable word sense in the knowledge base. Moro et al.[17] present Babelfy, a unified graph-based approach to EL and WSD based on a loose identification of candidate meanings coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations. We can find also application of random walks[18] and topic models[19] for knowledge-based WSD

Developing WSD system requires much effort and as a result, very few open source WSD systems are publicly available. Zhong et al.[20] present an English all-words WSD system, IMS (It Makes Sense), built using a supervised learning approach that is written in Java and completely open source. Following Lee and Ng[21], they adopt support vector machines (SVM) as the classifier and integrate multiple knowledge sources including parts-of-speech (POS), surrounding words, and local collocations as features.

3 Dataset

For the purpose of testing word embeddings as a method to differentiate between different meanings, we gathered examples for 6 ambiguous words, 4-7 meanings each (28 meanings in total). Ambiguous word together with its meaning constitues a *keyword*, which we use as a separate class when identyfing the closest meaning given some context. All keywords can be seen on Figure 1.

Fig. 1. Ambiguous words with their meanings (keywords) from the dataset



Examples were mostly gathered from Wikipedia, using *What links here* utility for each keyword. If usage examples from Wikipedia were not enough, other websites were used (or even the Wikipedia article on specific keyword itself).

The dataset is split into training and test set, with 5 training and 10 test examples for each keyword. Each example is stored in plain text, with the ambiguous word marked with "*" on both sides.

The correct keyword for each example, together with a path to file and a link, where the original text was taken from, are stored in CSV files: *train.csv* for training set, *test.csv* for test set (columns: path,keyword,link). Keywords themselves, together with links to their Wikipedia articles, are stored in *keywords.csv* file. Dataset, together with the code to execute experiments from this paper, can be found on our GitHub repository [22].

<u>Example for *bolt crossbow* keyword:</u>

Keyword definition: keyword.csv

keyword,link

...

bolt crossbow,https://en.wikipedia.org/wiki/Crossbow_bolt

Test set: test.csv

path,keyword,link

...

texts/test/bolt_crossbow_5.txt,bolt crossbow,https://en.wikipedia.org/wiki/Incendiary_device

Text: texts/test/bolt_crossbow_5.txt

*"Sulfur- and oil-soaked materials were sometimes ignited and thrown at the enemy, or attached to spears, arrow and *bolts* and fired by hand or machine. Some siege techniques—such as mining and boring—relied on combustibles and fire to complete the collapse of walls and structures."*

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
8. Mihalcea, R.: Semcor semantically tagged corpus. (1998)
9. : Wordnet, Princeton University (2010)
10. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
11. Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics (2017) 99–110
12. Raganato, A., Delli Bovi, C., Navigli, R.: Neural sequence learning models for word sense disambiguation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2017) 1156–1167

13. Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning generic context embedding with bidirectional lstm. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics (2016) 51–61
14. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Embeddings for word sense disambiguation: An evaluation study. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2016) 897–907
15. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics (2018) 2227–2237
16. Luo, F., Liu, T., Xia, Q., Chang, B., Sui, Z.: Incorporating glosses into neural word sense disambiguation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2018) 2473–2482
17. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics **2** (2014) 231–244
18. Agirre, E., López de Lacalle, O., Soroa, A.: Random walks for knowledge-based word sense disambiguation. Computational Linguistics **40** (2014) 57–84
19. Chaplot, D.S., Salakhutdinov, R.: Knowledge-based word sense disambiguation using topic models. In: AAAI. (2018)
20. Zhong, Z., Ng, H.T.: It makes sense: A wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 System Demonstrations, Association for Computational Linguistics (2010) 78–83
21. Lee, Y.K., Ng, H.T.: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 41–48
22. Beringer, G., Jablonski, M., Januszewski, P., Szymański, J.: https://github.com/gberinger/automatic-wiki-links (2018)