Supporting Frame Analysis using Text Mining

S. Ananiadou¹, D. Weissenbacher¹, B. Rea¹, E. Pieri², F. Vis², Y. Lin², R. Procter², P. Halfpenny²

¹National Centre for Text Mining, University of Manchester, 131 Princess Street, M1 7DN, UK
²National Centre for e-Social Science, University of Manchester, Oxford Rd, Manchester M13 9PL, UK

Corresponding author: Sophia.Ananiadou@manchester.ac.uk

Abstract. In this paper, we describe how text mining (TM) has been used to support frame analysis within the context of the ASSIST project. This technology is capable of retrieving knowledge from unstructured text and presenting it to researchers in a concise form. In the course of the ASSIST project, which is a collaboration between social researchers and text mining experts, we have designed and prototyped a specialized search engine to help researchers who specialise in the frame analysis of news articles. After describing each of the TM modules which compose our system and the functionalities they provide, we present the results of their evaluation.

Introduction

The rapid increase of digital content in newswires and other sources allows information to be made immediately available to a wide audience. Social scientists, faced with an overwhelming amount of information, are turning to new automated techniques to support their research methodology and to handle the huge amount of information. Text mining (TM) is a novel technology which retrieves knowledge from unstructured text and presents the distilled knowledge to users in a concise form (Ananiadou et al., 2009). The advantage of text mining is that it enables researchers to collect, maintain, interpret, curate and discover knowledge needed for research or education in an efficient and systematic matter (Ananiadou and McNaught, 2006). Text mining annotates documents with entities and facts of interest to the user, thus enabling information extraction (IE). Its goal is to extract important information from textual sources without requiring the end user of the information to read the text themselves (McNaught and Black, 2006).

In this paper, we describe the ASSIST project¹ in which text mining has been used to support
a particular analytical technique known as frame analysis. This project is joint work between the National Centre for Text Mining (NaCTeM), which provides the text mining technology, and the National Centre for e-Social Science (NCeSS), which provides user requirements and expertise in frame analysis.

In the next section we define the frame analysis process and justify the use of text mining to support this process. In the third section, after commenting the solutions proposed by the existing software used for Frame analysis to integrate TM, we present the ASSIST search engine and the evaluation realized during the frame analysis of the introduction of the ID cards in UK. The last section concludes with the evaluation of the system, identifies possible improvements and outlines the next stage in its development.

Frame Analysis

A key feature of frame analysis is its heterogeneous nature. Frames were first envisaged by (Goffman, 1974) but since then a wide range of disciplines has theorised them in different ways. Some scholars take them to be like cognitive schemata (Snow and Benford, 1988), similar to mental scaffolding through which we perceive, organise and communicate experience more or less unconsciously. At the other end of the spectrum are those scholars who believe frames are better understood as conscious devices used strategically for casting ‘events’ in a certain light, defining ‘the issues’ that we ought to attend to, and prioritising some interventions and responses over other possible ones (Entman, 1993).

The different disciplines that use frame analysis – including cognitive psychology, sociology, politics, journalism, media, communication, and cultural studies – emphasise different accounts of framing. These approaches vary in their understanding of how frames are produced, whether these frames are more or less conscious and deliberate, and whether they are mainly the result of individual or culturally-embedded practices. Furthermore, studies using frame analysis can focus on identifying and discussing the frames as recovered by the analysts in certain discourses or genres, they can include an exploration of how these frames may relate to the ‘intentions’ of the frame producers, or an investigation of how certain frames may be perceived by communities of readers and ‘consumers’ of texts and discourses. This means that frame analysis can be thought of a multi-method approach and one that relies and lends itself to making use of a variety of data sources too – including, for instance, written texts, oral accounts and interviews, images and visual representations.

The focus of this frame analysis was the stories in UK national newspapers on the topic of ID cards (Pieri, 2009).

The presence of frames in the text is not immediately visible or evident, however. The type of
frame analysis examined here traditionally requires an intensive research process, performed mostly manually by social scientists. Researchers usually read texts piece by piece, sorting out textual elements manually, and tagging them with the selected category’s name (or code). This is a laborious and time-consuming task, even with the use of current computer-assisted qualitative data analysis (CAQDAS) packages, such as Atlas.ti2, to assist in tagging the text elements. While document length will vary and the number of analysts involved in a study will also vary, it is very uncommon for a single frame analyst to work with datasets that include thousands of texts.

**Assisting Frame Analysis by Computer**

While there are strands of frame analysis that focus on quantitative content analysis, the appeal of applying text mining is to try and combine qualitative and quantitative research questions and methodologies – for example, to explore whether certain frames identified qualitatively can be recovered also in larger datasets. Furthermore, text mining offers the possibility of finding interesting associations among disparate facts. Some of the key benefits of text mining are therefore in the area of hypothesis generation, in highlighting and bringing to the fore hidden or not immediately intuitive patterns in the data. Text mining might also be used to explore whether certain frames are more generalizable, looking at whether they can be identified over long periods of time, for instance, or over different debates.

In this project, we set out to explore exactly how frame analysis might benefit from the application of text mining techniques, and has involved close and continuous collaboration between social researchers who specialise in frame analysis and text mining experts.

**CAQDAS Software vs Text Mining**

There is a long tradition of using qualitative analysis programs for social science research, some of these go under the name of Computer Assisted Qualitative Data Analysis Software (CAQDAS)3. CAQDAS software can facilitate analysis by allowing visualization of large quantities of extracts that the analyst has recognized and (manually or not) annotated. So it can assist in hypothesis generation, validation and possible discovery through visualization.

Amongst the more recent qualitative software packages released, QDA Miner4 claims to incorporate text mining functionalities. Although practical in use, the text mining functionalities of this system mainly rely on word frequencies, co-occurrences, stoplists and standard data mining techniques (document clustering). Text mining modules have a different

---

3 http://caqdas.soc.surrey.ac.uk/index.html
4 http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html
scope of analysis. Firstly, as some CAQDAS software (Atlas.ti for instance) they are interactive, e.g. feedback from the users can be used to improve and customise the performance of the components. Secondly, they are modular, e.g. text mining workflows can use any text mining component to provide the best result for the user. Text mining workflows like U-Compare\(^5\) provide an interoperable environment which allows users to mix and match the best of breed text mining tools for a specific application (Kano \textit{et al.}, 2009). This infrastructure allows the user to evaluate the performance of a specific text mining workflow for a particular application. Although this infrastructure is currently adopted within the biomedical domain, it can be transferred to the social sciences. Finally, text mining tools can analyse a huge number of documents to extract important information from text in a matter of seconds.

To produce an efficient text mining platform that meets users’ requirements, text miners and social researchers have worked closely together throughout the project. As a starting point, we adapted an existing search engine developed during the ASSERT project to the social science domain.\(^6\) The ASSERT search engine is a modular platform which allows advanced TM tools (largely developed at NaCTeM) to be plugged in. By extending the ASSERT search engine with existing and reliable TM tools, we have first produced a simple prototype. This prototype has been used to elicit requirements in a collaborative way from social researchers involved in frame analysis. Based on their feedback, we have selected and improved the embedded TM tools to obtain the final ASSIST search-engine. The general architecture of the ASSIST search engine is illustrated in Figure 1. We describe each module individually, together with the rationale for integrating it into the system and its performance on our corpus.

**The ASSIST Search Engine Architecture**

The underlying core of the ASSIST system is a specialized search engine, customised to assist in some of the task associated with frame analysis’. When starting a new study the search engine simplifies the work of filtering the documents, allowing users to focus upon smaller sets of documents appropriate to topics of interest. Alternatively, when investigating the frames it facilitates exploration and browsing among the relevant subject matter through a combination of inter-operable and integrated text mining modules.

\(^5\) [http://u-compare.org/index.html](http://u-compare.org/index.html)
\(^6\) [http://www.nactem.ac.uk/assert/](http://www.nactem.ac.uk/assert/)
We have used the well known search engine *Lucene*, for its efficient and powerful software which is used by major international companies and eminent universities alike. The search interface provides familiar querying options, such as Boolean operators and wildcard characters for generic searching. This has been complemented by our addition of several new operators which are able to query the metadata extracted from our documents as well as the semantic information added by the TM modules presented below. This new functionality makes it possible to combine different operators in a query to express a precise idea like in the example: “*ID* AND *date*:2006 AND *CELEBRITY:*blair”. The query will return all documents written in 2006 mentioning the person Prime Minister Blair (as opposed to the company Blair) and containing ID card, ID scheme, etc.

As a concrete case of application we have used the search engine to study the debate surrounding the introduction of ID cards in the UK. We have constructed a corpus of 4889 documents from the LexisNexis newspaper database. This corpus covers a time span from 2003-2008 and has been built with the key words “*ID card, Identity card, National Identity Register, National Identity Scheme, NIS*”.

---

7 Documentation and evaluation of its performances can be found at http://lucene.apache.org/java/docs/features.html.
The traditional presentation of the search results returned in response to a free text query is a list of documents with a short context (called a snippet) showing where the words of the query occurred in the documents. However, when a search returns a large number of documents, the list of snippets can become too long to read and may distract the user with too much information. Our proposed solution for this is to cluster the documents retrieved according to their similarities and to associate a readable label with each cluster. This task generates smaller sets of documents surrounding an automatically identified topic area, and is referred as the search result clustering problem. To address this issue, the search result clustering algorithm Lingo (Osinski, 2003) has been selected and integrated into the ASSIST search engine. The preliminary qualitative evaluation carried out by the social researchers involved in the project has shown that most of the labels are readable and correctly describe the clusters. However, this method can occasionally produce a small number of spurious clusters, representing documents with similar words used in different contexts and without an underlying theme. For instance, the corpus presents some distinctive features which lead the algorithm to generate meaningless titles of clusters. Several documents have the title 'Dear Sun' or 'Letters to the editor' because they are specific pages of newspapers. These noun phrases are long enough to be selected as candidate labels and, as they appear in the document title, they are favoured by the algorithm and used as descriptors. A future version of the ASSIST system will provide the user an interface to fill a stoplist in order to remove potentially unsuitable candidates.

The Text Mining Components

The first semantic annotation module integrated within the platform is a named entity (NE) recognizer. Traditionally, NEs are a noun phrase used as a rigid designator (Kripke, 1982) to denote an existing object in a real or an imaginary world. This module allows the social scientist to locate the main actors and locations contained within documents. For our NE recognizer we have integrated as a baseline system an open source NER, BaLIE (Nadeau et al., 2006), which is based on semi-supervised machine learning. The semi-supervised strategy allowed us to extend the categories of named entities usually recognized. Three main categories are usually recognized:

- person names, e.g., George W. Bush, Mr Chirac, Harry Potter, etc.
- organization names, e.g., Microsoft, ONU, Ford, etc.
- numerical expressions, i.e., date, money, and percentage

Name entities have been judged as not particularly useful in the context of this project. Incidentally, these name entity categories could be of greater interest to social scientists working on network theory analysis for instance.

8 Incidentally, these name entity categories could be of greater interest to social scientists working on network theory analysis for instance.
categories proposed by (Sekine and Nobata, 2004). Our module recognized 26 categories of NEs which can be new subcategories of existing entities (e.g., names of people subdivided to separately identify the names of fictional characters, nicknames, etc.) or new types of NEs (like names of products, disease, etc.).

The extension of the hierarchy of NEs increases the probability of recognizing the main discourse objects in a document and refining the categories of the hierarchy provides precise NE types to the system, which are important for semantic annotations. However, the performance of a recognizer with an increased number of categories depends mainly of the disambiguation rules used to classify a NE which may belong to different categories (e.g., the city Washington rather than persons named Washington). The disambiguation rules of the BaLIE system are simple heuristics justified through experimental results and their performances, whilst in line with similar systems, are not guaranteed when they are applied to different types of NEs. Our evaluation of the BaLIE system across the categories chosen for our corpus confirms this limitation. The performance obtained, shown in Figure 2, is lower than performances published for the BaLIE system. In order to address this problem, we are planning as future work a hybrid approach combining dictionaries, named entity recognisers and disambiguators in a supervised manner (Sasaki et al., 2008).

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>72</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>Organisation</td>
<td>90</td>
<td>69</td>
<td>75</td>
</tr>
<tr>
<td>Location</td>
<td>85</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2 Results of the NE recognizer

NaCTeM’s term extractor TerMine10 complements the NE recognizer by computing the most salient terms in our corpus. A term is the linguistic realisation of a specialised concept. For example, ‘Big Brother’ and ‘DNA database’ are terms automatically extracted from newspaper articles relating to ID cards. Term-based browsing can be used to narrow the scope of search, thus reducing the collection to only those documents that are relevant for frame analysis. An example of this technique can be seen in the search engine NaCTeM has developed for the INTUTE repository11. TerMine is based on C-value, a hybrid and domain independent measure (Frantz et al., 2000) which combines linguistic filters (term formation patterns) as well as statistical information to identify the terms in documents. For our

---

9 The precision expresses the number of errors make by the system. The precision is the number of correct NEs annotated by the system divided by the number of NEs identified. The recall expresses the number of NEs not identified by the system. The recall is the number of NEs correctly annotated by the system divided by the total number of NEs in the document. The F-score combine these scores into a single score expressing the global performance of the system.

10 http://www.nactem.ac.uk/software/termine/

11 http://www.nactem.ac.uk/nactem_irs/doc_search
evaluation (precision) on our corpus we have extracted terms from 300 documents. We asked our social scientist to measure the quality of a sample\(^\text{12}\) of the terms extracted. The list of terms is ranked according to the importance score (C-value scores) assigned to terms by TerMine. The social researcher involved in the project evaluated the quality of a term using a score on a scale from 1 to 5 (1 denotes phrases which are not terms while 5 denotes very good terms). The social scientist has considerable experience across the subject areas discussed in the collection of documents, since a subset of our corpus was used by the social scientist to analyse the frame of ID cards in the traditional way. The results of this evaluation are shown in Figure 3. As expected, the Top-Ranked and the Middle-Ranked terms, demonstrate promising results since they capture an important topic for the social scientists. The drop in precision for terms lower down the list shows that the TerMine scoring algorithm is effective at identifying the core terms and ranking them highly.

<table>
<thead>
<tr>
<th>Top-Ranked Terms</th>
<th>Middle-Ranked Terms</th>
<th>Bottom-Ranked Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>78</td>
<td>64</td>
</tr>
</tbody>
</table>

**Figure 3** Precision of extracted terms

The last TM module integrated is a sentiment analyser, called HYSEAS (Piao et al., 2009), which computes automatically the sentiment (i.e., negative, positive, neutral) of a writer about a fact in an article.\(^\text{13}\) This module has been inserted to explore whether it facilitates the study of the arguments put forward by both supporters and opponents of ID cards. When the syntactic structure of sentences have been analysed, a subjective lexicon proposed by (Wilson et al., 2005) is applied to recognize the tone of the words in the clauses. This lexicon classifies the subjective words according various intensities (e.g., *intrude* is strongly negative and *achieve* weakly positive). The algorithm of our module expresses the overall subjective intensities attached to each word in a clause as a global score associated with the sentence. For instance, the overall sentiment of the writer towards the following event is automatically computed as negative: ‘*Identity cards will intrude on our private lives, MPs and peers said last night*’.

This module has been evaluated on the Multi-Perspective Question Answering (MPQA) corpus (Wiebe et al., 2005). As this corpus is very similar to our own, i.e., newspaper articles, we have not carried out the evaluation of this module on our own corpus. We report the general performance of this module. From a total of 4,706 sentences processed, 3,028 have

\(^{12}\) The sample included a few hundred of terms.

\(^{13}\) [http://text0.mib.man.ac.uk:8080/opminpackage/opinion_analysis](http://text0.mib.man.ac.uk:8080/opminpackage/opinion_analysis)
been correctly classified, which equates to a total precision and recall equal to 64.34% and 64.26% respectively. 40.17% of the errors are caused by neutral sentences that were misclassified. This is partially due to the limited coverage of the lexicon. For future work we will use machine learning methods to complement the rule-based module. However, the inability to get to the pragmatic level of utterance meaning remains an issue for this type of sentiment analysis. Another issue is whether the unit of the sentence is a meaningful way of understanding sentiment.

**Conclusions**

We have designed in a collaborative way a specialized search engine to help social scientists conduct frame analysis studies. Our system is based on an existing modular system ASSERT which we have extended with selected TM components. By extracting automatically the main discourse objects and the salient topics which occur in the text, these components provide a concise access to the content of the document without requiring the social scientist to read the full document. The system evaluation has led on a corpus of newspapers articles being created to analysis the frame of the debate caused by the proposed introduction of ID cards in the UK. We have presented the quantitative results for each module carried out on our corpus.

Qualitative evaluation is ongoing. By screening the functionalities of our system effectively used by the social scientists we will be able to enrich it with additional components. Preliminary conclusions show the needs for additional concise information and a larger interactivity between the system and the user. In the near future, we will integrate an interface dedicated to user annotation for social sciences, similar to the ones provided by the CAQDAS software (Tsuruoka et al., 2008). This interface, once customised to the specific domain, will assist social scientists during the semantic and pragmatic annotations process.

**Bibliography**


