Combining embedding methods for a word intrusion task

Finn Årup Nielsen and Lars Kai Hansen

Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark

Abstract. We report a new baseline for a Danish word intrusion task by combining pre-trained off-the-shelf word, subword and knowledge graph embedding models.

In the word intrusion task, a cognitive agent is presented with a set of words and is to determine the odd-one-out. In [5], we constructed a word intrusion dataset with Danish words and evaluated how well different machine-based methods could identify the intruded word. Explicit semantic analysis and a Word2vec-based word embedding with large corpora performed the best with performances of 73% and 71%, respectively, against a baseline of 25%. Since [5], new embedding methods have arisen with pre-trained models for non-English languages, e.g., fastText (FT) [2], Byte-Pair Encoding (BPE) [3] and BERT [1], and we have constructed Wembedder (W), a knowledge graph embedding based on the multilingual Wikidata knowledge base [4].

Our word intrusion dataset comprises 100 sets of 4 words each where one of 4 is the outlier to be detected [5].[1] The dataset contains common and proper nouns (named entities) and other word classes as well as a few numbers, years and phrases. Some sets of "words" require detailed Danish world knowledge, e.g., 1807, 1864, 1940, 1909, — the last being the outlier. There are also a number of homographs.

We use FastText through Gensim with the FastText `cc.da.300.bin` pre-trained model. For BERT, we use the currently recommended cased multilingual model[2] through the package bert-as-service.[3] The BPE model comes in various sizes of vocabulary and embedding dimensions and we test them all. Wembedder is an embedding of Wikidata items rather than words, and the use of Wembedder for natural language requires a translation from the word to the Wikidata item identifier. We use the Wikidata search API[4] and its `wbsearchentities` action to search for Wikidata items based on the queried word. Not all words can be found in Wikidata, e.g., adjectives and verbs are rarely present as Wikidata items, meaning words from such word classes are usually out-of-vocabulary. Over

[1] https://github.com/fnielsen/dasem/blob/master/dasem/data/four_words_2.csv. We use the second version correcting two spelling errors.

[2] `multi_cased_L-12_H-768_A-12` from https://github.com/google-research/bert/blob/master/multilingual.md

[3] https://github.com/hanxiao/bert-as-service

[4] https://www.wikidata.org/w/api.php

Model	FT	BPE	BERT	W	FT+W	FT+W+BERT	Baseline
Accuracy	78	69	32	47	82	83	25

Table 1. Odd-one-out detection percentage.

1000 entities for Danish words and phrases exist as lexemes on Wikidata, but the current Wembedder models has no Wikidata lexemes.

There are multiple ways of getting from a vectorial representation to a measure of outlierness. For Gensim-based models, we use Gensim's `doesnt_match` method. For the other embeddings, we sort the row sum of the correlation matrix of the concatenated embedding vectors of the four words and select the word associated with the lowest sum. The performance of a model is measured as the percentage of correctly detected outliers.

Results are displayed in Table 1. FastText alone can improve the benchmark to 78%, while the BPE embeddings cannot reach a better performance than our previous results. Its accuracies range from 33% to 69%, depending on dimension and vocabulary. Our current simple application of BERT does not yield good performance with only an accuracy of 32%. Wembedder neither performs well with just 47% accuracy. However, it tends to perform well on proper nouns, better than (sub-)word embedding models: We can attain an accuracy of 82% by combining fastText and Wembedder using Wembedder for entries with non-lower first letters (named entities). We can improve that performance slightly to 83% by using BERT for phrases which are not named entities (as we only have 100 tests these improvements are not statistically strong). The 17 errors made form a heterogeneous set. A handful of them may well be due to homographs, e.g., 'tog' (either 'train' or 'took') and 'kassen' ('the box'), where the Wembedder search identifies the latter as the surname 'Kassen' (Q37436530).

Acknowledgment: This research is funded by the Innovation Foundation Denmark through the DABAI and the ATEL projects.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018), https://arxiv.org/pdf/1810.04805.pdf
2. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages. LREC (2018), https://arxiv.org/pdf/1802.06893.pdf
3. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. LREC (2018), http://www.lrec-conf.org/proceedings/lrec2018/pdf/1049.pdf
4. Nielsen, F.Å.: Wembedder: Wikidata entity embedding web service (2017), https://arxiv.org/pdf/1710.04099
5. Nielsen, F.Å., Hansen, L.K.: Open semantic analysis: The case of word level semantics in Danish. LTC (2017), http://www2.compute.dtu.dk/pubdb/views/edoc_download.php/7029/pdf/imm7029.pdf