Open semantic analysis: The case of word level semantics in Danish

Finn Årup Nielsen and Lars Kai Hansen

Cognitive Systems, DTU Compute
Technical University of Denmark

19 November 2017

A open Danish semantic model

Collect Danish corpora.

Setup a few models for semantic analysis.

Construct small evaluation datasets.

Evaluate the collected system.

Build a Python package *Dasem* available https://github.com/fnielsen/dasem/.

Make the system free for companies with the Apache license.

Open Danish corpora

Danish Wikipedia (CC BY-SA)

Danish Wikisource (at least CC BY-SA)

Danish part of *Gutenberg* (PD). Old books.

Danish part of *Runeberg* (PD). Old books.

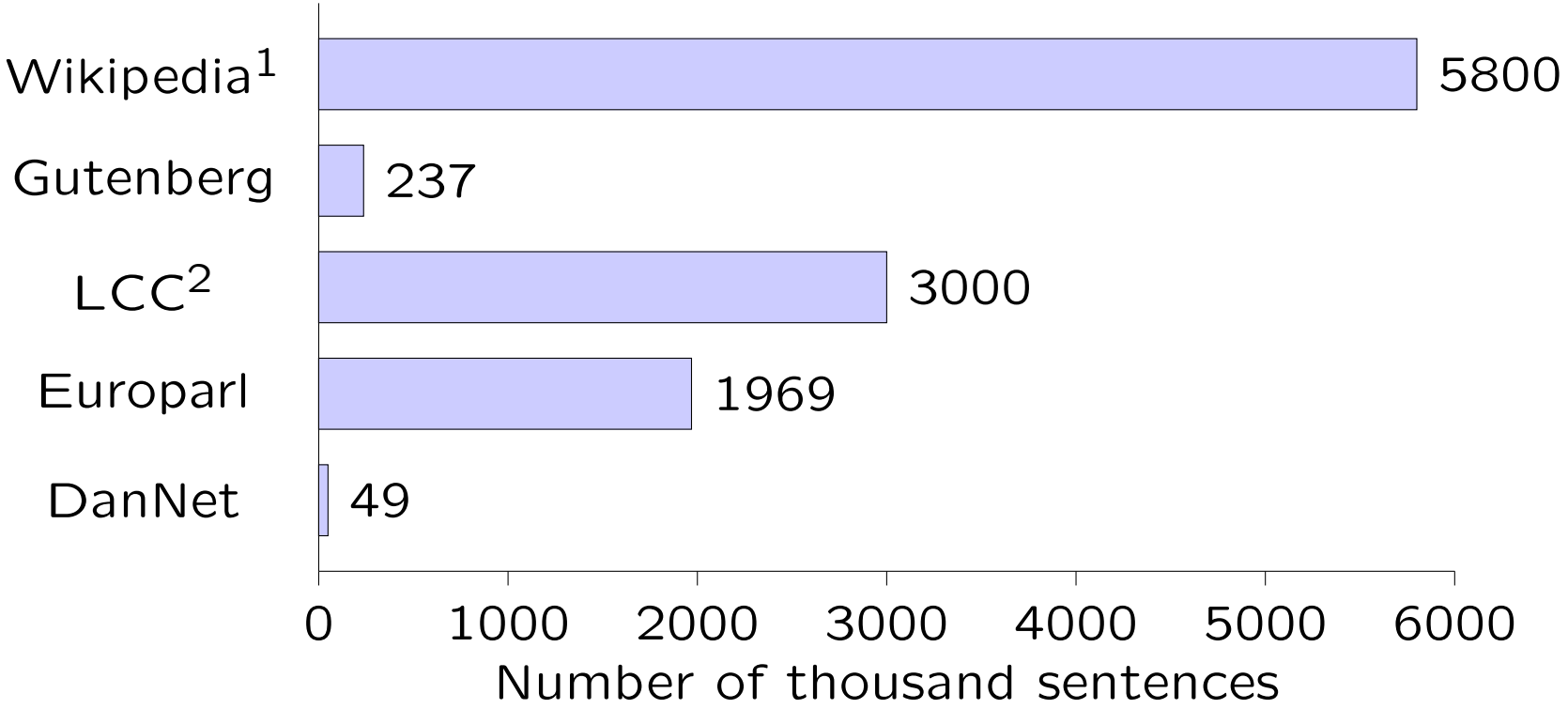
Danish part of *Leipzig Corpora Collection* (CC-BY). Various text from the Internet (Quasthoff et al., 2006).

Danish part of *Europarl* (PD). Parallel corpus from the EU Parliament (Koehn, 2005).

DanNet (DanNet license). Danish wordnet with example sentences (Pedersen et al., 2009).

Retsinformation.dk. Danish legal texts.

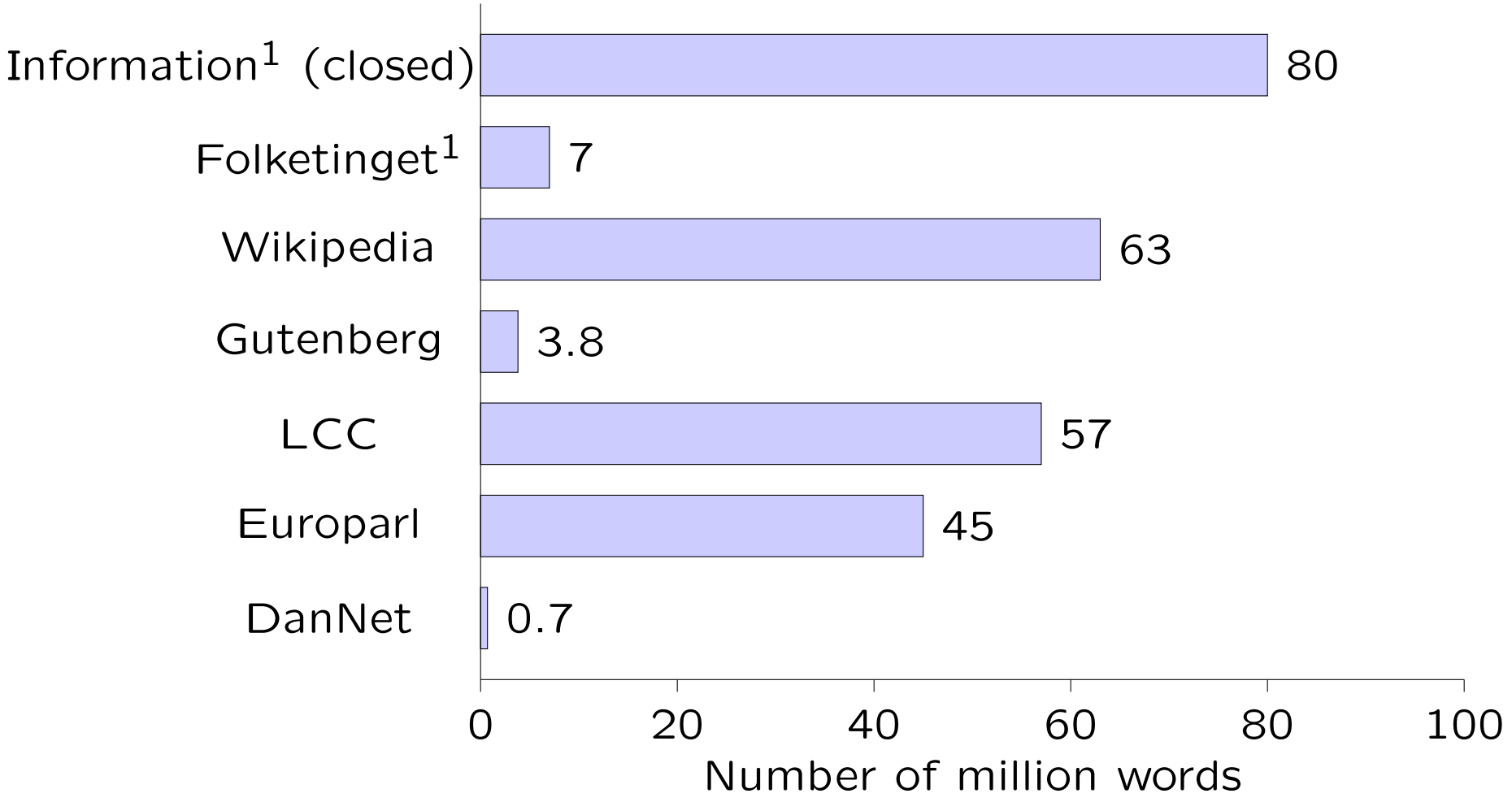
Open Danish corpora size wrt. sentences



[1]Wikipedia pages can be split into sentences in multiple ways.

[2]Only a part of the Danish part of LCC has been used so far.

Danish corpora size wrt. words



[1] According to https://visl.sdu.dk/corpus_linguistics.html

Two models

Explicit Semantic Analysis (ESA)

Word embedding with Word2vec in Gensim

Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) suggested back in 2007 (Gabrilovich and Markovitch, 2007).

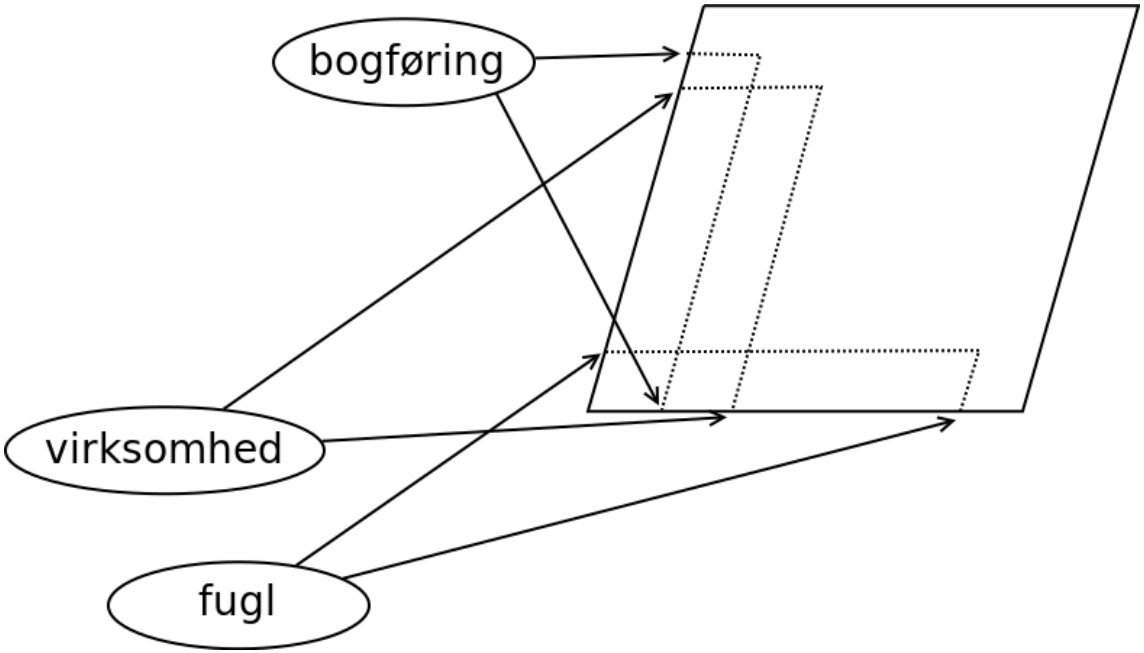
Represent each Wikipedia articles in a bag-of-words representation.

Tfidf-normalize the document-term matrix.

Representation of a term is the projection of the term so it becomes a weighting over Wikipedia articles: $\mathbf{y}_1 = \mathbf{X}\mathbf{q}_1$

Relatedness between two terms computer by distance function in projected space $d(\mathbf{y}_1, \mathbf{y}_2)$.

Word embedding



Word embedding: project words into a low dimensional subspace.

Word2vec: Predict word(s) from near word(s) with linear projection (Mikolov et al., 2013). Implemented in, e.g., Gensim (Řehůřek and Sojka, 2010)

Two types: Predict middle word from surrounding (CBOW), predict surrounding words (skipgram)

Semantically (and syntactically) similar words (should probably?) appear near each other in the projected space.

Three forms of evaluations

1. Semantic relatedness with a Danish version of **Wordsim353**

2. **Word intrusion** (Odd-one-out-of-four): Four terms where three of the terms share relatedness while the fourth term is the odd-one-out.

3. **AFINN** Word list for sentiment analysis. Predict the sign of the sentiment-labeled work with supervised learning based on word embedding as features.

Wordsim353-da

Danish translation of the classic English word list

Word 1	da1	Word 2	da2	Human (mean)	Problem
love	kærlighed	sex	sex	6.77	
tiger	tiger	cat	kat	7.35	
tiger	tiger	tiger	tiger	10	
book	bog	paper	papir	7.46	
computer	computer	keyboard	tastatur	7.62	
⋮					
football	fodbold	soccer	fodbold	9.03	1
⋮					

Only 319 word pairs used in the further analysis due to “problems”.

Compute similarity with the semantic models and compare with the human annotation.

Word intrusion

word1	word2	word3	word4
æble (apple)	pære (pear)	kirsebær (cherry)	stol (chair)
stol (chair)	bord (table)	reol (shelves)	græs (grass)
græs (grass)	træ (tree)	blomst (flower)	bil (car)
bil (car)	cykel (bike)	tog (train)	vind (wind)
vind (wind)	regn (rain)	solskin (sunshine)	mandag (Monday)

The first 5 rows of the odd-one-out-of-four dataset out of a total of 100 rows. The fourth column is the outlier. Distributed in *Dasem*: https://github.com/fnielsen/dasem/blob/master/dasem/data/four_words.csv

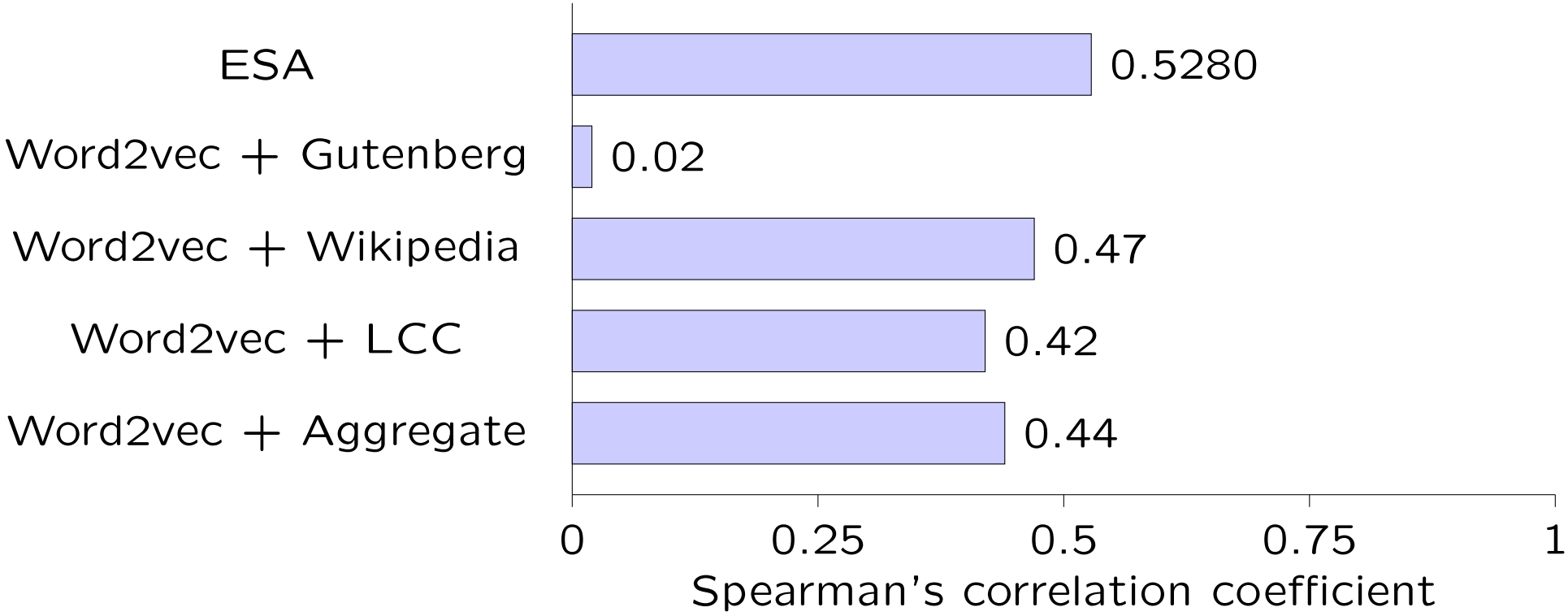
AFINN

AFINN word list with 3552 Danish words labeled with sentiment between −5 and +5 available at https://github.com/fnielsen/afinn/:

```
absorberet       1
acceptere        1
accepterede      1
...
flagskib         2
flerstrengede    2
flerstrenget     2
flop            -2
flot             3
flov            -2
fluekneppende   -3
flueknepperi    -3
```

Prediction of the sign of the sentiment label from AFINN word list.

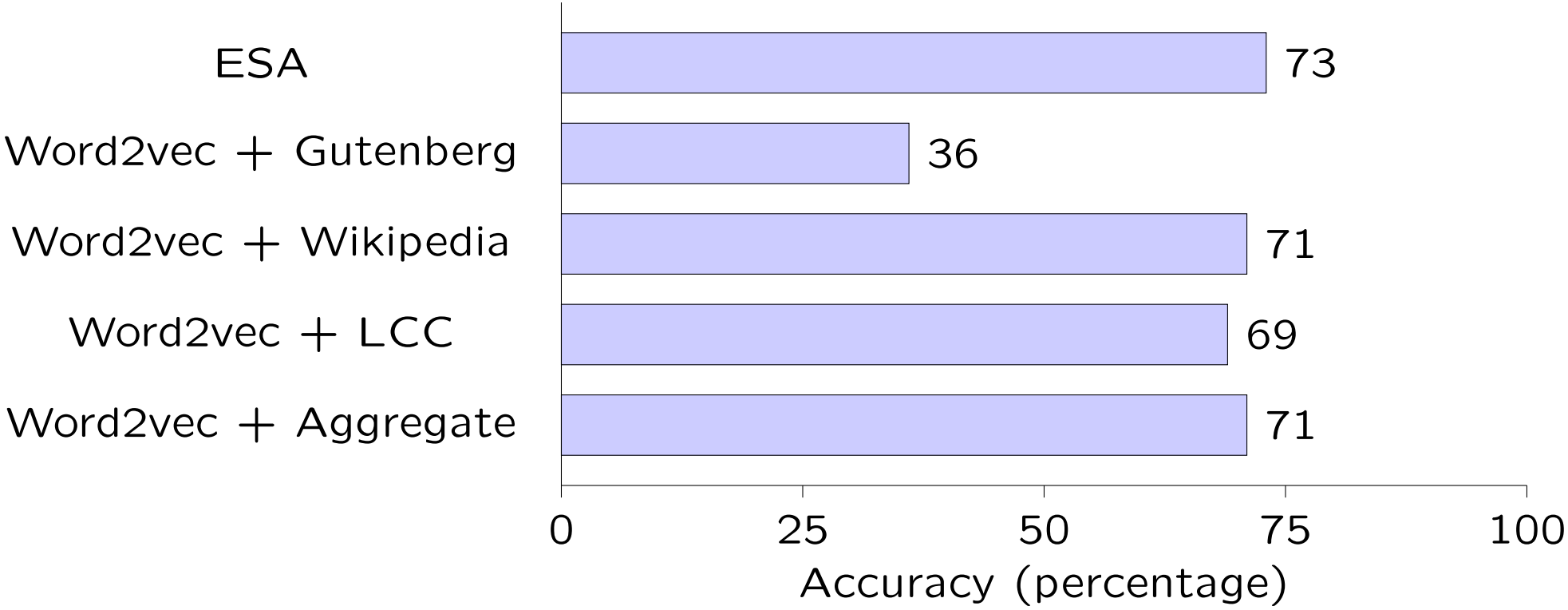
Wordsim353-da



Spearman's correlation coefficient between semantic model and human annotation on the wordsim353-da word pair data.

Bigger is better. ESA better than Word embedding.

Word intrusion



Accuracy in percentage for guessing the odd-one-out among four terms.

Bigger is better. ESA better than Word embedding.

Word intrusion

Detection of the odd-one-out with different semantic models.

word1	word2	word3	(outlier) word4	ESA	Wiki2vec Gutenberg	Wiki2vec LCC	Wiki2vec Wikipedia	Wiki2vec Aggregate
æble (apple)	pære (pear)	kirsebær (cherry)	stol (chair)	stol	stol	stol	stol	stol
stol (chair)	bord (table)	reol (shelves)	græs (grass)	græs	stol	bord	reol	bord
græs (grass)	træ (tree)	blomst (flower)	bil (car)	bil	træ	bil	bil	bil
bil (car)	cykel (bike)	tog (train)	vind (wind)	vind	tog	vind	tog	tog
vind (wind)	regn (rain)	solskin (sunshine)	mandag Monday	mandag	mandag	mandag	mandag	mandag

Five first rows in dataset: Here the ESA model detects all five correct, while the word2vec models selects the wrong term multiple times.

Predicting AFINN word sentiment

Accuracy for a number of classifiers trained to predict sign of AFINN sentiment score from the representation in the word embedding:

Classifier	Gutenberg	Wikipedia	LCC	Aggregate
MostFrequent	0.596 (0.019)	0.632 (0.027)	0.653 (0.006)	0.646 (0.013)
AdaBoost	0.644 (0.015)	0.754 (0.016)	0.806 (0.009)	0.829 (0.010)
DecisionTree	0.564 (0.018)	0.645 (0.019)	0.716 (0.011)	0.721 (0.020)
GaussianProcess	0.660 (0.020)	0.741 (0.022)	0.784 (0.014)	0.812 (0.011)
KNeighbors	0.615 (0.017)	0.711 (0.022)	0.765 (0.011)	0.796 (0.014)
Logistic	0.694 (0.015)	0.779 (0.016)	0.832 (0.011)	0.853 (0.009)
PassiveAggressive	0.624 (0.051)	0.723 (0.036)	0.792 (0.024)	0.830 (0.030)
RandomForest	0.622 (0.017)	0.722 (0.024)	0.774 (0.009)	0.791 (0.008)
RandomForest1000	0.672 (0.012)	0.777 (0.020)	0.825 (0.010)	0.860 (0.011)
SGD	0.653 (0.021)	0.758 (0.018)	0.808 (0.024)	0.836 (0.020)

Table 1: Classifier accuracy for sentiment prediction over *scikit-learn* classifiers with Project Gutenberg, Wikipedia, LCC and *aggregate* corpora Word2vec features. The *MostFrequent* classifier is a baseline predicting the most frequent class whatever the input might be. *SGD* is the stochastic gradient descent classifier. The values in the parentheses are the standard deviations of the accuracies of 10 training/test set splits.

Predicting AFINN word sentiment

Investigating wrong annotations.

"Ophidset" (excited or strongly irritated) and "udsigtsløs" (futile). Both labeled positively in AFINN. This should be corrected in the word list.

Implicit negativity: benådet (pardoned), tilgiver (forgives), præcisere (clarify), formilder (appeases), appellerer (appeals) and frikendt (acquitted). The AFINN word list has these terms as positive.

Schadenfreude or sarcasm like: "lol" and "hahaha"

Conclusion

We can obtain reasonable performance with semantic models on free Danish corpora.

Explicit Semantic Analysis works well for some tasks compared to Word2vec embedding.

Python package available at https://github.com/fnielsen/dasem

What next

More data? For instance, include non-free datasets?

fastText to handle Danish compounds.

Thanks

References

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.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *The Tenth Machine Translation Summit: Proceedings of Conference*, pages 79–86.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299. DOI: 10.1007/S10579-009-9092-1.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. *LREC 2006 Proceedings*, pages 1799–1802.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. *New Challenges For NLP Frameworks Programme*, pages 45–50.