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

Finn Årup Nielsen, Lars Kai Hansen

Cognitive Systems, DTU Compute
Technical University of Denmark
Kongens Lyngby, Denmark
faan,lkai@dtu.dk

Abstract

The present research is motivated by the need for accessible and efficient tools for automated semantic analysis in Danish. We are interested in tools that are completely open, so they can be used by a critical public, in public administration, non-governmental organizations and businesses. We describe data-driven models for Danish semantic relatedness, word intrusion and sentiment prediction. Open Danish corpora were assembled and unsupervised learning implemented for explicit semantic analysis and with *Gensim*'s Word2vec model. We evaluate the performance of the two models on three different annotated word datasets. We test the semantic representations' alignment with single word sentiment using supervised learning. We find that logistic regression and large random forests perform well with Word2vec features.

1. Introduction

Text mining for a language of the size of Danish, may be hindered by the lack of language resources (Derczynski et al., 2014; Pedersen et al., 2012). The state of the art and challenges related to tools for analysis of such languages in a European context have been described in the publications of META-NET[1]. In the Danish META-NET white paper, a 'range of experts' have rated existing language technology support for the Danish language using a scale from 6 (best) to 0 (worst). Among ten different fields *semantics* scores low on all dimensions, e.g, a score of 0 on availability. This represents a serious gap as typical applications of text mining relate one way or another to extraction of meaning. Here we will make an effort to mitigate this gap. In particular, we are interested in completely open tools, so they can be widely applied in society, including a critical public, public administration, non-governmental organizations and for data-driven innovation in business.

There exists a number of Danish corpora (Kirchmeier-Andersen, 2002; Norling-Christensen and Asmussen, 1998), some of which are annotated (Pedersen et al., 2014). These corpora are to varying degrees available or open, e.g., the recent Danish semantics resource SemDaX provides sense tagging (Pedersen et al., 2016), available 'through a CLARIN academic license'.[2] Our approach is based on aggregation of multiple open Danish corpora and unsupervised semantic word models constructed with open source tools, catering for societal use and business innovation.

Wikipedia is a key corpus and has been applied in numerous text mining applications (Medelyan et al., 2009; Mehdi et al., 2017), e.g., for inferring semantic relatedness of word pairs (Sajadi et al., 2015; Strube and Ponzetto, 2006; Gabrilovich and Markovitch, 2007). One particularly successful application of the Wikipedia corpus is *explicit semantic analysis* (ESA) in which classical information retrieval vector space models yielded state-of-the-art performance in a semantic relatedness task (Gabrilovich and Markovitch, 2007). More recent progress in word embedding models trained with large corpora has been reported to reflect semantics in several dimensions (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016).

2. Open data sources in Danish

We used five Danish corpora that were publicly available and free for general use:

Danish Wikipedia. We downloaded the Danish Wikipedia XML article dump from https://dumps.wikimedia.org/ and used the *mwparserfromhell* Python module to extract text from 351,186 raw article wiki-pages. For splitting the text into sentences we used the default Danish sentence tokenizer in *NLTK* (Bird et al., 2009) getting a total of 3,421,052 sentences.

Leipzig corpora collection. The Leipzig corpora collection (LCC) distributes several corpora in various languages (Quasthoff et al., 2006).[3] We downloaded a subset of the Danish corpora which included three of the largest corpora. Our combined LCC corpus had a total of 3,009,997 sentences.

Project Gutenberg. We downloaded all Danish texts from the digital text library of the Project Gutenberg.[4] Totally, we found 63 ebooks. Before the Danish 'spelling reform' double-a ('aa') was used instead of the modern "å", and common nouns had capital first letter. Our text processing translated the double-a to a "å", while no action was taken to handle noun's capital first letter as subsequent text mining was case insensitive. With *NLTK* preprocessing this dataset comprised 236,824 sentences.

DanNet is a Danish wordnet (Pedersen et al., 2009).[5] It has usage example sentences for many *synsets*. A total of 49,040 sentences were extracted.

[1] META-NET is a Network of Excellence dedicated to fostering the technological foundations of a multilingual European information society. See http://www.meta-net.eu/ for additional details and link to publications about other European Languages.

[2] https://www.clarin.eu/content/license-categories

[3] http://corpora.uni-leipzig.de/

[4] http://www.gutenberg.org

[5] http://www.wordnet.dk/

Europarl is a multilingual corpus (Koehn, 2005).[6] We extracted the 1,968,800 Danish sentences from the Danish–English parallel corpus.

2.1. Evaluation metrics and data

We use three labeled datasets for evaluation:

wordsim353-da. We translated the English *wordsim353* dataset, originally described as a "a diverse list of 350 noun pairs representing various degrees of similarity" (Finkelstein et al., 2002). The dataset obtained[7] had 353 noun pairs with manually assigned similarity scores. We kept the English similarity scores and translated the noun pairs to Danish. This process weakened the test as the semantics of the English words does not necessarily correspond to the semantics of the translations into Danish. We indicated noun pairs with a major semantic problem with an extra column in the dataset. As an example consider our translation of the pair "soccer" and "football": We translated both words to "fodbold" (the American version of football would in Danish be referred to as "amerikansk fodbold"). In the further analysis, we only used the 319 word pairs that were not indicated to have a major semantic problem. We leave it to future research to further assess the quality of the translation by use of multiple translators as in (Hassan and Mihalcea, 2009).

Word intrusion. We constructed a new odd-one-out-of-four dataset containing sets of four words/phrases where the last word is an outlier compared to the three others, e.g., ("æble", "pære", "kirsebær", "stol") corresponding in English to (apple, pear, cherry, chair), see Table 1. The task is to predict the outlier word, — a task related to the *word intrusion* task used in topic model evaluation (Chang et al., 2009). The dataset consists of 100 individual tests.

AFINN. This is an open word list with sentiment-labeled words. The sentiment scores range from -5 (most negative) to $+5$ (most positive. Originally established for English (Nielsen, 2011), a derived Danish word list is openly available. We used AFINN-da-32.txt.[8] The list holds sentiment scores for 3,552 Danish words and phrases.

3. Methods

We invoked two semantic models:

Word2vec (Mikolov et al., 2013) word embedding model as implemented in *Gensim* (Řehůřek and Sojka, 2010). We iterated over sentences, splitting each with *NLTK*'s *WordPunctTokenizer* to get a list of tokens (words and punctuation). After conversion of words to lowercase we fed the lists of tokens to the *Gensim* Word2vec training it with its default parameters, where the dimension of the word embedding space is 100. For the semantic relatedness we use the default *Gensim* similarity computed as the dot product between normalized word embedding vectors. We built several Word2vec models: Separate models for some of the individual corpora and one *aggregate* model, aggregating multiple corpora (Gutenberg, LCC, DanNet and Europarl) getting a total of 5,264,661 sentences and 124 million tokens.

ESA (Gabrilovich and Markovitch, 2007). We implemented this scheme using the *tfidf* vectorizer in *scikit-learn* (Pedregosa et al., 2011) and used the Wikipedia corpus for this model as in the original work. We fed the raw wikitext directly into the vectorizer and used its default word tokenizer. Though ESA has been based on other corpora than Wikipedia (Anderka and Stein, 2009), we only evaluate the ESA model within the original setting, i.e., using Wikipedia only.

To evaluate the semantic models for sentiment polarity prediction we use supervised machine learning, i.e., train classifiers based on the semantic representations for classifying the polarity (sign) of the *AFINN* sentiment score. We evaluated the performance on an independent test set, randomizing a 75%/25% training/test set split 10 times. This test is based on the assumption that semantically related words are also related by sentiment. The assumption will be discussed later. We used a range of machine learning classifiers from the *scikit-learn* package. We generally refrained from extensive optimization of the classifiers and used default hyper-parameters, except for the random forest classifier were we also included a large forest based on 1,000 estimators, exploring the contested claim that random forests *"is clearly the best family of classifiers"* (Fernández-Delgado et al., 2014; Wainberg et al., 2016). The number of estimators in the random forest classifier is a hyper-parameter that may increase performance at the expense of training time.

To probe the structure of the classification problem we noted that a random forest and a logistic regression classifier reached training set accuracies of 99.4% and 84%, respectively. As expected the random forest has sufficient flexibility to model the data (and possibly overfit), while the logistic regression's *linear decision surface* regularizes the fit.

Corpus processing and word embeddings methods are implemented in *Dasem*.[9] The Danish word list is distributed as part of this library.

4. Results

For the *wordsim353-da* task, the Spearman correlation between the human annotated semantic relatedness and the estimated relatedness shows the best performance for explicit semantic analysis reaching $\rho_{\mathrm{ESA}} = 0.52$. For the Word2vec models the Spearman correlations were $\rho_{\mathrm{Gutenberg=0.02}}$, $\rho_{\mathrm{Wikipedia}} = 0.47$, $\rho_{\mathrm{LCC}} = 0.42$, $\rho_{\mathrm{aggregate}} = 0.44$, respectively.

On our *word intrusion* task, the ESA method yielded an accuracy of 73%, while the Word2vec methods yield accuracies on 36%, 69%, 71% and 71% for the Gutenberg, Leipzig corpora collection, Wikipedia and *aggregate* corpora, respectively. Table 1 shows the first part of the dataset and the predicted outlier words for the five different combinations of corpora and methods we examined.

Table 2 displays the result for the *AFINN* sentiment prediction task across four different feature sets (from four

[6] http://www.statmt.org/europarl/

[7] Source: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[8] https://github.com/fnielsen/afinn/

[9] https://github.com/fnielsen/dasem/.

word1	word2	word3	(outlier) word4	ESA	W2V Gutenberg	W2V LCC	W2V Wikipedia	W2V 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

Table 1: The first 5 rows of the data and results for the *word intrusion* task. In each row, the first three columns contain three related words, while the fourth column is the outlier word. The English translation is show in parenthesis below the Danish word.

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 2: 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 *scikit-learn*'s stochastic gradient descent classifier which defaults to a linear support vector machine. The values in the parentheses are the standard deviations of the accuracies of 10 training/test set splits.

different corpora) and ten different *scikit-learn* classifiers, where one of the classifiers, *MostFrequent*, is a baseline. In tests, the logistic regression and the large random forest perform at par and best for the largest corpus *aggregate*.

5. Discussion

On the *wordsim353-da* semantic relatedness task, the small and old Gutenberg dataset performs very poorly, while the larger corpora somewhat better. Our best performance at 0.52 is still considerable lower than 0.76 reported for the GloVe model for the English wordsim353 trained on the several hundred times larger dataset (Pennington et al., 2014), but better than the correlations reported for the semantic relatedness task with non-English languages (Hassan and Mihalcea, 2009). While we note that there is a certain amount of noise injected in the translation process we also assume that larger data sets would improve the Danish relatedness results.

For the *word intrusion* task, we have found our best accuracies to be in the lower 70%s. These values may be compared to the results on the English *word intrusion* tasks (on entirely different corpora) reported to be in the ranges 0.65–0.8 and 0.70–0.82 (Chang et al., 2009, Figure 5).

We found accuracies in the 80%s for the larger corpora and best classifiers in the *AFINN* sentiment prediction task. An earlier study reported 77% accuracy for predicting the sentiment polarity in an English sentiment lexicon (Qin et al., 2014) essentially using a K-nearest neighbor scheme averaging the polarity of the top $K = 100$ closest words in a word embedding. A study on predicting word category based on word embedding features and using logistic regression for the supervised binary classification reached between 67% and 75% in accuracy on a 'positive emotion or not' task with the relatively small English PERMA word list (Dhillon et al., 2015).

We examined particular misclassified *AFINN* words, and found cases in which the annotation may be questioned. For instance, 'ophidset', labeled as positive, may be translated to 'excited', but also 'strongly irritated' and a cursory glance on Twitter shows that it is mostly used in the negative sense. The Danish word 'udsigtsløs' is another positive-labeled word. The scoring of the corresponding English word 'futile' in the English version of *AFINN* has previously been noted as disagreeing with other sentiment lexicons (Bravo-Marquez et al., 2014). Some of the positive words for which the classifiers dis-

agree with the *AFINN* score involve 'implicit' negativity: benådet (pardoned), tilgiver (forgives), præcisere (clarify), formilder (appeases), appellerer (appeals) and frikendt (acquitted). Yet other cases reflect classical challenges in single word based sentiment analysis, e.g., containing elements of schadenfreude or sarcasm like: 'lol' and 'hahaha'. Conversely the semantic classifiers can be seen as tool for understanding and cleaning the sentiment annotated word list. For some of the cases we found that words with opposite sentiment polarity are mapped by neighboring Word2vec vectors, — this has been noted also by (Qin et al., 2014). For instance, god (good) and dårlig (bad) are mapped closely, and among the 10 most similar words for 'accepteret' (accepted) approximately half have negative sentiment and the other half positive.

We have presented three tasks and applied two common semantic analysis methods on five corpora. Overall, we conclude that the semantic tasks benefit from large corpora, — a not surprising observation. For sentiment prediction, we find that a large aggregated corpus performs the best and that a supervised classifier, predicting from a word embedding feature space, can point to words that may be suboptimally labeled.

Acknowledgments

This work was support by the Danish Innovation Foundation (Innovationsfonden) through the project Danish Center for Big Data Analytics and Innovation (DABAI).

6. References

Anderka, Maik and Benno Stein, 2009. The ESA retrieval model revisited. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*:670–671.

Bird, Steven, Ewan Klein, and Edward Loper, 2009. Natural Language Processing with Python.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2016. Enriching Word Vectors with Subword Information.

Bravo-Marquez, Felipe, Marcelo Mendoza, and Barbara Poblete, 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86–99.

Chang, Jonathan, 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*:288–296.

Derczynski, Leon, Camilla Vilhelmsen Field, and Kenneth S. Bøgh, 2014. DKIE: Open Source Information Extraction for Danish. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*:61–64.

Dhillon, Paramveer S., Dean P. Foster, and Lyle H. Ungar, 2015. Eigenwords: Spectral Word Embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim, 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Gabrilovich, Evgeniy and Shaul Markovitch, 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on Artifical intelligence*:1606–1611.

Hassan, Samer and Rada Mihalcea, 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*:1192–1201.

Kirchmeier-Andersen, Sabine, 2002. Dansk korpusbaseret forskning. Hvordan kommer vi videre? *Studies in Modern Danish*:11–26.

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

Medelyan, Olena, David Milne, Catherine Legg, and Ian H. Witten, 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67:716–754.

Mehdi, Mohamad, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki, 2017. Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. *Information Processing & Management*, 53:505–529.

Mikolov, Tomas, Jeff Dean, and Greg Corrado, 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26*:3111–3119.

Nielsen, Finn Årup, 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*:93–98.

Norling-Christensen, Ole and Jørg Asmussen, 1998. The Corpus of the Danish Dictionary. *Lexikos*, 8:223–242.

Pedersen, Bolette Sandford, 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.

Pedersen, Bolette Sandford, Sanni Nimb, Sussi Olsen, Anders Søgaard, Anders Trærup Johannsen, Anna Braasch, Hector Martinez Alonso, and Nicolai Hartvig Sørensen, 2016. The SemDaX Corpus - sense annotations with scalable sense inventories. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*:842–847.

Pedersen, Bolette Sandford, Sanni Nimb, Sussi Olsen, Anders Søgaard, and Nicolai Sørensen, 2014. Semantic annotation of the Danish CLARIN Reference Corpus. *Proceedings 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*:25–28.

Pedersen, Bolette Sandford, Jürgen Wedekind, Steen

Bøhm-Andersen, Peter J. Henrichsen, Sanne Hoffensetz-Andresen, Sabine Kirchmeier-Andersen, Jens Otto Kjærum, Louise Bie Larsen, Bente Maegaard, Jens-Erik Rasmussen, Peter Revsbech, Hanne E. Thomsen, and Sanni Nimb, 2012. Det danske sprog i den digitale tidsalder.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher Manning, 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*:1532–1543.

Qin, Bing, Duyu Tang, Furu Wei, Ming Zhou, Nan Yang, and Ting Liu, 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*:1555–1565.

Quasthoff, Uwe, Matthias Richter, and Christian Biemann, 2006. Corpus Portal for Search in Monolingual Corpora. *LREC 2006 Proceedings*:1799–1802.

Sajadi, Armin, Evangelos E. Milios, Vlado Kešelj, and Jeannette C. M. Janssen, 2015. Domain-Specific Semantic Relatedness from Wikipedia Structure: A Case Study in Biomedical Text. *Computational Linguistics and Intelligent Text Processing*:347–360.

Strube, Michael and Simone Paolo Ponzetto, 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*:1419–1424.

Wainberg, Michael, Babak Alipanahi, and Brendan J. Frey, 2016. Are Random Forests Truly the Best Classifiers? *Journal of Machine Learning Research*, 17:1–5.

Řehůřek, Radim and Petr Sojka, 2010. Software framework for topic modelling with large corpora. *New Challenges For NLP Frameworks Programme*:45–50.