Automatic thematic classification of election manifestos

Suzan Verberne
Radboud University Nijmegen
S.Verberne@cs.ru.nl

Eva D'hondt
Radboud University Nijmegen

Antal van den Bosch
Radboud University Nijmegen

Maarten Marx
University of Amsterdam

March 25, 2013

1 Introduction

Isaac Lipschits (1930–2008) was a Dutch historian and political scientist. One of his works is an annotated collection of election manifestos (party programmes) for the Dutch elections between 1977-1998 (Lipschits, 1977). For each election year he compiled a book with the manifestos published by all parties that participated in that year's elections. Lipschits manually labelled the manifestos with themes: he segmented the manifestos into coherent text fragments, numbered them, and added an index of themes in the back of the book referring to these text numbers.

In the Political Mashup project (Marx, 2009), Dutch political data from 1814 onwards is being digitized and indexed. The data are not only digitized and integrated but also disclosed to the public. The aims of the work presented in the current paper are: (1) to digitize the 1977–1998 Lipschits collections and (2) to build an automatic classifier for more recent, unclassified election manifestos. The starting points for our work are the Lipschits books, scanned as PDF files.

We took the following approach: We first converted the scanned PDFs to XML data in which each text fragment has been annotated with the Lipschits themes. We then used these data to build a classifier that is suited for classifying election manifestos from 2002 onwards using the data from the 1980s and 1990s. We evaluated the results by having a domain expert manually judge a sample of the classified data.

2 Converting the PDF books to enriched publications

The body of Lipschits data are the manifesto texts, which have been labelled with numbers at the places in the text where the topic changes. At the back of the book is a register of alphabetically sorted themes. For each theme, Lipschits has listed the associated parts of the election manifestos per party by topic number.

The PDFs were converted to XML. Due to OCR errors, both the body of the election manifestos and the register needed a substantial amount of clean-up before the data could be published digitally, or used for classification. We used a Perl script for textual clean-up. We were able to convert all texts from the election manifestos of 1998, 1994 and 1986 to annotated XML files of acceptable quality. In the XML data, each text fragment has been annotated with the themes from the Lipschits register. See Table 1 for the statistics of the data per year. The table shows that although the texts are only a few hundred words long, Lipschits assigned more than six themes per text on average.

Table 1: Statistics on the classification data

year	# words	vocabulary size	# texts	# themes	# words/text	# themes/text
1986	290,942	32,836	797	214	365	10.5
1994	269,270	32,499	951	210	283	6.9
1998	244,697	31,162	826	218	296	8.3

3 Classification experiments

We aim at developing a classifier that can assign themes to unseen Dutch election manifestos that were written after Lipschits' work. Since we do not have manually labelled data from these recent years, we have to rely on the older data from the eighties and nineties for training and optimization of the classifier. Therefore, our initial experiments are directed at obtaining the best classification results for the 1998 manifestos, using the data from older years in the training phase.

The classification task we consider is a multi-label classification task, with each Lipschits theme being a class. We selected the implementation of the Balanced Winnow classifier (Littlestone, 1988; Dagan et al., 1997) in the Linguistic Classification System (LCS, Koster et al. (2003)) for this purpose. As evaluation measures we use precision (the proportion of assigned themes that is correct), recall (the proportion of correct themes that have been assigned) and F1. We will also report Mean Average Precision (MAP), which gives the quality of the generated ranking of the themes per text.

While having the 1998 texts as test data, we evaluated the use of the years 1986 and 1994 as training data, both separately and in combination. Of course, better results could be obtained if we would use the theme set from 1998 to train the classifier on. However, this is not realistic since in the future application of the classifier it should be able to classify texts from years for which the theme set is unknown. Therefore, we not only aim at selecting the best training set but also the best theme set from the training data. We compared the use of all themes from 1986 and 1994 to the use of only the themes from 1994. The results are in Table 2 below. We continue our experiments with the data from both 1986 and 1994 as training data, but using only the theme set from 1994.

Table 2: Results for the automatic classification of the 826 texts fragments from the 1998 election manifestos using different sets of training data and a standard bag-of-words representation.

training data	# training texts	# themes	P	R	F1	MAPl
1986 data	797	214	34.9%	29.1%	31.7%	0.546
1994 data	951	211	68.8%	39.2%	49.9%	0.798
1986 and 1994 data	1748	320	68.8%	37.3%	48.4%	0.835
1986 and 1994 data, only 1994 themes	1718	211	72.2%	37.1%	49.0%	0.841

There is evidence that text classification can be improved by adding word bigrams (Braga et al., 2009; Bekkerman and Allan, 2004), also for Dutch (Gaustad and Bouma (2002), p. 11). We extracted all within-sentence bigrams from the texts in our corpus and added them to the bag of unigrams for each text. We also lemmatized the texts using the output of the Dutch morpho-syntactic analyzer Frog (Van den Bosch et al., 2007) and experimented with lemmatized unigrams and bigrams. Finally, we experimented with removing stopwords from the feature set. The results of classification experiments with the variants of the data are in Table 3.

Table 3: Results for the automatic classification of the 826 texts fragments from the 1998 election manifestos using the text examples from the 1986 and 1994 data (themes from 1994) with different text representations. The asterisk denotes that the F-scores for the individual texts are significantly different ($P < 0.01$) from to the unigram baseline.

Text representation	P	R	F1	MAP
unigrams	72.2%	**37.1%**	**49.0%**	0.841
unigrams lemmatized	**73.3%**	36.8%	**49.0%**	**0.850**
unigrams lemmatized without stopwords	72.9%	36.0%	48.2%	0.839
unigrams + bigrams	68.8%	36.4%	47.6%*	0.834
unigrams + bigrams lemmatized	73.1%	36.1%	48.3%	0.845
unigrams + bigrams lemmatized without stopwords	72.4%	36.0%	48.1%	0.837

A paired t-test on F-scores for individual texts ($n = 826$) showed that the only settings that gives significantly different results from the unigram baseline is *unigrams + bigrams* ($P < 0.0001$) — yielding a lower average F-score than the baseline. Since none of the alternative text representations gives results above the unigram baseline, we decide to continue our experiments with unigrams only, wordforms rather than lemmas and without stopword removal.

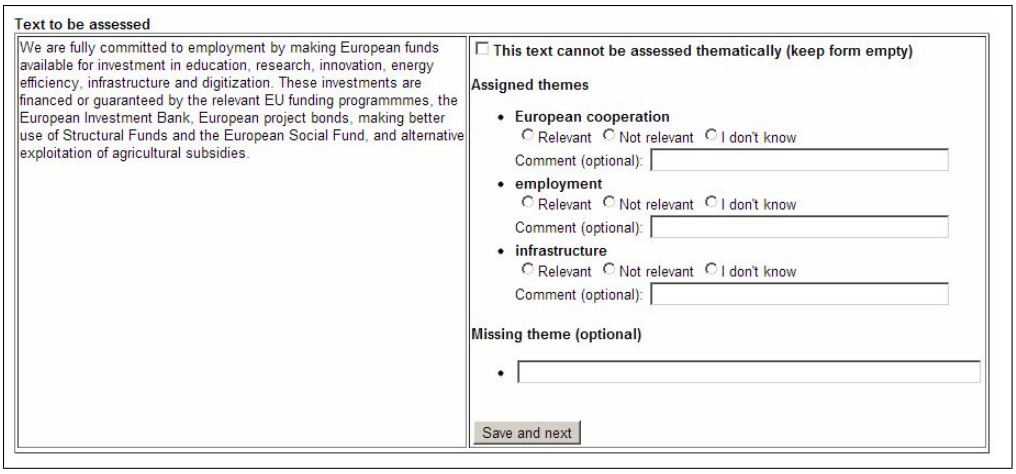

Figure 1: The assessment interface for the evaluation of the classifier output by an expert (translated to English for the reader's convenience). The text fragment is shown on the left side. On the right side are from top to bottom: a checkbox for "this text cannot be assessed thematically"; the themes assigned by the classifier (in this case without theme merging) with radiobuttons for 'relevant', 'not relevant' and 'I don't know' and the option for leaving a comment; and a textbox for missing themes. When typing a string here, matching themes from the set of (merged) themes are shown.

4 Labelling unseen election manifestos

We built an automatic Lipschits classifier that can assign themes to unseen election manifestos. We trained the classifier on all available data from 1986, 1994 and 1998, but only kept the themes from the most recent year: 1998. We used a standard bag-of-words representation without lemmatization and removal of stopwords. We obtained the unlabelled election manifestos from 2006, 2010 and 2012. Statistics on these data are in Table 4.

Table 4: Statistics on the unseen data

year	# words	vocabulary size	# texts	# words/text
2006	235,949	35,547	4,771	49.5
2010	756,254	35,547	21,329	35.5
2012	246,004	30,477	5,880	41.8

We automatically labelled the texts using our Lipschits classifier. Then we asked a domain expert, a political journalist, to manually evaluate a sample of the automatic annotations. The assessment interface is shown in Figure 1.

The expert assessed 200 text fragments picked randomly from the data. She judged seven of these as not assessable because they were too short, so 193 texts remain. In addition, she judged 50 text fragments from the 1998 data, manually labelled by Lipschits (while she was under the assumption that these were also labelled automatically). The results are in Table 5.

Table 5: Results for the automatic classification of 193 text fragments from 2006 onwards into the 218 themes defined by Lipschits in 1998. The themes were manually evaluated by an expert.

Year	# texts	# themes/text	P	R	F1	MAP
2006 (trained on 86+94+98)	39	1.7	74.2%	45.8%	56.6%	0.683
2010 (trained on 86+94+98)	122	1.9	77.4%	55.0%	64.3%	0.714
2012 (trained on 86+94+98)	32	2.0	84.6%	56.1%	67.5%	0.766
1998 (manually by Lipschits)	50	7.7	71.5%	89.0%	79.3%	0.843

When we compare the results obtained with the automatic classification of the 2006–2012 data to the evaluation of the manual Lipschits labelling (the last row), we see that although recall is substantially lower, its precision is higher. This means that our automatic Lipschits classifier is able to emulate Lipschits'

precision, but is not as complete as Lipschits himself. One of the likely reasons for the difference is the effect of the 'theme gap': for themes in the 2012 data that did not exist yet in 1998 we did not have any training data.

5 Conclusions

We digitized three years of Dutch election manifestos annotated by the Dutch political scientist Isaac Lipschits: 2,574 short (∼300-words) texts that Lipschits labelled. There are more than six political themes per text on average. We used these data to train a classifier that can automatically label new, unseen election manifestos with themes.

The general conclusions that we draw from work are (1) in the thematic classification of political texts, a high level of detail seems to be preferred by domain experts; (2) change of themes over the years affects recall of the learned classifier, but (3) precision is comparable to the precision obtained by a human expert labeller. Thus when using old political texts to classify new texts, work is needed to link and expand the set of themes to newer topics.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 380-52-005 (PoliticalMashup).

References

Lipschits, I.. Verkiezingsprogrammas 1977 : verkiezingen voor de Tweede Kamer der Staten-Generaal. 1977. Subsequent editions appeared for the elections in 81, 86, 89, 94, 98.

Marx, M.. Advanced information access to parliamentary debates. J Digit Inf 2009;10(6).

Littlestone, N.. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine learning 1988;2(4):285–318.

Dagan, I., Karov, Y., Roth, D.. Mistake-driven learning in text categorization. In: Proceedings of the Second Conference on Empirical Methods in NLP. 1997, p. 55–63.

Koster, C.H.A., Seutter, M., Beney, J.. Multi-classification of patent applications with winnow. In: Broy, M., Zamulin, A.V., editors. Ershov Memorial Conference; vol. 2890 of *Lecture Notes in Computer Science*. Springer; 2003, p. 546–555.

Braga, I., Monard, M., Matsubara, E.. Combining unigrams and bigrams in semi-supervised text classification. In: EPIA; vol. 9. 2009, p. 489–500.

Bekkerman, R., Allan, J.. Using bigrams in text categorization. Department of Computer Science, University of Massachusetts, Amherst 2004;1003.

Gaustad, T., Bouma, G.. Accurate stemming of dutch for text classification. Language and Computers 2002;45(1):104–117.

Van den Bosch, A., Busser, B., Canisius, S., Daelemans, W.. An efficient memory-based morphosyntactic tagger and parser for dutch. In: Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting. 2007, p. 99–114.