Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Combining embedding methods for a word intrusion task

Finn Årup Nielsen
Cognitive Systems, DTU Compute
Technical University of Denmark
Kongens Lyngby, Denmark

Lars Kai Hansen
Cognitive Systems, DTU Compute
Technical University of Denmark
Kongens Lyngby, 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. We test fastText, Byte-Pair Encoding, BERT and the knowledge graph embedding in Wembedder, finding fastText as the individual model with the superior performance, while a simple combination of the fastText with other models can slightly improve the accuracy of finding the odd-one-out words in the word intrusion task.

In the word intrusion task, see, e.g., (Chang et al., 2009), a cognitive agent is presented with a set of words and is to determine the odd-one-out. Such a test has been used to evaluate unsupervised topic models (Chang et al., 2009) and human subjects in experimental psychology, see, e.g., (Crutch et al., 2008). The test somewhat resembles *Test of English as a Foreign Language* (TOELF), where the task is to select the semantically most similar one among four words given a query word (Turney, 2006). A convenient method (`doesnt_match`) is implemented in the distributed semantics models of Gensim (Řehůřek and Sojka, 2010; Wohlgenannt et al., 2019), giving users of this Python package a straightforward way to test trained machine learning models in odd-one-out tasks.

(Nielsen and Hansen, 2017) 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 random choice baseline of 25%. Since (Nielsen and Hansen, 2017), new embedding methods have appeared with pre-trained models for non-English languages, e.g., fastText (FT) (Bojanowski et al., 2016; Grave et al., 2018), Byte-Pair Encoding (BPE) (Heinzerling and Strube, 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), and Wembedder (W), a knowledge graph embedding based on the multilingual Wikidata knowledge base (Nielsen, 2017). We note that some of the best performing semantic models have combined corpus-based and explicit lexicon-/knowledge graph-based methods (Turney, 2006; Speer and Lowry-Duda, 2017), and we will also pursue such a combination here.

Below we will describe the Danish word intrusion task dataset used for evaluation, the applied new off-the-shelf methods, their results in terms of accuracy of detecting the odd-one-out and finally we discuss what further approaches are needed to handle the remaining misclassified cases.

1 Evaluation dataset

The word intrusion dataset comprises 100 sets of 4 words each where one of 4 is the outlier to be detected (Nielsen and Hansen, 2017),[1] see the left part of Figure 1 for a small excerpt of the dataset. 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 as the three first years relates to Danish military defeats. The dataset contains also several homographs/polysemous words. Most of the words are common nouns. There are 11 word sets with proper nouns and 11 with verbs. Further sets includes sets of adjective and other word classes. A few of the word sets mix lexical categories, e.g., the set (halvsyg, forkølelse, hoster, vej) corresponding to the English ("half-sick", flu, the verb "coughs", road).

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

word 1	Word 2	word 3	word 4	FT	BERT	W	FT+W+BERT
æble (apple)	pære (pear)	kirsebær (cherry)	stol (chair)	stol	kirsebær	kirsebær	stol
bil (car)	cykel (bike)	tog (train)	vind (wind)	tog	bil	bil	tog
Finland (Finland)	Sverige (Sweden)	Norge (Norway)	Kina (China)	Kina	Norge	Kina	Kina
tres (sixty)	60 (60)	LX (LX)	3 (3)	tres	LX	LX	LX

Table 1: Excerpt of the evaluation dataset and individual results from fastText (FT), BERT, Wembedder (W) and the combined system of fastText, BERT and Wembedder (FT+W+BERT). The ground truth outlier is in the *word 4* column.

While word intrusion tasks may be based on the sound of words, see, e.g., (Oakhill et al., 2003), the Danish dataset contains none of this kind, so the methods we employ need no phonological information.

2 Methods

We use fastText (Bojanowski et al., 2016; Grave et al., 2018) through the Gensim 3.6.0 implementation (Řehůřek and Sojka, 2010) with the fastText `cc.da.300.bin` pre-trained model.[2] This model has been trained on the *Common Crawl* and the *Danish Wikipedia* with the continuous bag-of-words setup. In terms of training corpus size, it may be the largest publicly available linear word embedding model and as such should be regarded as a baseline model. We downloaded it from its homepage.[3]

For BERT, we use the currently recommended cased multilingual model[4] through the package bert-as-service.[5]

The BPE model comes in various sizes of vocabulary and embedding dimensions and we test them all.[6] The size of the vocabulary of the pre-trained distributed models ranges from 1,000 to 200,000 while the embedding dimension ranges from 25 to 300.

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[7] and its `wbsearchentities` action to search for Wikidata items based on the queried word or phrase. 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. The Wembedder model we use is the one trained on the 2017-06-13 truthy dump of Wikidata with an embedding dimension of 100 and using the continuous bag-of-word Word2vec approach implemented in Gensim.[8] The use of the Wikidata API means that results may not necessarily reproduce between runs of our evaluation, because Wikidata is continuously expanded and modified.

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.

Our computations are available in a public Jupyter Notebook.[9]

[2] https://fasttext.cc/docs/en/crawl-vectors.html.

[3] https://fasttext.cc/docs/en/crawl-vectors.html

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

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

[6] https://nlp.h-its.org/bpemb/da/.

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

[8] The Wembedder Gensim model was downloaded from https://zenodo.org/record/827339.

[9] https://gist.github.com/fnielsen/93f3b68941e74c468522f187e2dbe9a7.

Model	FT	BPE	BERT	W	FT+W	FT+W+BERT	Random
Accuracy	78	64	32	47	82	83	25

Table 2: Odd-one-out detection percentage for fastText (FT), BPE, BERT, Wembedder (W), fastText and Wembedder (FT+W) and the combined model of fastText, Wembedder and BERT (FT+W+BERT) against the random choice.

3 Results

The results are displayed in Table 2. 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. Generally, the performance increases considerably as the vocabulary increases, see Table 3. However, for the largest vocabulary (200,000) the accuracy decreases for the models with the largest embedding dimension. With respect to the dimension, the largest embeddings with sizes 200 and 300 yield the best performance. The increase in performance from low to high dimensional models is smaller than when the vocabulary size is changed. This difference could be explained by the different range: The vocabulary sizes differs by 200 times, while the embedding dimension only differ by 12 times.

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).

4 Discussion

Our best model detects 83 outliers out of 100. What is needed to improve the performance, handling the misclassified cases?

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) for the set (Nielsen, Jensen, Olsen, kassen). If we are to improve the model, it may be necessary to handle the homography/polysemy of words. It is likely that even larger corpora with the non-context embedding models such as the ordinary application of fastText may not be able to handle the cases with homographs.

Voc. \ Dim.	25	50	100	200	300
1,000	36	34	34	36	33
3,000	45	42	48	47	47
5,000	52	50	51	54	55
10,000	56	59	59	63	59
25,000	58	58	62	63	67
50,000	58	63	65	**69**	**69**
100,000	58	63	63	**69**	**69**
200,000	60	64	67	67	64

Table 3: BPE results. Percentage of correctly spotted outliers among four words for BPE models of varying sizes: vocabulary from 1,000 to 200,000 words and dimensions from 25 to 300.

Numbers pose a common problem for all the models. One of the tests is (tres, 60, LX, 3), where *tres* is the Danish word for sixty, *LX* is the Latin number 60 and 3 is the outlier. FastText chooses *tres*, while BERT, the largest BPE model and Wembedder report LX as the outlier, so modifying any ensemble weighting will not help. It is possible that a larger corpora could learn the relations, or that explicit entry of such information in the Wikidata knowledge graph could help.

The low performance of BERT may come as a surprise given that BERT has been reported with a string of state-of-the-art results (Devlin et al., 2018). We note that the benchmarks used in the original BERT report had input that was longer than a word (e.g., sentences), while our current application of BERT only submits one word at a time to the model. It is tempting to think that some form of multiple word input to BERT may perform better, e.g., where two or three of the four words in a word set are submitted at a time. Such an approach could also handle the homography/polysemy problem.

The measure of outlierness is based on

the cosine similarity implemented in Gensim's `doesnt_match` function and the correlation matrix. We note that an exploration and a more careful selection of the metric for comparison may yield different results.

Over 1,800 entities for Danish words, affixes and phrases exist as lexemes on Wikidata (Nielsen, 2019), but the current Wembedder models have no Wikidata lexemes. Knowledge graph embedding that includes the relatively new Wikidata lexemes and its connection to the Danish wordnet DanNet (Pedersen et al., 2009) may be a fruitful avenue for further study.

5 Acknowledgment

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

References

[Bojanowski et al.2016] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2016. Enriching Word Vectors with Subword Information. July. https://arxiv.org/pdf/1607.04606.pdf.

[Chang et al.2009] Jonathan Chang, Jordan Boyd-Graber, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, pages 288–296.

[Crutch et al.2008] Sebastian J Crutch, Sarah Connell, and Elizabeth K Warrington. 2008. The different representational frameworks underpinning abstract and concrete knowledge: evidence from odd-one-out judgements. *Quarterly Journal of Experimental Psychology*, 62:1377–88, 1388–90, December.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. October. https://arxiv.org/pdf/1810.04805.pdf.

[Grave et al.2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning Word Vectors for 157 Languages. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, February.

[Heinzerling and Strube2018] Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 2989–2993, May.

[Nielsen and Hansen2017] Finn Årup Nielsen and Lars Kai Hansen. 2017. Open semantic analysis: The case of word level semantics in Danish. *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 415–419, October.

[Nielsen2017] Finn Årup Nielsen. 2017. Wembedder: Wikidata entity embedding web service. October. https://arxiv.org/pdf/1710.04099.

[Nielsen2019] Finn Årup Nielsen. 2019. Danish in Wikidata lexemes. *Global WordNet Conference 2019*.

[Oakhill et al.2003] J.V. Oakhill, K. Cain, and P.E. Bryant. 2003. The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, 18:443–468, August.

[Pedersen et al.2009] Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, August.

[Speer and Lowry-Duda2017] Robert Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. April. https://arxiv.org/pdf/1704.03560.pdf.

[Turney2006] Peter D. Turney. 2006. Similarity of Semantic Relations. *Computational Linguistics*, 32:379–416, September.

[Wohlgenannt et al.2019] Gerhard Wohlgenannt, Ekaterina Chernyak, Dmitry Ilvovsky, Ariadna Barinova, and Dmitry Mouromtsev. 2019. Relation Extraction Datasets in the Digital Humanities Domain and their Evaluation with Word Embeddings. March. https://arxiv.org/pdf/1903.01284.pdf.

[Řehůřek and Sojka2010] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. *New Challenges For NLP Frameworks Programme*, pages 45–50, May. https://radimrehurek.com/gensim/lrec2010_final.pdf.