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

Andre Blessing[1] **Jonathan Sonntag**[2] **Fritz Kliche**[3]
Ulrich Heid[3] **Jonas Kuhn**[1] **Manfred Stede**[2]
[1]Institute for Natural Language Processing
Universitaet Stuttgart, Germany
[2]Institute for Applied Computational Linguistics
University of Potsdam, Germany
[3]Institute for Information Science and Natural Language Processing
University of Hildesheim, Germany

Abstract

We develop a pipeline consisting of various text processing tools which is designed to assist political scientists in finding specific, complex concepts within large amounts of text. Our main focus is the interaction between the political scientists and the natural language processing groups to ensure a beneficial assistance for the political scientists and new application challenges for NLP. It is of particular importance to find a "common language" between the different disciplines. Therefore, we use an interactive web-interface which is easily usable by non-experts. It interfaces an active learning algorithm which is complemented by the NLP pipeline to provide a rich feature selection. Political scientists are thus enabled to use their own intuitions to find custom concepts.

1 Introduction

In this paper, we give examples of how NLP methods and tools can be used to provide support for complex tasks in political sciences. Many concepts of political science are complex and faceted; they tend to come in different linguistic realizations, often in complex ones; many concepts are not directly identifiable by means of (a small set of) individual lexical items, but require some interpretation.

Many researchers in political sciences either work qualitatively on small amounts of data which they interpret instance-wise, or, if they are interested in quantitative trends, they use comparatively simple tools, such as keyword-based search in corpora or text classification on the basis of terms only; this latter approach may lead to imprecise results due to a rather unspecific search as well as semantically invalid or ambigious search words. On the other hand, large amounts of e.g. news texts are available, also over longer periods of time, such that e.g. tendencies over time can be derived. The corpora we are currently working on contain ca. 700,000 articles from British, Irish, German and Austrian newspapers, as well as (yet unexplored) material in French.

Figure 1 depicts a simple example of a quantitative analysis.[1] The example shows how often two terms, *Friedensmission*('peace operation'), and *Auslandseinsatz*('foreign intervention') are used in the last two decades in newspaper texts about interventions and wars. The long-term goal of the project is to provide similar analysis for complex concepts. An example of a complex concept is the evocation of *collective identities* in political contexts, as indirect in the news. Examples for such *collective identities* are: the Europeans, the French, the Catholics.

The objective of the work we are going to discuss in this paper is to provide NLP methods and tools for assisting political scientists in the exploration of large data sets, with a view to both, a detailed qualitative analysis of text instances, and a quantitative overview of trends over time, at the level of corpora. The examples discussed here have to do with (possibly multiple) collective identities. Typical context of such identities tend to report communication, as direct or as indirect speech. Examples of such contexts are given in 1.

(1) Die Europäer würden die Lücke füllen,
The Europeans would the gap fill,

[1]The figure shows a screenshot of our web-based prototype.

Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 55–64,
Sofia, Bulgaria, August 8 2013.

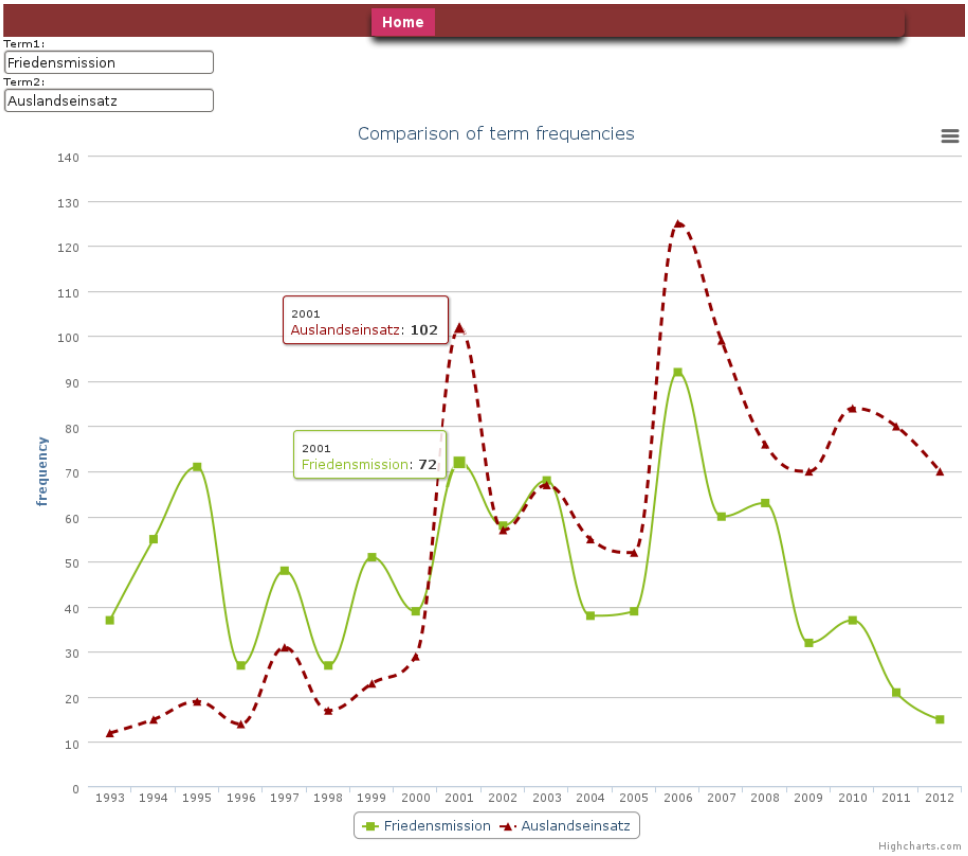

Figure 1: The screenshot of our web-based system shows a simple quantitative analysis of the frequency of two terms in news articles over time. While in the 90s the term Friedensmission (peace operation) was predominant a reverse tendency can be observed since 2001 with Auslandseinsatz (foreign intervention) being now frequently used.

sagte Rühe.
said Rühe.

„The Europeans would fill the gap, Rühe said."

The tool support is meant to be semi-automatic, as the automatic tools propose candidates that need to be validated or refused by the political scientists.

We combine a chain of corpus processing tools with classifier-based tools, e.g. for topic classifiers, commentary/report classifiers, etc., make the tools interoperable to ensure flexible data exchange and multiple usage scenarios, and we embed the tool collection under a web (service) - based user interface.

The remainder of this paper is structured as follows. In section 2, we present an outline of the architecture of our tool collection, and we motivate the architecture. Section 3 presents examples of implemented modules, both from corpus processing and search and retrieval of instances of complex concepts. We also show how our tools are related to the infrastructural standards in use in the CLARIN community. In section 4, we exemplify the intended use of the methods with case studies about steps necessary for identifying evocation: being able to separate reports from comments, and strategies for identifying indirect speech. Section 6 is devoted to a conclusion and to the discussion of future work.

2 Project Goals

A collaboration between political scientists and computational linguists necessarily involves finding a common language in order to agree on the precise objectives of a project. For example, social scientists use the term codebook for manual annotations of text, similar to annotation schemes or guidelines in NLP. Both disciplines share methodologies of interactive text analysis which combine term based search, manual annotation and learning-based annotation of large amounts of data. In this section, we give a brief

summary of the goals from the perspective of each of the two disciplines, and then describe the text corpus that is used in the project. Section 3 will describe our approach to devising a system architecture that serves to realize the goals.

2.1 Social Science Research Issue

Given the complexity of the underlying research issues (cf. Section 1) and the methodological tradition of manual text coding by very well-trained annotators in the social science and particular in political science, our project does not aim at any fully-automatic solution for empirical issues in political science. Instead, the goal is to provide as much assistance to the human text analyst as possible, by means of a workbench that integrates many tasks that otherwise would have to be carried out with different software tools (e.g., corpus preprocessing, KWIC searches, statistics). In our project, the human analyst is concerned specifically with manifestations of collective identities in newspaper texts on issues of war and military interventions: who are the actors in political crisis management or conflict? How is this perspective of responsible actors characterized in different newspapers (with different political orientation; in different countries)? The analyst wants to find documents that contain facets of such constellations, which requires search techniques involving concepts on different levels of abstraction, ranging from specific words or named entities (which may appear with different names in different texts) to event types (which may be realized with different verb-argument configurations). Thus the text corpus should be enriched with information relevant to such queries, and the workbench shall provide a comfortable interface for building such queries. Moreover, various types and (possibly concurrent) layers of human annotations have to complement the automatic analysis, and the manual annotation would benefit from automatic control of codebook[2] compliance and the convergence of coding decisions.

2.2 Natural Language Processing Research Issue

Large collections of text provide an excellent opportunity for computational linguists to scale their methods. In the scenario of a project like ours, this becomes especially challenging, because standard automatic analysis components have to be combined with manual annotation or interactive intervention of the human analyst.

In addition to this principled challenge, there may be more mundane issues resulting from processing corpora whose origin stretches over many years. In our case, the data collection phase coincided with a spelling reform in German-speaking countries. Many aspects of spelling changed twice (in 1996 and in 2006), and thus it is the responsibility of the NLP branch of the project to provide an abstraction over such changes and to enable today's users to run a homogeneous search over the texts using only the current spelling. While this might be less important for generic web search applications, it is of great importance for our project, where the overall objective is a combination of quantitative and qualitative text analysis.

In our processing chain, we first need to harmonize the data formats so that the processing tools operate on a common format. Rather than defining these from scratch, we aim at compatibility with the standardization efforts of CLARIN[3] and DARIAH[4], two large language technology infrastructure projects in Europe that in particular target eHumanities applications. One of the objectives is to provide advanced tools to discover, explore, exploit, annotate, analyse or combine textual resources. In the next section we give more details about how we interact which the CLARIN-D infrastructure (Boehlke et al., 2013).

3 Architecture

The main goal is to provide a web-based user-interface to the social scientist to avoid any software installation. Figure 2 presents the workflow of the different processing steps in this project. The first part considers format issues that occur if documents from different sources are used. The main challenge is to recognize metadata correctly. Date and source name are two types of metadata which are required for analyses in the social sciences. But also the separation of document content (text) and metadata is important to ensure that only real content is processed with the NLP methods. The results are stored in a repository which uses a relational database as a back-end. All further modules are used to add more annotations to the textual data. First a complex linguistic pro-

[2] or, in NLP terms: annotation scheme.

[3] http://www.clarin.eu/

[4] http://www.dariah.eu/

cessing chain is used to provide state-of-the-art corpus linguistic annotations (see Section 3.2 for details). Then, to ensure that statistics over occurrence counts of words, word combinations and constructions are valid and not blurred by the multiple presence of texts or text passages in the corpus, we filter duplicates. Duplicates can occur if our document set contains the same document twice or if two documents are very similar, e.g. they differ in only one sentence.

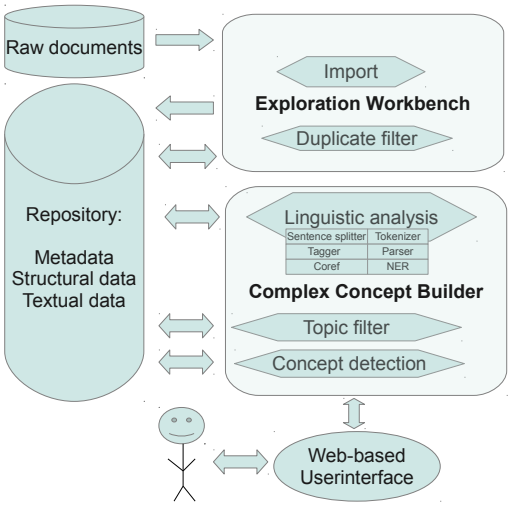

Figure 2: Overview of the complete processing chain.

We split the workflow for the user into two parts: The first part is only used if the user imports new data into the repository. For that he can use the exploration workbench (Section 3.1). Secondly, all steps for analyzing the data are done with the Complex Concept Builder (Section 3.2).

3.1 Exploration Workbench

Formal corpus inhomogeneity (e.g. various data formats and inconsistent data structures) are a major issue for researchers working on text corpora. The web-based "Exploration Workbench" allows for the creation of a consistent corpus from various types of data and prepares data for further processing with computational linguistic tools. The workbench can interact with to existing computational linguistic infrastructure (e.g. CLARIN) and provides input for the repository also used by the Complex Concept Builder.

The workbench converts several input formats (TXT, RTF, HTML) to a consistent XML representation. The conversion tools account for different file encodings and convert input files to Unicode (UTF-8). We currently work on newspaper articles wrapped with metadata. Text mining components read out those metadata and identify text content in the documents. Metadata appear at varying positions and in diverse notations, e.g. for dates, indications of authors or newspaper sections. The components account for these variations and convert them to a consistent machine readable format. The extracted metadata are appended to the XML representation. The resulting XML is the starting point for further computational linguistic processing of the source documents.

The workbench contains a tool to identify text duplicates and semi-duplicates via similarity measures of pairs of articles (Kantner et al., 2011). The method is based on a comparison of 5-grams, weighted by significance (tf-idf measure (Salton and Buckley, 1988)). For a pair of documents it yields a value on a "similarity scale" ranging from 0 to 1. Values at medium range (0.4 to 0.8) are considered semi-duplicates.

Data cleaning is important for the data-driven studies. Not only duplicate articles have a negative impact, also articles which are not of interest for the given topic have to be filtered out. There are different approaches to classify articles into a range of predefined topics. In the last years LDA (Blei et al., 2003; Niekler and Jähnichen, 2012) is one of the most successful methods to find topics in articles. But for social scientists the categories typically used in LDA are not sufficient. We follow the idea of Dualist (Settles, 2011; Settles and Zhu, 2012) which is an interactive method for classification. The architecture of Dualist is based on MALLET (McCallum, 2002) which is easily integrable into our architecture. Our goal is to design the correct feature to find relevant articles for a given topic. Word features are not sufficient since we have to model more complex features (cf. Section 2.1).

The workbench is not exclusively geared to the data of the current project. We chose a modular set-up of the tools of the workbench and provide user-modifiable templates for the extraction of various kinds of metadata, in order to keep the workbench adaptable to new data and to develop tools suitable for data beyond the scope of the current corpus.

3.2 Complex Concept Builder

A central problem for political scientists who intend to work on large corpora is the linguistic variety in the expression of technical terms and complex concepts. An editorial or a politician cited in a news item can mobilize a collective identity which can be construed from e.g. regional or social affiliation, nationality or religion. A reasonable goal in the context of the search for collective identity evocation contexts is therefore to find all texts which (possibly) contain collective identities. Moreover, while we are training our interactive tools on a corpus on wars and military interventions the same collective identities might be expressed in different ways in a corpus i.e. on the Eurocrisis.

From a computational point of view, many different tools need to be joined to detect interesting texts. An example application could be a case where a political scientist intends to extract newspaper articles that cite a politician who tries to rally support for his political party. In order to detect such text, we need a system to identify direct and indirect speech and a sentiment system to determine the orientation of the statement. These systems in turn need various kinds of preprocessing starting from tokenization over syntactic parsing up to coreference resolution. The Complex Concept Builder is the collection of all these systems with the goal to assist the political scientists.

So far, the Complex Concept Builder implements tokenization (Schmid, 2009), lemmatisation (Schmid, 1995), part-of-speech tagging (Schmid and Laws, 2008), named entity detection (Faruqui and Padó, 2010), syntactical parsing (Bohnet, 2010), coreference analysis for German (Lappin and Leass, 1994; Stuckardt, 2001), relation extraction (Blessing et al., 2012) and sentiment analysis for English (Taboada et al., 2011).

It is important for a researcher of the humanities to be able to adapt existing classification systems according to his own needs. A common procedure in both, NLP and political sciences, is to annotate data. Therefore, one major goal of the project and the Complex Concept Builder is to provide machine learning systems with a wide range of possible features — including high level information like sentiment, text type, relations to other texts, etc. — that can be used by non-experts for semi-automatic annotation and text selection. Active learning is used to provide immediate results that can then be improved continuously. This aspect of the Complex Concept Builder is especially important because new or adapted concepts that may be looked for can be found without further help of natural language processing experts.

3.3 Implementation

We decided to use a web-based platform for our system since the social scientist needs no software installation and we are independent of the used operating system. Only a state-of-the-art web-browser is needed. On the server side, we use a tomcat installation that interacts with our UIMA pipeline (Ferrucci and Lally, 2004). A HTML-rendering component designed in the project (and parametrizable) allows for a flexible presentation of the data. A major issue of our work is interaction. To solve this, we use JQuery and AJAX to dynamically interact between client- and server-side.

4 Case Study

In this section we explore the interaction between various sub-systems and how they collaborate to find complex political concepts. The following Section 4.1 describes the detection of direct and indirect speech and its evaluation follows in Section 4.2. Section 4.3 is a general exploration of a few selected sub-systems which require, or benefit from direct and indirect speech. Finally, Section 4.4 discusses a specific usage scenario for indirect speech.

4.1 Identifying Indirect Speech

The Complex Concept Builder provides analyses on different linguistic levels (currently morphosyntax, dependency syntax, named entities) of annotation. We exploit this knowledge to identify indirect speech along with a mentioned speaker. Our indirect speech recognizer is based on three conditions: i) Consider all sentences that contain at least one word which is tagged as subjunctive (i.e. "*.SUBJ") by the RFTagger. ii) This verb has to be a direct successor of another verb in the dependency tree. iii) This verb needs to have a subject.

Figure 3 depicts the dependency parse tree of sentence 2.

(2) Der Einsatz werde wegen der Risiken für die unbewaffneten Beobachter ausgesetzt, teilte

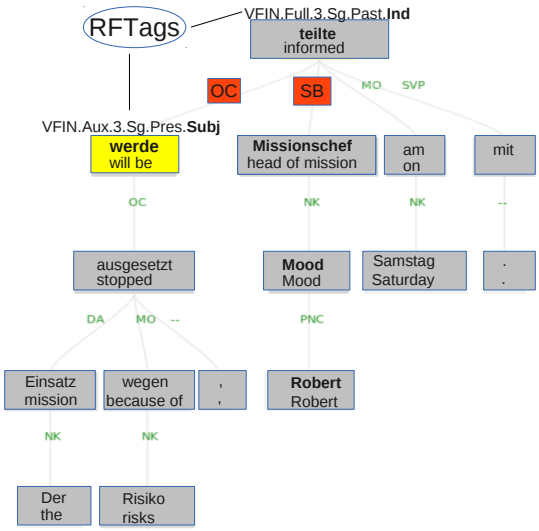

Figure 3: Dependency parse of a sentence that contains indirect speech (see Sentence 2).

Missionschef Robert Mood am Samstag mit.

The mission will be stopped because of the risks to the unarmed observers, informed Head of Mission Robert Mood on Saturday.

The speaker of the indirect speech in Sentence 2 is correctly identified as *Missionschef* (Head of Mission) and the corresponding verb is *teilte mit* (from *mitteilen*) (to inform).

The parsing-based analysis helps to identify the speaker of the citation which is a necessary information for the later interpretation of the citation. As a further advantage, such an approach helps to minimize the need of lexical knowledge for the identification of indirect speech. Our error analysis below will show that in some cases a lexicon can help to avoid false positives. A lexicon of verbs of communication can easily be bootstrapped by using our approach to identify candidates for the list of verbs which then restrict the classifier in order to achieve a higher precision.

4.2 Indirect Speech Evaluation

For a first impression, we present a list of sentences which were automatically annotated as positive instances by our indirect speech detector. The sentences were rated by political scientists. Additionally, for each sentence we extracted the *speaker* and the used *verb of speech*. We manually evaluated 200 extracted triples ($sentence$, $speaker$, $verb\ of\ speech$): The precision of our system is: 92.5%

Examples 2, 3 and 4 present good candidates which are helpful for further investigations on collective identities. In example 3 Cardinal Lehmann is a representative speaker of the Catholic community which is a collective identity. Our extracted sentences accelerate the search for such candidates which amounts to looking manually for needles in a haystack.

example	speaker	verb of speech
(2)	Robert Mood	teilte (told)
(3)	Kardinal Karl Lehmann	sagte (said)
(4)	Sergej Ordzhonikidse	sagte (said)
(5)	Bild (picture)	trüben (tarnish)
(6)	sein (be)	sein (be)

Examples 5 and 6 show problems of our first approach. In this case, the speaker is not a person or an organisation, and the verb is not a verb of speech.

(3) Ein Angriffskrieg jeder Art sei ” sittlich verwerflich ”, sagte der Vorsitzende der Bischoffskonferenz, Kardinal Karl Lehmann.

Any kind of war of aggression is ”morally reprehensible,” said the chairman of the Bishops’ Conference, Cardinal Karl Lehmann.

(4) Derartige Erklärungen eines Staatschefs seien im Rahmen der internationalen Beziehungen inakzeptabel, sagte der UN-Generaldirektor Sergej Ordzhonikidse gestern in Genf.

Such statements of heads of states are unacceptable in the context of international relations, said UN General Director Sergei Ordzhonikidse in Geneva yesterday.

(5) Würden die Wahlen verschoben, trübte sich das geschönte Bild.

Would the elections be postponed, the embellished image would tarnish.

(6) Dies sei alles andere als einfach, ist aus Offizierskreisen zu hören.

This is anything but simple, is to hear from military circles.

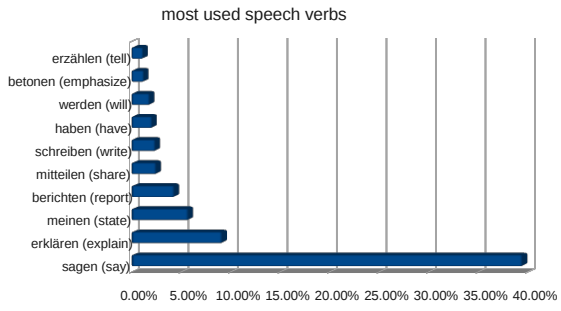

Figure 4: 10 most used verbs (lemma) in indirect speech.

4.3 Using Indirect Speech

Other modules benefit from the identification of indirect speech, as can be seen from Sentence 7. The sentiment system assigns a negative polarity of -2.15 to the sentence. The nested sentiment sources, as described by (Wiebe et al., 2005), of this sentence require a) a direct speech with the speaker "Mazower" and b) an indirect speech with the speaker "no one" to be found.[5]

(7) "There were serious arguments about what should happen to the Slavs and Poles in eastern Europe," says Mazower, "and how many of them should be sent to the camps and what proportion could be Germanised . . . No one ever came out and directly said Hitler had got it wrong, but there was plenty of implied criticism through comparisons with the Roman empire. [...]"[6]

A collective identity evoked in Sentence 7 is "the Germans"— although the term is not explicitly mentioned. This collective identity is described as non-homogeneous in the citation and can be further explored manually by the political scientists.

The following are further applications of the identified indirect speeches a) using the frequency of speeches per text as a feature for classification; e.g. a classification system for news reports/commentaries as described in Section 4.4 b) a project-goal is to find texts in which collective identities are mobilised by entities of political debate (i.e. persons, organisations, etc.); the detection of indirect speech is mandatory for any such analysis.

4.4 Commentary/Report Classification

A useful distinction for political scientists dealing with newspaper articles is the distinction between articles that report objectively on events or backgrounds and editorials or press commentaries.

We first extracted opinionated and objective texts from DeReKo corpus (Stede, 2004; Kupietz et al., 2010). Some texts were removed in order to balance the corpus. The balanced corpus contains 2848 documents and has been split into a development and a training and test set. 570 documents were used for the manual creation of features. The remaining 2278 documents were used to train and evaluate classifiers using 10-fold cross-validation with the WEKA machine learning toolkit (Hall et al., 2009) and various classifiers (cf. Table 1).

The challenge is that the newspaper articles from the training and evaluation corpus come from different newspapers and, of course, from different authors. Commentaries in the yellow press tend to have a very different style and vocabulary than commentaries from broadsheet press. Therefore, special attention needs to be paid to the independence of the classifier from different authors and different newspapers. For this reason, we use hand-crafted features tailored to this problem. In return, this means omitting surface-form features (i.e. words themselves).

The support vector machine used the SMO algorithm (Platt and others, 1998) with a polynomial kernel $K(x, y) =< x, y >^e$ with $e = 2$. All other algorithms were used with default settings.

	precision	recall	f-score
SVM	0.819	0.814	0.813
Naive Bayes	0.79	0.768	0.764
Multilayer Perceptron	0.796	0.795	0.794

Table 1: Results of a 10-fold cross-validation for various machine learning algorithms.

A qualitative evaluation shows that direct and indirect speech is a problem for the classifier. Opinions voiced via indirect speech should not lead to a classification as 'Commentary', but should be ignored. Additionally, the number of

[5]The reported sentiment value for the whole sentence is applicable only to the direct speech. The indirect speech (i.e. "Hitler had got it wrong") needs a more fine-grained polarity score. Since our Complex Concept Builder is very flexible, it is trivial to score each component separately.

[6]http://www.guardian.co.uk/education/2008/jul/01/academicexperts.highereducationprofile

uses of direct and indirect speech by the author can provide insight into the intention of the author. A common way to voice one's own opinion, without having to do so explicitly, is to use indirect speech that the author agrees with. Therefore, the number of direct and indirect speech uses will be added to the classifier. First experiments indicate that the inclusion of direct and indirect speech increase the performance of the classifier.

5 Related Work

Many approaches exist to assist social scientists in dealing with large scale data. We discuss some well-known ones and highlight differences to the approach described above.

The Europe Media Monitor (EMM) (Steinberger et al., 2009) analyses large amounts of newspaper articles and assists anyone interested in news. It allows its users to search for specific topics and automatically clusters articles from different sources. This is a key concept of the EMM, because it collects about $100,000$ articles in approximately 50 languages per day and it is impossible to scan through these by hand. EMM users are EU institutions, national institutions of the EU member states, international organisations and the public (Steinberger et al., 2009).

The topic clusters provide insight into "hot" topics by simply counting the amount of articles per cluster or by measuring the amount of news on a specific topic with regards to its normal amount of news. Articles are also data-mined for geographical information, e.g. to update in which geographical region the article was written and where the topic is located. Social network information is gathered and visualised as well.

Major differences between the EMM and our approach are the user group and the domain of the corpus. The complex concepts political scientists are interested in are much more nuanced than the concepts relevant for topic detection and the construction of social networks. Additionally, the EMM does not allow its users to look for their own concepts and issues, while this interactivity is a central contribution of our approach (cf. Sections 1, 2.1 and 3.2).

The CLARIN-D project also provides a web-based platform to create NLP-chains. It is called WebLicht (Hinrichs et al., 2010), but in its current form, the tool is not immediately usable for social scientists as the separation of metadata and textual data and the encoding of the data is hard for non-experts. Furthermore, WebLicht does not yet support the combination of manual and automatic annotation needed for text exploration in the social science. Our approach is based on the webservices used by WebLicht. But in contrast to WebLicht, we provide two additional components that simplify the integration (exploration workbench) and the interpretation (complex concept builder) of the research data. The former is intended, in the medium term, to be made available in the CLARIN framework.

6 Conclusion and Outlook

We developed and implemented a pipeline of various text processing tools which is designed to assist political scientists in finding specific, complex concepts within large amounts of text. Our case studies showed that our approach can provide beneficial assistance for the research of political scientists as well as researcher from other social sciences and the humanities. A future aspect will be to find metrics to evaluate our pipeline. In recently started annotation experiments on topic classification Cohen's kappa coefficient (Carletta, 1996) is mediocre. It may very well be possible that the complex concepts, like multiple collective identities, are intrinsically hard to detect, and the annotations cannot be improved substantially.

The extension of the NLP pipeline will be another major working area in the future. Examples are sentiment analysis for German, adding world knowledge about named entities (e.g. persons and events), identification of relations between entities.

Finally, all these systems need to be evaluated not only in terms of f-score, precision and recall, but also in terms of usability for the political scientists. This also includes a detailed investigation of various political science concepts and if they can be detected automatically or if natural language processing can help the political scientists to detect their concepts semi-automatically. The definition of such evaluation is an open research topic in itself.

Acknowledgements

The research leading to these results has been done in the project eIdentity which is funded from the Federal Ministry of Education and Research (BMBF) under grant agreement 01UG1234.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Andre Blessing, Jens Stegmann, and Jonas Kuhn. 2012. SOA meets relation extraction: Less may be more in interaction. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities*, pages 6–11.

Volker Boehlke, Gerhard Heyer, and Peter Wittenburg. 2013. IT-based research infrastructures for the humanities and social sciences - developments, examples, standards, and technology. *it - Information Technology*, 55(1):26–33, February.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational*, pages 89–97.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

D. Ferrucci and A. Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

Cathleen Kantner, Amelie Kutter, Andreas Hildebrandt, and Mark Puettcher. 2011. How to get rid of the noise in the corpus: Cleaning large samples of digital newspaper texts. *International Relations Online Working Paper*, 2, July.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The german reference corpus dereko: a primordial sample for linguistic research. In *Proceedings of the 7th conference on international language resources and evaluation (LREC 2010)*, pages 1848–1854.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Andreas Niekler and Patrick Jähnichen. 2012. Matching results of latent dirichlet allocation for text. In *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling*, pages 317–322. Universitätsverlag der TU Berlin.

John Platt et al. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. *technical report msr-tr-98-14, Microsoft Research*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK, August.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Helmut Schmid, 2009. *Corpus Linguistics: An International Handbook*, chapter Tokenizing and Part-of-Speech Tagging. Handbooks of Linguistics and Communication Science. Walter de Gruyter, Berlin.

Burr Settles and Xiaojin Zhu. 2012. Behavioral factors in interactive training of text classifiers. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 563–567. Association for Computational Linguistics.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics.

Manfred Stede. 2004. The potsdam commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, pages 96–102, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, and Erik Van Der Goot. 2009. An introduction to the europe media monitor family of applications. In *Proceedings of the Information Access in a Multilingual World-Proceedings of the SIGIR 2009 Workshop*, pages 1–8.

Roland Stuckardt. 2001. Design and enhanced evaluation of a robust anaphor resolution algorithm. *Computational Linguistics*, 27(4):479–506.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.