UNIVERSITY OF COPENHAGEN

The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification

Costanza Navarretta, costanza@hum.ku.dk
Dorte Halltrup Hansen, dorteh@hum.ku.dk

Background and aims

Subject (policy areas) annotations are used by political science researchers for analysing and comparing different parties' policies in various countries and over time.

The aims of our work are to present:

- a corpus of Danish parliament speeches annotated with subjects (policy areas).
- multi-label classification experiments act to verify the consistency of the annotations of two co-occurring subjects in the data.

The Corpus

- The Danish Parliament Corpus (2009-2017) v.2 contains the transcripts of parliamentary speeches of the Danish Parliament, *Folketinget,* from 6/10-2009 to 7/9-2017.
- It was recently released under the CLARIN-DK infrastructure
  https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/44
- The transcripts of the speeches from the parliament come with information about the speeches' timing and the speakers (name, role, title and party).
- We have added age and gender of the speakers using external sources, and subject annotations (19 categories).
- The corpus is in csv format and consists of 40,841,226 words, 381,949 speeches, 886 files, corresponding to the 886 meetings in the period

Subject annotation method

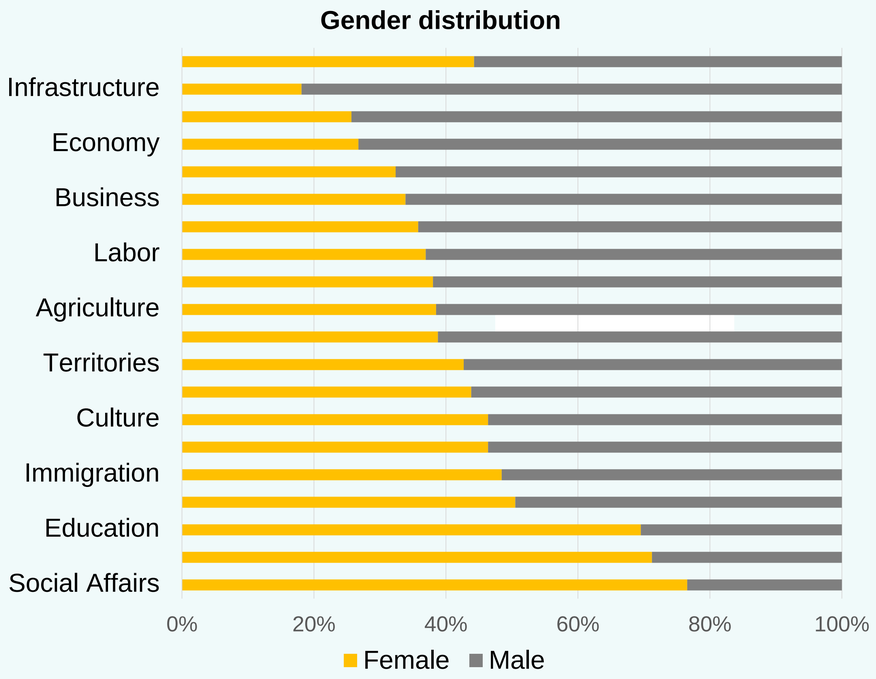
- Extract the titles of the agendas,
- Normalize them, e.g. "First reading of bill 193: XYZ" becomes "XYZ",
- Manually annotate the agenda titles with up to two subjects,
- Add these subjects automatically to each speech under the agenda titles.

Subject categories

Roughly corresponding to the areas of responsibility in the Danish Parliament.

These are mapped to the CAP classification (Comparative Agenda Project):

Agriculture
Business
Culture
Defense
Economy
Education
Energy
Environment
European Integration
Foreign Affairs
Health Care
Housing
Immigration
Infrastructure
Justice
Labour
Local and Regional Affairs
Social Affairs
Territories



The multi-label classification's aims

- Determine whether the annotations of main and secondary subjects in the Danish parliament corpus are consistent and can be reproduced by classifiers.
- Determine the performance of multi-label classifiers on different training data:
  a) BOW and TF*IDF values obtained from the titles of the agenda meetings.
  b) BOW and TF*IDF values obtained from the lemmas of the speeches, and
  c) the data as a) and b), but enriched with information about the speakers (gender, party, role and age).
- Investigate the performance of a number of classifiers on the task.

Multi-label classification

- Multi-label classification made with python 3's scikit-learn and scikit-multilearn modules.
- Classifiers: a) multinomial Naive Bayes (NB), b) support vector machine (SVM) with a linear kernel and c) multilayer perceptron (MLP).
- Baseline: majority classifier that accounts for the classes' frequency.

Multi-label classification: Experiment I

- Dataset: BOW and TF*IDF of agenda titles +/- speaker information.
- Best results: F1-score = 0.997 by SVM trained on BOW values and speaker information. 5-fold cross validation: F1-score = 0.95.
- All classifiers perform significantly better than the baseline (F1-score of 0.022).
- The results confirm that the agenda titles of the Danish parliament can be used for automatically assigning subjects to the speeches without human intervention.

Multi-label classification: Experiment II

- Dataset: BOW and TF*IDF of lemmas of speeches +/- speaker information.
- Best results by MLP trained on the BOW values: F1-score = 0.681.
- 5-fold cross-validation gave an F1-score of 0.631.
- Difference of the results when using BOW or TF*IDF values with or without speaker information is not large for SVM and MLP.
- NB's performance falls dramatically when trained on TF*IDF values.
- Good results given the large subject 1 and subject 2 combinations (186).
- The most frequently occurring combination are those best identified.

Conclusion

- We have accounted for the subject annotation of the Danish Parliament Corpus 2009-17.
- The subject which is more often discussed alone or with other subjects is Economy.
- Subjects addressed more by women than by men are the "softer" ones: Social Affairs, Health care, and Education. Danish female politicians talked as much as male politicians about Environment and Immigration.
- Multi-label classification using the speeches' lemmas gave good results, F1-score just under 0.7. Classification using agenda titles gave F1-score near 1.
- The results indicate that the politicians in the Danish Parliament follow the meetings' agendas during the debates. Therefore, the strategy proposed by political scientists who use agenda titles for annotating the policy areas of political meetings works well for our data.

Future work

- Apply subject classification to other political data and in other languages.
- Use the corpus to test more advanced classification strategies and NLP methods.