Towards a Tool for Interactive Concept Building for Large Scale Analysis in the Humanities

J. Sonntag¹  A. Blessing²  F. Kliche³  U. Heid³  
J. Kuhn²  M. Stede¹

¹Institute for Applied Computational Linguistics
University of Potsdam, Germany

²Institute for Natural Language Processing
Universitaet Stuttgart, Germany

³Institute for Information Science and Natural Language Processing
University of Hildesheim, Germany

ACL 2013 workshop on: Language Technology for Cultural Heritage, Social Sciences, and Humanities; 8.8.2013
Outline

1. General Organisation and Planning
2. Architecture
3. Case Study
4. Complex Concepts Revisited
The project e-Identity\textsuperscript{1}

Organisation

- is a cooperation between the universities of Stuttgart, Hildesheim and Potsdam
- is a cooperation between political scientists and computational linguists

→ requires lots of communication!

\textsuperscript{1}funded by the Federal Ministry of Education and Research as part of the e-humanities initiative
Complex Concepts

Find **Complex Concepts** in a large collection of newspaper articles (English, German, French).

a. provide automatic classification for as many problems as possible

b. provide a framework to use all available information on the texts to allow for custom interactive learning of further classifications
What are Complex Concepts?

Example

The German leader said it had been an illusion to think that Germans and foreign workers could "live happily side by side".

- declarative-like speech acts from influential politicians
- (multiple) collective identities
- ...
Starting Point

What did political scientists do before?
Starting Point

What did political scientists do before?

- token-based search!
- qualitative research on a dozen articles/texts
What did political scientists do before?

- token-based search!
- qualitative research on a dozen articles/texts
- the token-based search is nevertheless implemented
Token-based search – “peace mission” vs. “foreign intervention”
Outline

1. General Organisation and Planning
2. Architecture
3. Case Study
4. Complex Concepts Revisited
Architecture

What is required to deliver the information the political scientists need?

- *acquisition and filtering*: creation of the corpus, filter all irrelevant articles
- *preprocessing*: tokenization, part-of-speech tagging, syntactic parsing, coreference analysis
- *classification*: sentiment, relation extraction, knowledge acquisition for politicians, active learning for complex concepts

What else is required?

- annotation tools
- useful visualisation
- avoidance of software installation
Figure: Overview of the complete processing chain.
Current state

implemented and running:

- tokeniser, part-of-speech tagger, syntactic parser
- indirect speech detection
- genre classification (for German)
- coreference resolution (for German)
- sentiment analysis (for English)
- active learning of Complex Concepts

to be done:

- knowledge acquisition for politicians
- more tools for French
- classifiers for complex concepts (?)
Integrating documents of different data sources is challenging.

- **Goal** extract and unify all metadata, separate textual content, keep document structure
- **Vision** create general import tools including wizards to stay adaptive and simplify usage for scientists of the humanities

- current state (implemented methods):
  - encoding guesser
  - umlaut restorer
  - integrity check
  - duplicates detection
Technical aspects

- based on Apache UIMA
- newspaper articles are stored in a relational database in the UIMA format
- annotations are added via a web interface
- political scientists access and visualise their texts and information with an HTML rendering component
- the web technology includes JQuery, AJAX, tomcat
1 General Organisation and Planning

2 Architecture

3 Case Study

4 Complex Concepts Revisited
Case Study: Indirect Speech Detection

symbolic approach based on the syntactic analysis based on three conditions:

- look for words tagged as subjunctive
- the verb has to be a direct successor of another verb in the sentence (i.e. it is in an embedded phrase)
- the main verb needs to have a subject
The mission will be stopped because of the risks [to the unarmed observers], informed head of mission Robert Mood on Saturday.
Benefits of indirect speech detection

Which modules benefit or depend on that information?

- „Sentiment“ and finding the correct sources
- coreference analysis
- as features for machine learning algorithms: Commentary/Report Classification
The information itself is important!

Example

„As a government spokesman said, Chancellor Merkel imposed economic sanctions on Syria yesterday.“

- indirect speech (as well as direct speech) is important for the political scientists
- a requirement is a specific vocabulary: declare, initiate, impose, condemns
- depending on the organisation, specific formulations have declarative power
- the speaker is also required to have the legal power to declare sth.

→ similar to declarative speech acts
1. General Organisation and Planning
2. Architecture
3. Case Study
4. Complex Concepts Revisited
Figure: Annotation UI showing possible attributes for a collective identity.
A prime example for a Complex Concept are Multiple Collective Identities.

Example

Laut Perle müssen die Vereinigten Staaten die Osterweiterung der Europäischen Union nutzen, um eine Mehrheit innerhalb der EU für die eigenen Zwecke einzuspannen.

According to Perle the United States have to use the eastern enlargement of the European Union to create a majority within the EU for their own sake.
Complex Concepts

Example

Laut Perle müssen die Vereinigten Staaten die Osterweiterung der Europäischen Union nutzen, um eine Mehrheit innerhalb der EU für die eigenen Zwecke einzuspannen.

According to Perle the United States have to use the eastern enlargement of the European Union to create a majority within the EU for their own sake.

Properties of the MCI:

- attributed by others
- national identity: USA
- commercium
Example

Welch wunderbarer Rollentausch für eine Nation, die im Laufe ihrer tausendjährigen Geschichte so oft von seinen östlichen und westlichen Nachbarn gedemütigt wurde! Dieses Mal sind die Polen nicht die Besetzten, sondern die Besatzer. What a wonderful role reversal for a nation which has been humiliated by their eastern and western neighbors in their thousand year old history! This time Poland is not the occupied territory but the occupying force.

Properties of the MCI:

- national identity: other country
- communio (,,culture“ in a strict sense: history, civilisatory achievements, cultural specifics (food, music, literature))
Active Learning

Why is active learning necessary?

MCI and other complex concepts from the humanities are extremely hard to find.

Why is active learning necessary?

\(a\) our annotations are very **expensive**!

\(b\) the system is always prepared for **new complex concepts**!
Active Learning

What do we do?
Use all possible features from the linguistic preprocessing (tokens, lemmas, grammatical relations, knowledge about entities, sentiment, etc.).
The political scientists are then able to:

- annotate a few instances
- train, classify and annotate a few more instances
- rinse and repeat!
What is the Complex Concept Builder, and what can we learn from it?

- an extended NLP pipeline to aid political scientists to do quantitative research
- building extended pipelines can be difficult, but offers good opportunities to explore synergy effects between the different tools
- the specific scientific concepts researchers of the humanities are looking for are very subtle
- annotating these complex concepts is very expensive and active learning is an ideal concept to reduce these annotation efforts