Topic Coherence for Dutch

Janneke M. van der Zwaan
Netherlands eScience Center
j.vanderzwaan@esciencecenter.nl

Maarten Marx
University of Amsterdam
maartenmarx@uva.nl

Jaap Kamps
University of Amsterdam
j.kamps@uva.nl

Topic modeling has become a popular technique for exploring and analyzing the contents of large text corpora. The method finds latent topics by grouping together words that co-occur in documents. Topic modeling is unsupervised and probabilistic, and as a result of this topics are not always semantically coherent or meaningful. Because creating manual usefulness ratings is costly, researchers have been trying to come up with ways to automatically assess the interpretability of topics. A potential solution to this problem are topic coherence measures, which are calculated by taking into account word co-occurrence statistics of an external corpus, such as Wikipedia. A recent study proposes a framework for topic coherence measures, and systematically evaluates the correlation between manual usefulness ratings and a wide variety of topic coherence measures (Röder et al. 2015). However, all calculations are done for English topics, using the English Wikipedia. Having a resource to calculate topic coherence for Dutch could benefit researchers in digital humanities. In this paper, we describe how we created such a resource from the Dutch Wikipedia, using Palmetto, a tool provided by Röder et al. (2015). In addition, we present the results of a case study to determine the best topic coherence measure for Dutch. The Palmetto database for Dutch was generated from a Wikipedia dump containing articles[1]. Preprocessing of the articles consisted of lemmatization and stopword removal. Palmetto and the database based on the Dutch Wikipedia are available online[2].

In order to determine what topic coherence measure works best for Dutch, we conducted a case study. Topics were learned from the proceedings of the Dutch house of parliament and senate from parliamentary years 1999/2000–2011/2012[3]. Because in our project we were interested in a special form of topic modeling, cross-perspective topic modeling (Fang et al. 2012), topics were learned from the nouns in the corpus only. We extracted 100 topics from 20594 documents. Manual useful ratings were gathered for the top 10 topic words of these topics. Three independent judges were asked to rate

[1] Wikipedia dump from September 2, 2015: https://dumps.wikimedia.org/nlwiki/20151102/
[2] Palmetto: http://aksw.org/Projects/Palmetto.html, Palmetto position storing Lucene index of Dutch Wikipedia: https://doi.org/10.5281/zenodo.46377
[3] http://ode.politicalmashup.nl/data/summarise/folia/

topics on a 3-point Likert scale. A score of 1 indicates a 'Useless' topic (i.e., words appear to be random and unrelated to each other), 2 indicates 'Average quality' (i.e., some of the topic words are coherent and interpretable but others are not), and 3 indicates a 'Useful' topic (i.e., one that is semantically coherent, meaningful and interpretable) (Aletras & Stevenson 2013). Table 1 presents example topics for each of these categories (topics were rated with the indicated score by all three judges).

Usefulness rating	Topic	Words
1 ('Useless')	6	border, abroad, traffic, principle, line, difference, idea, past, construction, circumstance
2 ('Average quality')	61	book, schoolbook, price, method, human trafficking, prostitution, book price, registration, prostitute, supply
3 ('Useful')	96	student, education, institution, university, study, quality, stipend, higher professional education, tuition, college

Table 1: Example topics.

Inter-rater reliability was calculated using Krippendorff's alpha (Krippendorff 1980). Krippendorff's alpha for the topic ratings is 0:72.

	C_A	C_P	C_V	*NPMI*	*UCI*	*UMass*
Correlation	0.092	0.368*	0.129	0.416*	0.364*	0.076

Table 2: Pearson correlation coefficients between the usefulness score and topic coherence measures (* statistically significant at $p < 0.05$).

To determine what topic coherence measure works best, we calculated Pearson correlation between the mean of the usefulness ratings and the different topic coherence measures provided by Palmetto. The results are listed in table 2. Generally, correlation between topic coherence scores and mean usefulness score is much lower than the correlations reported by Röder et al. (2015). Röder et al. (2015) report average correlation coefficients between 0.358 and 0.731, whereas the maximum correlation coefficient in our study is 0.416. Moreover, for C_A, C_V, and *UMass*, the correlation coefficients are not statistically significant. According to the data, *NPMI* is the best topic coherence score for Dutch. However, since it is not clear why our correlation coefficients are lower than the ones reported in the original study, additional research is required.

References

Aletras, N. & Stevenson, M. (2013), 'Evaluating topic coherence using distributional semantics', in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pp. 13–22.

Fang, Y., Si, L., Somasundaram, N. & Yu, Z. (2012), 'Mining contrastive opinions on political texts using cross-perspective topic model', in *the fifth ACM international conference on Web search and data mining*, pp. 63–72.

Krippendorff, K. (1980), *Content Analysis: an Introduction to its Methodology*, Sage.

Röder, M., Both, A. & Hinneburg, A. (2015), 'Exploring the space of topic coherence measures', in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408.