Storting, so far

Emanuelle Lapponi & Martin Søyland

April 19, 2016

We have compiled a dataset with storting speeches from 1998 to 2015. The data was obtained by downloading the relevant pages directly from `http://www.stortinget.no` and processing them to extract relevant information. Should you be curious about the process, keen on replicating the automatic part, peek at the data or read some word counts, see `https://github.com/emanlapponi/storting`.

1 Linguistic pre-processing

Part of the mission of this project is to investigate the usefulness of linguistic features in "classical" text-as-data experiments in political science. In order to gather the relevant annotations (that will be used as a basis for feature extraction), the data we gathered will be pre-processed within LAP (`http://lap.clarino.uio.no`). As of now, LAP is able to conduct pre- processing of single text files; part of Eman's PhD work in the context of this experiment is to design an input processor for LAP that is suitable for this kind of experiment.

Currently, the idea is to accept tsv files as input and output a collection of files containing the linguistically annotated text in CoNLL format (`http://ilk.uvt.nl/conll/#dataformat`), using the id of each line in the input tsv file as a filename for the output conll file. We plan to be done with the pre-processing part of the project before summer 2016.

2 Research questions brainstorming

There are also numerous avenues of research questions that can be paved out based on these kind of data: firstly, a sophisticated classifier is interesting in itself. Second, it is interesting to compare its performance with less sophisticated classifiers ("bag of words"/expert surveys). Third, as we have individual level speech data, we can try to trace within-party unity that is not based on as heavily monitored legislative activities such as roll-calls; are some representatives talking more loosely than others, and do some parties have more drifters than others?

In terms of experimentation, one possible setup here is similar to Høyland et al. (2014), where we classified party affiliation based on representative speeches. Measuring party-wise performance of the classifier could be used as a quantitative way to support ideas such as: AP representatives are generally more in line with their party than FrP ones.

The same setup could be trained on the party data and tested on minister data, to further investigate party drifting when representatives are in government positions.

Another idea is to measure classifier performance when labels are collapsed into groups, for instance using only two LEFT and RIGHT or PRO-DRILL and ANTI-DRILL labels.

Unsupervised vector space models are another possible machine learning tool that could be useful in this task, for instance by measuring average cartesian distance among representatives of one party, seeing whether certain parties are "closer" than others. This might potentially complement the supervised classification setup. It could also be done both with the whole 1999-2015 period or over time, allowing us to track how MPs change their way of communicating over different parliamentary roles – for example, will the same individual MPs of Venstre and KrF talk similarly when (a) in a minority coalition government, (b) opposition to a majority coalition government, and (c) as support parties for a minority coalition?

In light of these experiments, further research avenues could open up, for instance using the position measure as an explanatory variable in secondary research: does position cause voting behavior (obvious)? Will drifters lose their seat at the next election? And, will loyalty be rewarded with higher positions (ministry)? Are bills introduced by drifters less likely to pass?

3 Experimentation

As part of Eman's PhD work, all experimentation will have to work within the LAP ecosystem (though not necessarily through the LAP gui :); this entails further development of the current functionality to enable end-to-end classification experiments. In practical terms, this means integrating suitable tools for supervised and unsupervised classification into LAP, and augmenting the current data model to be able to describe all of the metadata needed for the experiments (party labels, government terms and so on), as well as the data structures needed by the classifiers (e.g. vectors).