

Deploying Natural Language Processing for Social Science Analysis

Karin Verspoor*, Antonio Sanfilippo[◇], Mark Elmore[‡] and Ed MacKerrow*

*Los Alamos National Laboratory, Los Alamos, NM
[◇]Pacific Northwest National Laboratory, Richland, WA
[‡]Oak Ridge National Laboratory, Oak Ridge, TN

Abstract

We explore the use of natural language processing technologies to assist in content and communication analysis, and argue that there is significant synergy between the goals of this social science analysis and the aims and capabilities of computational linguistics research. We discuss specific technologies that can be deployed for use in social science analysis, and describe the key components of a proposed system in which the use of such technologies can result in a significant benefit to the social science researcher interested in analyzing and formalizing the meaning in documents.

Social scientists often analyze textual data for indicators of the source, purpose, and consequences of communications. In media and political analysis, for instance, texts are scrutinized for evidence of thematic trends and *framing*, or the packaging of information with the intent of creating a particular interpretation [1]. The methodology of *content analysis* has been developed for systematic analysis of the characteristics of messages [1] in support of identification and categorization of texts or text segments relative to the core questions of communication theory: “Who says what, to whom, why, to what extent, and with what effect?”. This methodology includes both qualitative analysis through the coding of document segments in terms of previously established data theories and quantitative analysis of word and code frequencies. It is a methodology that can clearly benefit from automation, and indeed tools known collectively as Computer-Assisted Qualitative Data Analysis Software (CAQDAS) tools have been developed to support coding of documents, frequency analysis, and searching for patterns of words and/or codes in large document sets [3]. However, the bulk of content analysis still proceeds manually, with the social scientist interpreting the results of frequency analysis and searches over documents, and assigning category labels to one document segment at a time. Needless to say, the heavy reliance on manual annotation diminishes the appeal of content analysis as data sets grow larger. Natural Language Processing (NLP) methods can help overcoming this limitation by enabling computer-assisted identification of codes in large document collections and online sources during content analysis. We present a use case describing some specific ways in which technologies from natural language processing can be used to augment CAQDAS tools and facilitate more efficient content analysis. The approach to content analysis takes the notion of *computer-assisted* analysis seriously, where that assistance goes beyond software tools for recording manual annotations and searching for *ad hoc* textual patterns to tools that take advantage of linguistic processing.

Ontologies and lexical management

Efficiency gains in content analysis can be realized by formalizing the concepts and their relations to be searched in documents, through specification of a taxonomy of concepts, or more generally an *ontology*. Different domains require different categories of terms, phrases, and concepts [1]. Development of an ontology for specifying and relating document characteristics and concepts of interest aids in formalizing the coding schemes used and organizing the knowledge extracted. Integration of an ontological backbone into a content analysis tool can additionally result in better coding consistency across coders, as the target categories are clearly defined and the ontology establishes a common controlled vocabulary for concepts.

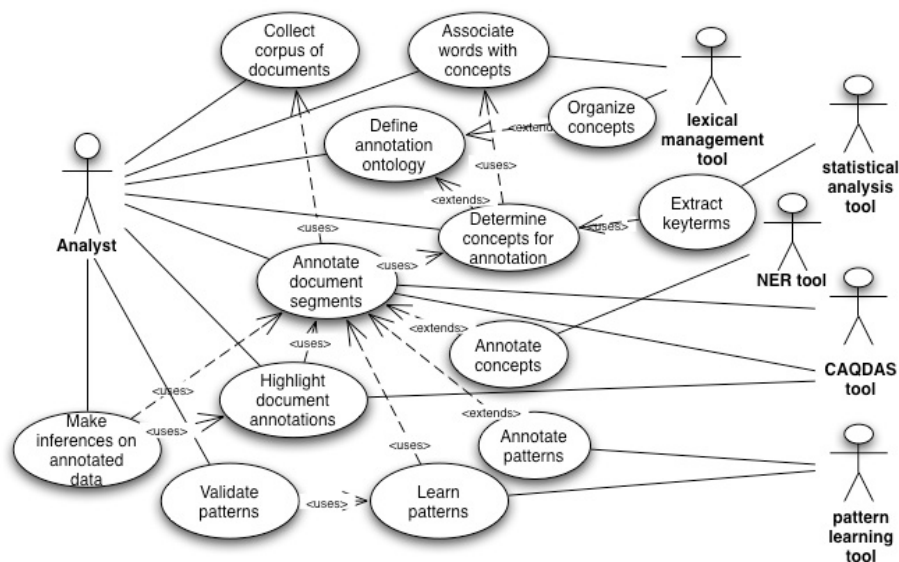
In the use case, the analyst develops a core ontology for a set of frames of interest and the concepts which are indicative of those frames, and ties in associated vocabulary terms with the assistance of existing lexical resources such as WordNet (wordnet.princeton.edu) or Roget’s thesaurus. These resources group words according to their semantic relations: both use synonymy as the key organizing paradigm, which allows easy identification of all of the terms that express a particular concept of interest. WordNet further represents hierarchical relations (hypernyms/hyponyms, or *is-a* relations and meronyms, or *part-of* relations) which can be used to support generalization or specification in concept definitions.

As an example, the concept of “religion” may be important to recognizing a particular frame. The thesaurus indicates that the terms “faith”, “creed”, and “belief” are synonyms of “religion”, and that the adjective “religious” is a syntactic variant of this word – terms that might not immediately come to mind for the researcher or be evidenced in the dataset under analysis, but clearly are relevant and important to searching for this concept in as yet unseen data. Furthermore, the set of hyponyms of “religion” includes “Buddhism”, “Christianity”, and “Islam” *inter alia*, and the investigation of these relations may cause the researcher to refine the concept relevant to the frame to one of these more specific concepts, or he may incorporate those terms into the concept specification.

Named Entity Recognition tool

There are several term categories that have primary importance for content analysis and for which there are automated recognizers available. These are the “named entity” categories of people, places, and organizations, which are particularly relevant for attribution of the sources and targets of communications. Named entity recognition (NER) tools are

Figure 1: Use case for augmenting content analysis with NLP technology.



computational tools that accurately identify such entities in documents. In the use case, a NER tool is used to automatically annotate the document entities. These annotations are presented via the CAQDAS tool's user interface to provide a first-pass coding of the document, which can be accepted as-is, modified, or thrown away by the analyst.

Statistical analysis tool

Statistical tests can be applied to text to discover key words and phrases. Tests such as TF.IDF (Term Frequency/Inverse Document Frequency [4]), pointwise mutual information, and chi-squared distributions go beyond simple frequency counting to identify words and phrases which are the main content-bearing terms in a document. Frequency alone is insufficient since words may be frequent in a document and yet unimportant (consider syntactic function words such as "and", "the", etc. or words that are ubiquitous in a domain such as "television" or "newspaper" in media analysis).

Key words and phrases in a document are likely to be relevant to the content analysis of that document. As such, the terms extracted by a statistical analysis tool can be used by the analyst as a basis for identifying concepts that are important to the domain, which can in turn be used for defining the annotation ontology. The statistical analysis provides insight into the topics represented by documents in the corpus, and can be used by the analyst to quickly get a sense of the range of those topics. Since the keyterms are also automatically annotated in the source document, the analyst can explore the document context of specific words to resolve any ambiguities.

Pattern learning tool

Once an ontology of concepts and terms of interest has been developed, along with a set of texts manually or semi-automatically (e.g. using NER) annotated with those ontological concepts, it is possible to develop algorithms which aim for automated coding. The existing annotations serve as input to a learning algorithm that aims to generalize from the original examples by determining commonalities among them through their linguistic properties [5]. Abstract patterns are inferred which then can be applied to annotate concepts in new documents automatically. These new annotations can be verified and corrected as necessary by the analyst, which facilitates further refinement of the pattern definitions.

Conclusions

The integration of these NLP tools with existing CAQDAS tools enables a content analyst to rapidly explore the domain of interest through a corpus of documents, define the core concepts to be coded, draw in terminology not directly evidenced in the corpus, and drive learning of patterns in support of automated coding. This provides a powerful advance over traditional manual content analysis tools while still providing a document exploration environment which will support in-depth analysis and inference.

References

1. Goffman, Erving. 1974. *Frame Analysis: An Essay on the Organization of Experience*. London: Harper and Row.
2. Holsti, Ole R. 1969. *Content Analysis for the Social Sciences and Humanities*. Reading, Mass.
3. Koenig, T. 2004. "Reframing Frame Analysis: Systematizing the empirical identification of frames using qualitative data analysis software." Presented at the ASA Annual Meeting, San Francisco, CA, August 14-17, 2004.
4. Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5): 513-523.
5. Thelen, M. and Riloff, E. 2002. "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts", *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.