Sources of Evidence for Automatic Indexing of Political Texts

Mostafa Dehghani, Hosein Azarbonyad, Maarten Marx, and Jaap Kamps
University of Amsterdam

Motivation

POLITICAL TEXTS on the Web, documenting laws and policies and the process leading to them, are of key importance to government, industry, and every individual citizen.

Laws
What laws are being made?

Questions
What committees are working on?

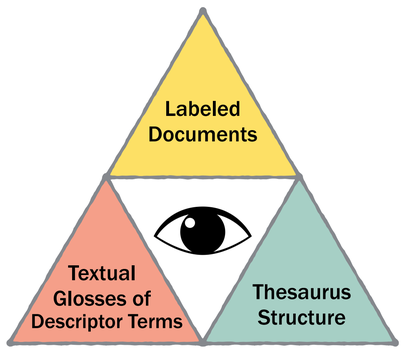
Debates
What parliaments is talking about?

Research Questions:

- What are the different possible sources of evidence for automatic indexing of political text?
- How effective is a learning to rank (LTR) approach integrating a variety of sources of information as features?
- What is the relative importance of each of these sources of information for indexing political text?
- Can we build a lean-and-mean system that approximate the effectiveness of the large LTR system?

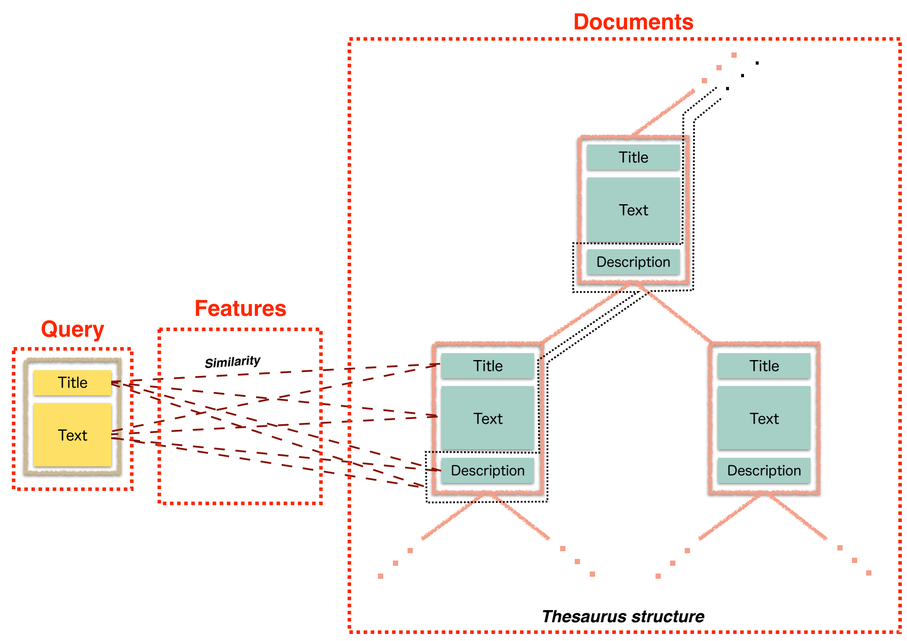
Learning to Rank for Sources of Evidence Combination

- Different **sources of evidence** for the selection of appropriate indexing terms for political documents:



- **Mapping Indexing Problem to the Retrieval Problem:**
  - Features for reflecting the similarity of descriptor terms and documents:
    - Formal models of documents and descriptor terms:

$$Model_D = \langle M(title_D), M(text_D) \rangle$$
$$Model_{DT} = \langle M(title_{DT}), M(text_{DT}), M(gloss_{DT}), M(anc\_gloss_{DT}) \rangle$$



    - For each combination, three IR measures are employed: a) language modeling similarity based on KL-divergence using Dirichlet smoothing, b)the same run using Jelinek-Mercer smoothing, and c) Okapi-BM25.
  - Features for reflecting the characteristics of descriptor terms independent of documents:
    - **Popularity**
    - **Generality**
    - **Ambiguity**

Experiments

Data Collection

- Annotated Data: JRC-Acquis
  - from 2002 to 2006
  - 16,824 documents
- Taxonomy: EuroVoc
  - 6,796 hierarchically structured concepts

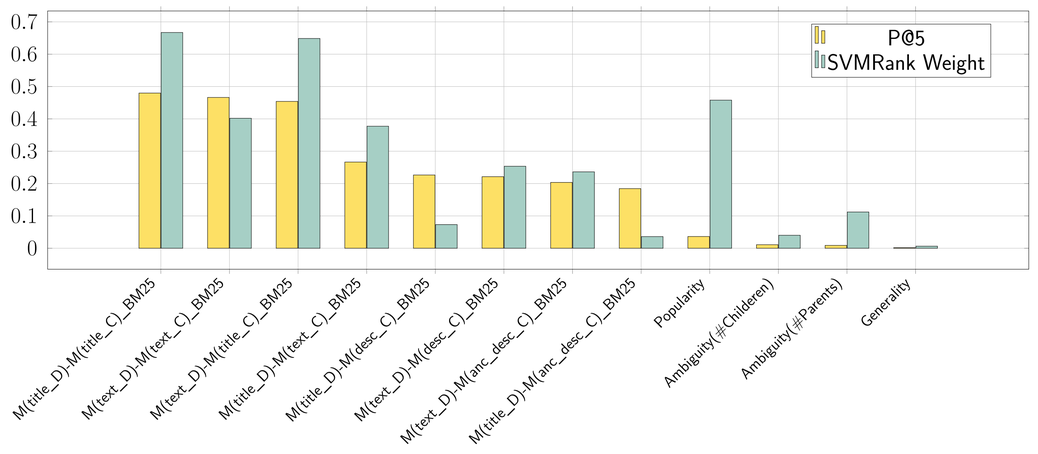
Experiment 1: Performance of LTR

Performance of JEX, best single feature, and LTR methods.

Method	P@5 (%Diff.)	Recall@5 (%Diff.)
JEX	0.4353	0.4863
BM25-TITLES	0.4798 (10%)	0.5064 (4%)
LTR-ALL	0.5206 (20%)	0.5467 (12%)

- Exploiting LTR enables us to learn an effective way to combine features from different sources of evidence.

Experiment 2: Feature Analysis



Feature importance: 1) $P@5$ of individual features, 2) weights in SVM-Rank model

- **Findings:**
  - **Titles** can be considered as the most succinct predictor of classes (Titles of political documents tend to be directly descriptive of the content)
  - **Popularity** is not an effective feature itself, but increases the performance along with other features.

Experiment 3: Lean and Mean System

- **Selected Features:**
  - Similarity of text submodels
  - Similarity of title submodels
  - Similarity of text and textual glosses of descriptor terms
  - Popularity of descriptor terms

Performance of LTR on all features, and on four selected features

Method	P@5 (%Diff.)	all@5 (%Diff.)
LTR-ALL	0.5206 (-)	0.5467 (-)
LTR-TTGP	0.5058 (-3%)	0.5301 (-3%)

- Selective LTR approach is a computationally attractive alternative to the full LTR-ALL approach.

Conclusion

- Using a **learning to rank** approach integrating all features has significantly better performance than previous systems.
- The analysis of feature weights reveals the **relative importance of various sources of evidence**, also giving insight in the underlying classification problem.
- A **lean and mean system** using only four features (text, title, descriptor glosses, descriptor term popularity) is able to perform at 97% of the large LTR model.

EXPOSE: EXPLORATORY POLITICAL SEARCH, Funded by NWO (project# 314-99-108)
NWO Netherlands Organisation for Scientific Research
UNIVERSITY OF AMSTERDAM

{dehghani,h.azarbonyad}@uva.nl