Detecting the odd-one-out among Danish words and phrases with embeddings

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

6 November 2019

Odd-one-out task

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

Odd-out-out or word intrusion task (Chang et al., 2009).

Detect the word that is an outlier compared to the other (here: four) words.

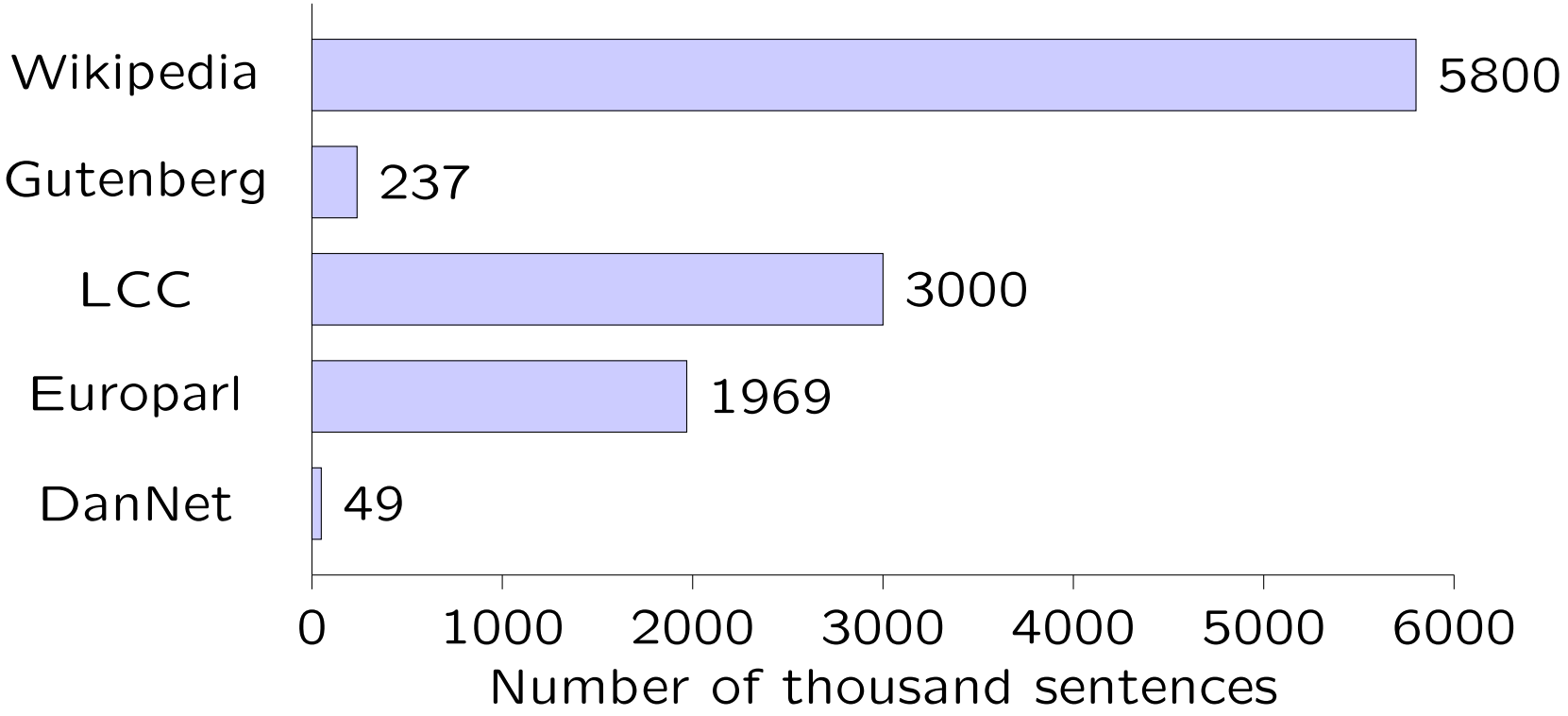
Resembles *Test of English as a Foreign Language* (TOELF)

Outlierness: Semantics, word class, sentiment, world knowledge, etc.

https://github.com/fnielsen/dasem/blob/master/dasem/data/four_words_2.csv

2017

Open semantic analysis: The case of word level semantics in Danish (Nielsen and Hansen, 2017): Assemble large Danish corpora and build various distributional semantics models.



2017

Accuracy	Corpus	Method
	Wikipedia	Explicit Semantic Analysis
	Gutenberg	Word2vec
	LCC	Word2vec
	Wikipedia	Word2vec
	Aggregate	Word2vec

Explicit Semantic Analysis (ESA) projects words into a Wikipedia article-spanned subspace (Gabrilovich and Markovitch, 2007).

Corpora handling and word2vec construction handled in Gensim-based Dasem at https://github.com/fnielsen/dasem

2017: Results

Accuracy	Corpus	Method
73	Wikipedia	Explicit Semantic Analysis
36	Gutenberg	Word2vec
69	LCC	Word2vec
71	Wikipedia	Word2vec
71	Aggregate	Word2vec

Explicit Semantic Analysis (ESA) projects words into a Wikipedia article-spanned subspace (Gabrilovich and Markovitch, 2007).

Corpora handling and word2vec construction handled in Gensim-based Dasem at https://github.com/fnielsen/dasem

2019

New pre-trained model available:

Facebook's pre-trained FastText model `cc.dan.300` from https://fasttext.cc (Bojanowski et al., 2016; Grave et al., 2018), includes subword modeling.

BERT (Devlin et al., 2018), deep learning

Byte-pair encoding (BPE) embedding (Heinzerling and Strube, 2018)

Wembedder (Nielsen, 2017) knowledge graph embedding from Wikidata data

2019: 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.

2019: 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.

Best combination model uses knowledge graph embedding (Wembedder) on named entities, BERT on phrases and FastText as a fall back (Nielsen and Hansen, 2019).

https://gist.github.com/fnielsen/93f3b68941e74c468522f187e2dbe9a7

2019: 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 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.

2020?

Better performance with a new version of Gensim with FastText!

What to do about homographs, e.g., (bil, cykel, *tog*, vind), (bibliotek, bog, *låner*, flag) or (*går*, spadserer, vandrer, siger)?

Wikidata lexemes (Nielsen, 2019): https://tools.wmflabs.org/ordia/language/Q9035

New corpora? retsinformation.dk: 8'395'616 sentences.

New attempt with deep learning models.

Thanks

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information.

Chang, J., Boyd-Graber, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, pages 288–296.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. N. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*.

Heinzerling, B. and Strube, M. (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.

Nielsen, F. Å. (2017). Wembedder: Wikidata entity embedding web service. DOI: 10.5281/ZENODO.1009127.

Nielsen, F. Å. (2019). Danish in Wikidata lexemes.

Nielsen, F. Å. and Hansen, L. K. (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.

Nielsen, F. Å. and Hansen, L. K. (2019). Combining embedding methods for a word intrusion task. pages 237–240.