Semantics and sentiment in a small language

Finn Årup Nielsen

DABAI, DTU Compute
Technical University of Denmark

29 March 2017

The problem with bogføringsvirksomhed

“bogføringsvirksomhed”

The problem with bogføringsvirksomhed

Look up in big dictionary

Den Danske Ordbog http://ordnet.dk/ddo/ does not have the word "bogføringsvirksomhed": Should have been between bogføringspligt og boghandel

Neither in *Ordbog over det dansk Sprog* http://ordnet.dk/ods/.

Nor *KorpusDK* http://ordnet.dk/korpusdk/.

Decompounding

Decompounding: split word into its parts:

bogføringsvirksomhed $\rightarrow$ bogføring|s|virksomhed

Decompounding

Decompounding: split word into its parts:

bogføringsvirksomhed $\rightarrow$ bogføring|s|virksomhed

. . . and "bogføring" and "virksomhed" are to be found in dictionaries.

Decompounding

Decompounding: split word into its parts:

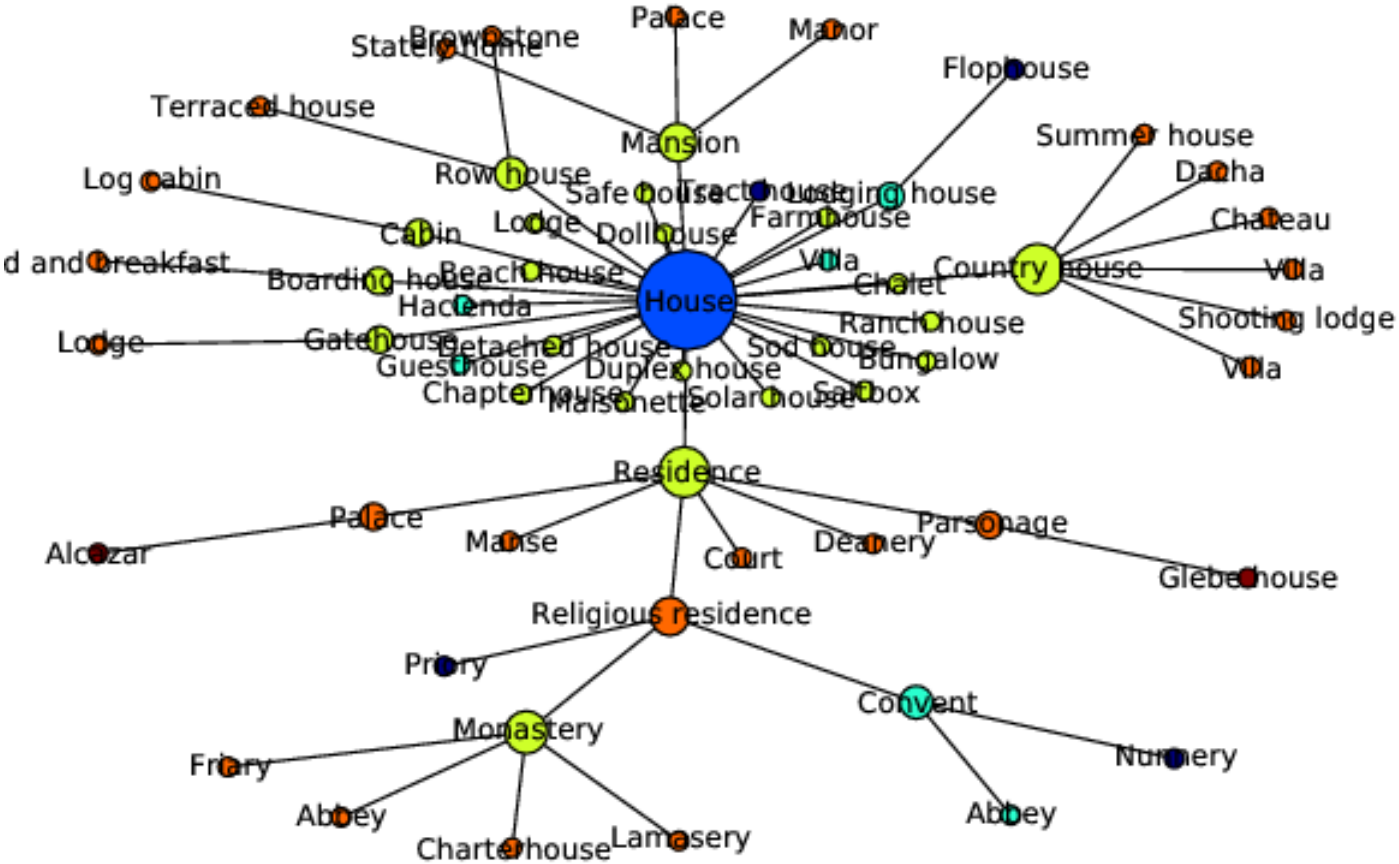
bogføringsvirksomhed $\rightarrow$ bogføring|s|virksomhed

. . . and "bogføring" and "virksomhed" are to be found in dictionaries.

. . . but what do "bogføring" and "virksomhed" mean?

DanNet

DanNet: a Danish wordnet (Pedersen et al., 2009).



Inspired from (Bird et al., 2009, section 4.8, pages 169+)

DanNet's understanding of "virksomhed"

organisation (som producerer og) sælger varer el. ... (Brug: "hun har netop startet sit eget firma, der designer børnetøj"; "Danmarks mest internationale virksomhed, ØK, fik i første halvår af 1991 et regnskabsresultat på 234 millioner kroner før skat")

beskæftigelse el. arbejde med noget (Brug: "det er en kendsgerning, at ordbogsarbejde og udgiverarbejde er mindre påagtet end anden videnskabelig virksomhed")

organisation (som producerer og) sælger varer el. ... (Brug: "Danmarks mest internationale virksomhed, ØK, fik i første halvår af 1991 et regnskabsresultat på 234 millioner kroner før skat")

DanNet's understanding of "bogføring"

DanNet's understanding of "bogføring"

"bog" "føring" ?

DanNet's understanding of "bog"

lille, trekantet frugt fra træer af slægten bøg (Brug: "Bog er nødfrugter, der bag en hård skal indeslutter et enkelt frø med tynd frøskal")

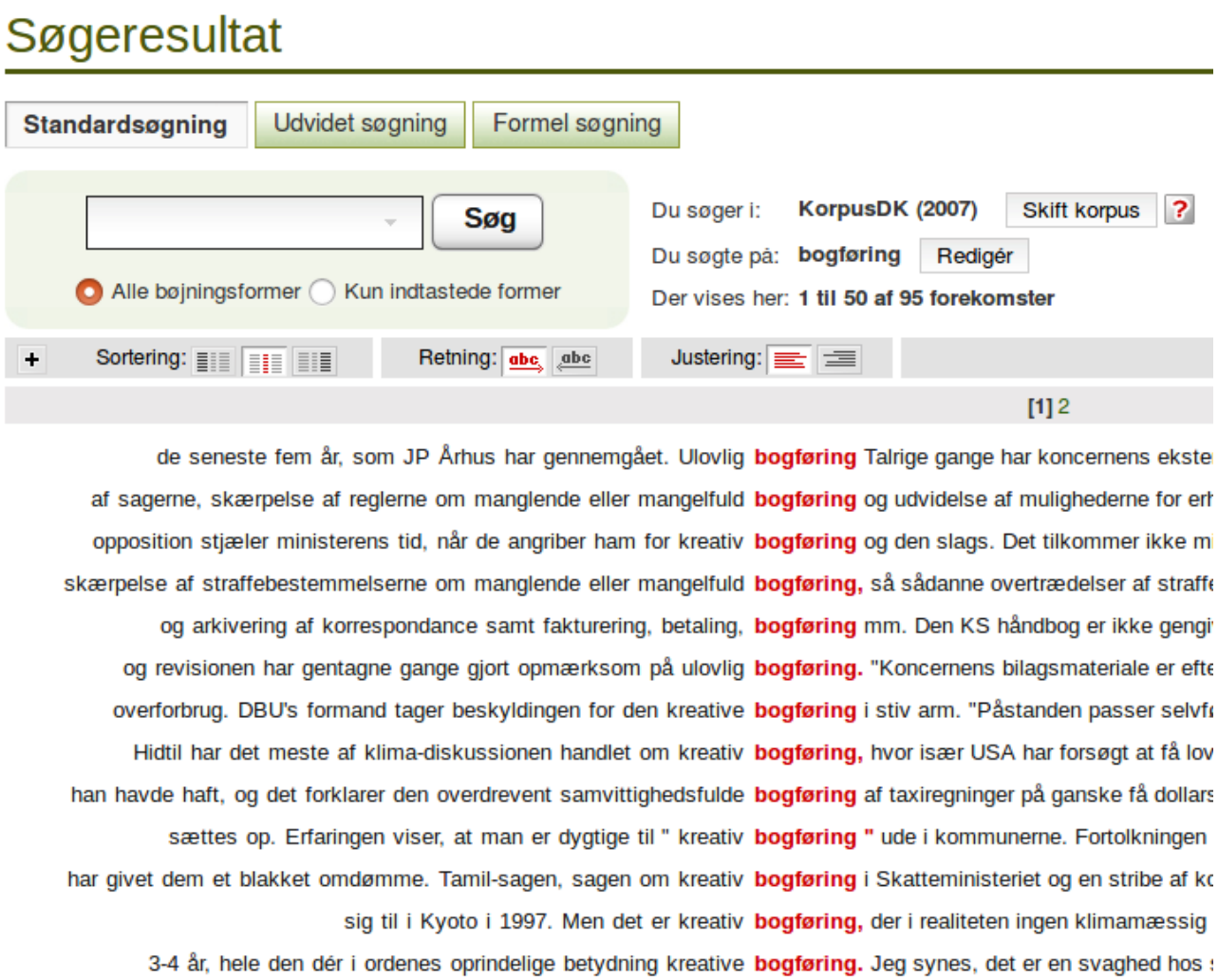
indbundne el. sammenhæftede blade beregnet til opt ... (Brug: "En logbog er ingen dagbog. Det er en bog, hvori føreren af et luftfartøj fører regnskab over sin flyvetid")

trykte el. beskrevne blade af papir indbundet el. ... (Brug: "Den indbundne og illustrerede bog er på 416 sider og koster 398 kr. || Så skal jeg på biblioteket og aflevere og genlåne bøger"; ...

større del af et værk, fx af en roman (Brug: "Jeg åbnede moppedrengen: 'Første bog Kapitel 1 HJEMKOMSTEN Den 24. februar 1815 blev det meldt fra søvagttårnet i Marseille at ..'")

del af Bibelen (Brug: "Dette uendeligt smukke citat stammer fra den sidste bog i Bibelen, Johannes Åbenbaring")

Corpora



KorpusDK has 95 occurrences of “bogføring”.

Google says “Ca. 510.000 resultater” for a query on “bogføring”.

So there are some context to get to “understand” “bogføring”

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.

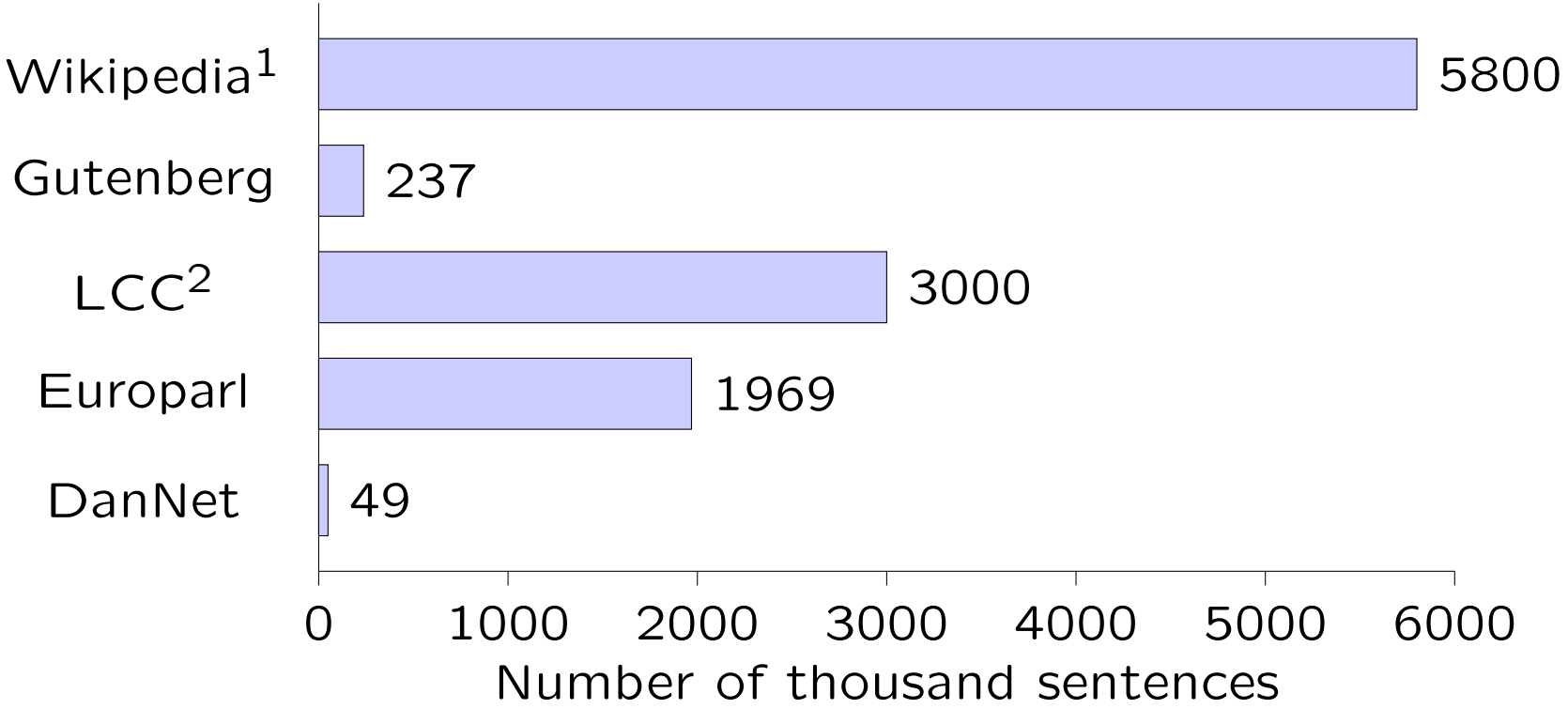
Danish part of *Leipzig Corpora Collection* (CC-BY). Various text from the Internet.

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

DanNet (DanNet license). Example sentences.

Retsinformation.

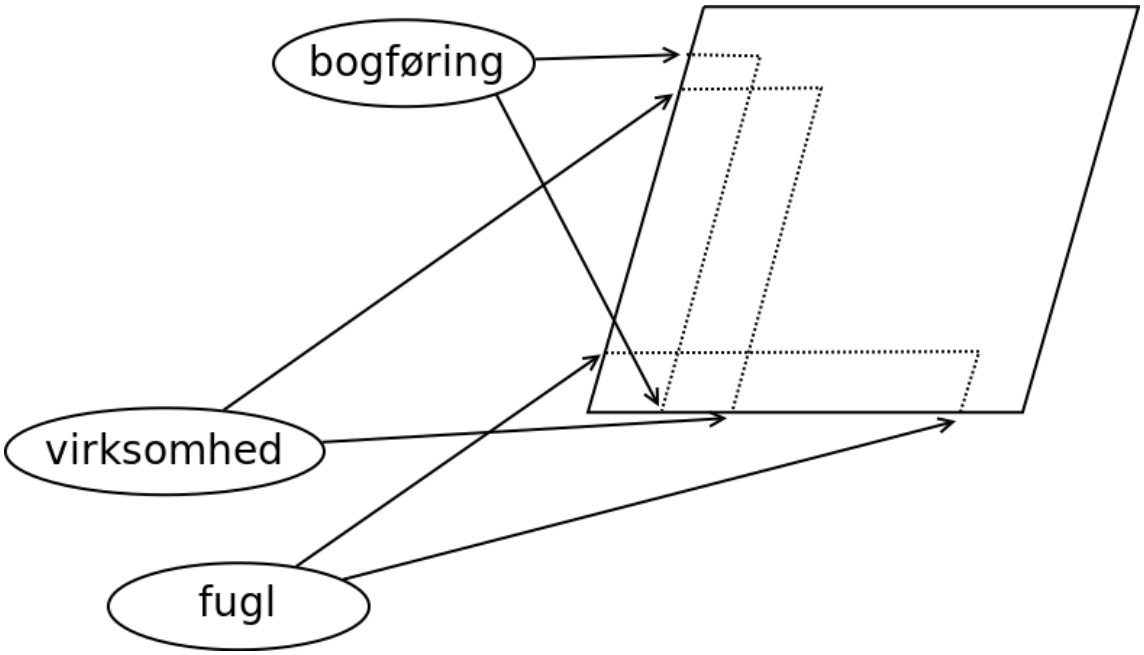
Open Danish corpora size



[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.

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.

Trained word embedding

Gensim-based CBOW word2vec embedding trained on an aggregate of LCC + Europarl + Dannet + Gutenberg corpora implemented in *Dasem*, — a Python package for Danish semantic analysis.

```
$ python -m dasem most-similar bogføring
budgetlægning
regnskabsføring
økonomistyring
programmering
finanskontrol
administration
budgetforvaltning
finansforvaltning
bogholderi
marketing
```

Supervised learning

But can we nevertheless use the word embedding to predict labels with supervised learning?

Supervised learning

But can we nevertheless use the word embedding to predict labels with supervised learning?

Our first attempt was predicting sentiment label from AFINN word list:

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

Supervised learning

Accuracy for a number of classifiers trained to predict sign of AFINN sentiment score from their 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.

Bogføringsvirksomhed is still a problem

“bogføringsvirksomhed” is still a problem because it does not exist in our corpus and thus cannot be projected into the word embedding.

Bogføringsvirksomhed is still a problem

"bogføringsvirksomhed" is still a problem because it does not exist in our corpus and thus cannot be projected into the word embedding.

Build a decompounder for splitting "bogføringsvirksomhed" into "bogføring" and "virksomhed"?

Bogføringsvirksomhed is still a problem

"bogføringsvirksomhed" is still a problem because it does not exist in our corpus and thus cannot be projected into the word embedding.

Build a decompounder for splitting "bogføringsvirksomhed" into "bogføring" and "virksomhed"?

Or represent the word with character n-grams, e.g., 4-grams example: bogf ogfø gfør føri ørin ring ings ngsv gsvi svir virk irks rkso ksom somh omhe mhed. And train an n-gram embedding?

Bogføringsvirksomhed is still a problem

"bogføringsvirksomhed" is still a problem because it does not exist in our corpus and thus cannot be projected into the word embedding.

Build a decompounder for splitting "bogføringsvirksomhed" into "bogføring" and "virksomhed"?

Or represent the word with character n-grams, e.g., 4-grams example: bogf ogfø gfør føri ørin ring ings ngsv gsvi svir virk irks rkso ksom somh omhe mhed. And train an n-gram embedding?

Or just use fastText (Joulin et al., 2016; Bojanowski et al., 2016)?

fastText: words and n-grams embedding

Fasttext + aggregate corpus:

```
python -m dasem.fullmonty fasttext-most-similar bogføring
bogføring
bogføring,
bogføringen
Bogføring
regnskabsføring
bogføring.
regnskabsføring,
bogføre
bogføringen,
regnskabsføring.
```

fastText

With "-en" postfixed:

```
python -m dasem.fullmonty fasttext-most-similar bogføringen
```

fastText

With "-en" postfixed:

```
python -m dasem.fullmonty fasttext-most-similar bogføringen
bogføringen
bogføring
bogføringen,
bogføring,
regnskabsføringen
regnskabsføring
regnskabsføring,
regnskabsførelsen
faktureringen
bogførte
```

Still ok!

fastText

Even spelling error are handle:

```
python -m dasem.fullmonty fasttext-most-similar bogføringn
bogføring
bogføring,
bogføringen
regnskabsføring
bogfører
bogføring.
regnskabsføring,
bogføre
Bogføring
bogført
```

fastText with bogføringsvirksomhed

"bogføringsvirksomhed" is now possible to project to the embedding.

```
investeringsvirksomhed
forretningsvirksomhed
rådgivningsvirksomhed
næringsvirksomhed
lovgivningsvirksomhed
børsvirksomhed
forsikringsvirksomhed
oplysningsvirksomhed
anlægsvirksomhed
forædlingsvirksomhed
```

fastText with bogføringsvirksomhed

"bogføringsvirksomhed" is now possible to project to the embedding:

```
investeringsvirksomhed
forretningsvirksomhed
rådgivningsvirksomhed
næringsvirksomhed
lovgivningsvirksomhed
børsvirksomhed
forsikringsvirksomhed
oplysningsvirksomhed
anlægsvirksomhed
forædlingsvirksomhed
```

Better, but not fully working yet.

Summary

We can build open embeddings for Danish semantics.

The embeddings can act as features in a supervised learning setting.

Thanks

References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly, Sebastopol, California. ISBN 9780596516499. *The canonical book for the NLTK package for natural language processing in the Python programming language. Corpora, part-of-speech tagging and machine learning classification are among the topics covered.*

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

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification.

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.

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