Combining embedding methods for a word intrusion task

Finn Årup Nielsen and Lars Kai Hansen

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

Summary

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 fastText with other models can slightly improve the accuracy of finding the odd-one-out words in the word intrusion task.

Methods

Semantic representations can be evaluated in a number of ways. One way is with an **word intrusion task** (odd-one-out task).

In the word intrusion task[1] a cognitive agent is presented with a set of words and is to determine the odd-one-out.

We evaluate a few off-the-shelf distributed semantic representations for the word intrusion task:

1. FastText `cc.da.300.bin` pre-trained model through Gensim.
2. Byte-pair encoding (BPE) embedding. BPE models of varying sizes: vocabulary from 1,000 to 200,000 words and dimensions from 25 to 300.
3. BERT[2] `multi_cased_L-12_H-768_A-12` model through the package bert-as-service.
4. Wikidata knowledge graph embedding: Wembedder running from `https://tools.wmflabs.org/wembedder`

Odd-one-out evaluation dataset

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

Excerpt of the evaluation dataset[3] with 100 word sets 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.

The dataset has different word classes: common nouns, proper nouns, numerals, verbs, etc. and Danish world knowledge is required for some word sets, e.g., (1807, 1864, 1940, 1909).

Results

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

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.

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).

This work is funded by the Innovation Fund Denmark through DABAI and ATEL projects.

BPE results

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

BPE odd-one-out detection percentage.

What next

- Further exploration of deep learning embedding
- Handling of homography/polysemy
- Exploration of different measures of outlierness
- Application of Wikidata lexemes[4] and its connection to the Danish wordnet DanNet.

References

[1] Chang J, et al. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 2009;pages 288–296.
[2] Devlin J, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018;.
[3] Nielsen FÅ and Hansen LK. Open semantic analysis: The case of word level semantics in Danish. *Human Language Technologies as a Challenge for Computer Science and Linguistics*, 2017;pages 415–419.
[4] Nielsen FÅ. Danish in Wikidata lexemes. 2019;.