Smart Qualitative Data (SQUAD): Information Extraction in a Large Document Archive

Maria Milosavljevic[*], Claire Grover[*] & Louise Corti[+]

[*]School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
{mmilosav, grover}@inf.ed.ac.uk

[+]ESDS Qualidata, UK Data Archive
University of Essex
Wivenhoe Park, Colchester CO4 3SQ, UK
corti@essex.ac.uk

Abstract

In this paper, we present the results of an investigation into methodologies and technical solutions for exposing the structured metadata contained within digital qualitative data, to make them more shareable and exploitable. In particular, we develop mechanisms for using Information Extraction (IE) technology to provide user-friendly tools for semi-automating the process of preparing qualitative data in the social science domain for digital archiving, in order to archive enriched marked-up data.

1 Introduction

The noble aim of digital archives is to archive documents in digital form in order to provide instant, ubiquitous and searchable long-term access. Social science data archives aim further to capture "moments in time", primarily for historical and sociological purposes. Obviously long-term storage and access is a noteworthy goal, however the objective of providing searchable content is equally important. The transition from physical to digital document archives most often involves the digitisation of texts, but the content of the repository remains primarily as natural language documents and little attention is given to improving long-term search capabilities by enriching the documents through the addition of metadata at the document level. Systematic metadata is usually only routinely used to describe collections of texts, using Dublin Core or in social science, the Data Documentation Initiative (DDI).

A primary goal of the Semantic Web (Berners-Lee *et al.*, 2001) is to augment natural language documents with machine-readable content that can be used by computer systems to perform reasoning about the content or meaning of those documents. In essence, this involves attaching metadata to documents that conform to some encoding scheme such as the Resource Description Framework (RDF). RDF aims to describe the relationship(s) between the things (or entities) in documents, for example, that person A in a document works-for organisation X in the document. In considering what metadata might be useful in the e-social science context, we need to consider the current and potential patterns of use of the data. This requires us to first identify what 'things' in a document might be identified as useful within the social science context.

In this paper we describe the SQUAD[1] project, an endeavour to build an infrastructure for semi-automating the process of metadata extraction for social-science documents contained in the UK Data Archive. The aim of the project is to provide the first step towards exploiting the semantic content found in qualitative data. We achieve this by using Information

[1] Smart Qualitative Data: Methods and Community Tools for Data Mark-Up -- a demonstrator project funded under the ESRC Qualitative Demonstrator Scheme (QUADS). For further details see http://www.data-archive.ac.uk/randd/squad.asp

Extraction (IE) techniques to extract meaningful structured data from unstructured texts. There are several areas of research within IE, and for this project, we concentrated on the utility of Named Entity Recognition (NER) for identifying the entities that will be most broadly useful within the social science context. These include people, locations, organisations, occupations and dates.

In the next section, we outline the current and anticipated needs of e-social science users. We then describe our framework and technological solutions to this problem. We evaluate the effectiveness of our techniques and draw some conclusions about the future of information extraction in e-social science.

2 The e-Social Science Repository

Over the past ten years ESDS Qualidata has contributed to the elucidation of some of the key barriers to re-using data, through extensive everyday contact with 2000 or more qualitative researchers, and the main (perceived) barriers, have been well rehearsed (Corti and Thompson, 2004; Corti, 2000). Today, we see a new generation of qualitative researchers who accept the ESRC's Datasets Policy and its efforts to promote the value of sharing data.

Fielding's (2003) scoping study examined issues for the role of qualitative data in e-social science. He emphasised the need for 'tools that allow data to be published to the Web more easily and support online interrogation of data via standard Web browsers'. E-science offers huge potential for shareable qualitative data, especially linking of multiple data and information sources. We are developing tools for publishing marked-up enriched data and associated linked research materials (such as researcher observation or audio materials) to the web and for longer-term archiving. Sustained and successful collaboration of social scientists and computational scientists is still a new phenomenon, and this project's shared cross-disciplinary innovation will allow a consideration of the role that new technologies can offer in enabling the sharing, archiving and presentation of qualitative data.

3 Information Extraction Technology for e-Social Science

3.1 Identifying Useful Social Science Entities

Information Extraction (IE) is a sub-field of computational linguistics that aims to identify key pieces of information in unstructured texts using 'shallow' text analysis techniques. These techniques include a series of sub-tasks such as tokenisation and sentence boundary detection (Grover *et al.*, 2000), part-of-speech (POS) tagging, and chunking (Grover and Tobin, 2006) and Named Entity Recognition (Curran and Clark, 2003).

A typical IE system will employ Named Entity Recognition (NER) to identify, classify and mark-up particular kinds of proper names and terms. Examples of named entities include the names of people (individuals or groups), organisations, places (both physical locations and geo-political entities), occupations, dates, times and quantities such as sums of money or distances. For specific domains, other entities might also be required, for example genes, proteins and drugs in biomedical text, and the names of vehicles or weapons in defence systems. A second stage in a typical IE system will construct an information template by identifying and marking up particular facts relating to the entities that have been recognised (e.g. facts pertaining to employment such as employer's name, length of service, salary etc.)

The two most common methods for identifying references to entities within a document are: (i) a rule-based approach in which entity recognition grammars are written; and (ii) a machine

learning approach, which includes the deployment and training of statistical taggers. In both cases, *test data* is used to evaluate how well a system identifies and classifies named entities. Both training data for machine learning and test data are human-annotated.

3.2 Improved Search

One of our aims in identifying and classifying named entities in qualitative data collections, is to provide a framework for more precise and efficient web indexing and search. Full IE can be viewed as a means to annotate documents with semantic metadata, creating a machine-readable semantics for use in fully automated reasoning (Berners-Lee *et al*., 2001) or highly sophisticated browsing and search. We use a range of XML-based language processing tools (Thompson et. al., 1997; Grover et. al., 2000; Grover and Tobin, 2006). that are exploited to reduce the manual efforts that are typically required to create marked-up data with shallow semantic information (including entities such as person names, company names, place names, and temporal information).

For the vision of the semantic web to be possible, computerised systems need to be able to identify references to entities in documents (Dill *et al.,* 2003). This is the first step towards building relations between entities and between documents. Annotation tools allow users to manually annotate documents with semantic information and in this work, we developed a solution in which the NITE XML Toolkit (Carletta *et al*., 2003) was integrated with our IE tools in order to semi-automate the process of annotating important entities in social science documents. This relieves some of the burden on researchers in moving towards fully exploiting the semantic content found in their qualitative data.

In this effort, we concentrate on five types of entities that we believe are broadly useful within the social science domain. These include the names of people, organisations, locations, occupations and dates. The names of entities are primarily identified in order to: (i) link between entities within and between documents and (ii) anonymise the names (as described next). A social science researcher can highlight entities that can be used to improve search, or to anonymise them if they are considered confidential. The names of occupations can also be anonymised, however they are a useful feature in searching for studies performed with people of particular occupations.

3.3 Automated Anonymisation

Our tools provide another potentially useful extension of IE research. While the law is complex, data providers/publishers must ensure that data are exposed in ways that respect ethical and legal considerations (Corti *et. al.*, 2000). Anonymisation techniques are a key option for protecting data to maintain respondents' confidentiality. Effective editing of data can involve using pseudonyms, abstract systems of coding or simply the crude removal of text. However, researchers must be aware of the potential for distortion. For example, deleting all identifiers is a simple but blunt tool that creates data that is confidential but also unusable.

Manual anonymisation is time-consuming and labour-intensive. Providing user-friendly tools to semi-automate this process when preparing data for archives would be extremely beneficial in increasing the flow of web-enabled data. Currently in the international realm only a few new projects are utilising IE tools in this context (Guo *et. al.*, 2006; Poesio *et. al.*, 2006). As IE and named entity recognition techniques become more mature, the opportunities for automatic anonymisation are greatly enhanced as many cases can be covered by, for instance, replacing names with dummy forms. However, just identifying named entities is not enough; many texts include co-references (as when "John Smith" is later referred to as "Mr Smith"), and true anonymisation should consider this.

4 Annotation and Evaluation

In evaluating the performance of our IE tools, we follow standard practice, by comparing our system output to that of a gold-standard test set. In this section, we outline how we created our gold-standard and calculate our inter-annotator agreement scores. We then compare our system's performance to the gold-standard. We perform these comparisons by computing the *f-score*, which is the harmonic mean of *precision* and *recall*.

4.1 Creating the Gold-Standard

Seven collections of interview transcripts were annotated with the names of people, locations, organisations, occupations and dates. The corpus contains thirty documents with an average of 12,019 words per document. Eleven of the documents were annotated by two people in order to calculate inter-annotator agreement scores. These scores are shown in Table 1.

Type	DAT	LOC	OCC	ORG	PER	**Total**
Precision	60.9	81.9	46.5	63.2	88.0	**64.9**
Recall	77.8	90.5	57.1	63.8	96.1	**78.9**
F-score	**68.3**	**86.0**	**51.3**	**63.5**	**91.8**	**71.2**

Table 1: Inter-Annotator Agreement Scores

The agreement on person names and locations were as high as one might expect. However, the guidelines for annotating the other entity types were obviously not as clear-cut and, in particular, occupations are significantly lower than any other entity type. The annotators were asked only to mark up specific references to job titles held by people in the study, and not generic descriptions of occupations. This is done because specific references are more useful in search and are also more likely to need to be anonymised. However, specific and generic cases are not always easy to distinguish. Table 2 contains some examples from our corpus that contain occupations. The first is a specific reference to the current occupation of the interview subject. The other examples were highlighted by only one of the annotators and, on first inspection, they may seem to be generic. However they are valid jobs the interviewee has held in the past which would need to be highlighted for search or anonymisation purposes.

(1)	"My job title is **Public Affairs Correspondent**..."
(2)	"I used to be a **health visitor** years ago..."
(3)	"I was also a **carer**..."
(4)	"after having about 9 months as a **general news reporter**..."

Table 2: Occupation Examples

4.2 Named Entity Recognition Evaluation

In this section, we compare our system's performance on recognising the names of entities in social science data. Our system uses the LT-XML2 and LT-TTT2 tools to preprocess the data (tokenisation, sentence boundary identification, part-of-speech tagging, lemmatisation and chunking) and uses Curran and Clark's (2003) maximum entropy tagger for named entity recognition. The statistical models for NER were trained on data from the MUC (Chinchor, 1998) and ACE-05[2] challenges using the tagger's built-in features. Table 3 shows f-scores for all named entity types.

[2] The ACE (Automatic Content Extraction) website: http://www.nist.gov/speech/tests/ace/index.htm

Type	DAT	LOC	OCC	ORG	PER	**Total**
Precision	65.0	55.5	63.7	41.1	59.3	**55.4**
Recall	51.1	56.1	78.1	54.2	82.7	**65.0**
F-score	**57.2**	**55.8**	**70.2**	**46.8**	**69.0**	**59.9**

Table 3: System Output Scores

In general, both precision and recall are lower than we would hope to achieve because of differences between the training and test sets. The names of entities are primarily extracted using NER models trained on newswire data. However, documents within the social science context are entirely different in both content and presentation to the annotated documents available for training machine learning models. A fundamental difference concerns the density of entity occurrences in the two data types: the ratio of words to entity mentions is significantly lower for the SQUAD data at 71:1 as compared to the ACE-derived data we used for training which has a ratio of 14:1. This represents a 80% difference in NER density.

The context of mention is very different compared to the newswire articles used for training our models. Hence, much of the gap in our results is due to a failure to properly disambiguate entity types. As shown in Table 4, if we treat the location, organisation and person entities as the same type, our F-score is significantly higher, and close to the combined F-score for these entity types in the inter-annotator agreement (80.4). Hence, our system is detecting entity mentions relatively well, but the disambiguation of their types is clearly difficult. We believe this is due to the difference between the sentence structures found in social science data compared to the training material.

Type	LOC/ORG/PER
Precision	74.1
Recall	80.3
F-score	**77.1**

Table 4: System Bare Output Scores (LOC/PER/ORG Combined)

The presentation style of social science documents is entirely different from newswire text. In particular, due to space restrictions, newswire articles are content-rich and have a high term density, whereas having been transcribed from speech, social science interview data has low content density and contains all the nuances of speech including pauses, interruptions, turn-taking idiosyncrasies such as correction, and so on.

Similarly, the occupations found in news texts are entirely different from those found in social science documents. Most occupations in the ACE corpus are US government positions rather than the cleaners and health workers found in our corpus. In addition, occupations are generally expressed as modifiers in news texts e.g. “Attorney General John Ashcroft”. In our social science data, they are expressed using less dense constructions, similar to the examples in Table 2. We therefore constructed a gazetteer of occupations from different sources, and used this as a look-up list to identify occupations. This is not ideal, since many occupations in our data are described using imprecise language (“she get me into carpentry, see”).

Moving from newswire articles to social science data is thus a substantial digression both in terms of content and in terms of writing style. Further, newswire articles tend to be composed of well-written sentences, whereas social science data consists primarily of transcripts of interviews, which have an entirely different style.

5 Conclusions and Future Work

In this paper, we have demonstrated that the application of existing computational linguistics methodologies and tools to real social science data can result in a useful contribution to interdisciplinary collaborative practice. Further, we have presented some preliminary findings that indicate that information extraction can be a useful innovation in this domain, particularly in identifying the entities within data that might be useful for either search or anonymisation. Our initial results are preliminary and there is certainly scope for improvement. Applying NER models to data from a different domain produces less-than-optimal results. In the future, we aim to develop rule-based techniques that are specific to the social science domain and/or annotate further data that can be used to train NER models that are specific for this purpose.

There are a number of interesting future directions that could be explored based on this work. Automatically identifying the relationships between documents on the basis of their entities is a particularly interesting area. This is important for automatic cross-document anonymisation in order to ensure that common pseudonyms are used for the same entities across documents. Moving on from named entity recognition to activities even closer to the ultimate goals of the Semantic Web initiative, such as key word extraction based on chosen ontologies or folksonomies, is also a direction the UKDA would like to explore. This can be employed to reduce a piece of text to its key attributes and might prove useful for efficient automatic comparison of documents for a qualitative researcher (Milosavljevic 2003).

References

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web, *Scientific American*, 284(5), 28--37.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J. and Voormann, H. (2003). The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, Special Issue on Measuring Behavior, 35(3).

Chinchor, N.A. (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Corti, L., (2000), Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research - The International Picture of an Emerging Culture, *Forum Qualitative Social Research* [Online Journal], 1(3).

Corti, L., Day, A. and Backhouse, G. (2000). Confidentiality and Informed Consent: Issues for consideration in the preservation of and provision of access to qualitative data archives, *Forum Qualitative Social Research* [On-line Journal], *1*(3).

Corti, L. and Thompson, P. (2004). Secondary Analysis of Archive Data. In C. Searle *et al*. (eds.), *Qualitative Research Practice*, London: Sage Publications.

Curran, J. R., Clark S. (2003). Language Independent NER Using a Maximum Entropy Tagger. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-03)*, 164-167.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, K., Rajagopalan, S., Tomkins, A., Tomlin, J.A. and Zien, J.Y. (2002). SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. *12th International Conference on World Wide Web*.

Fielding, N. (2003). *Qualitative Research and E-Social Science: appraising the potential*. University of Surrey.

Grover C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT - a flexible tokenisation tool. In *Proceedings of LREC-2000*.

Grover, C. and Tobin, R. (2006). Rule-Based Chunking and Reusability. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G. and Hepple, M. (2006). Identifying Personal Health Information Using Support Vector Machines . In *Proceedings of the AMIA 2006 Workshop on Challenges in Natural Language Processing for Clinical Data*. Washington. Nov 2006.

Milosavljevic, M. (2003). Defining Comparison. In P. Slezak, (Ed.), *Proceedings of the Joint International Conference on Cognitive Science with the Australasian Society for Cognitive Science*, Sydney: University of New South Wales.

Poesio, M., Kabadjov, M.A., Goux, P., Corti, L. and Bishop, E. (2006). An Anonymization Module Based on Anaphora Resolution. In *Proceedings of LREC'2006*, Genoa, Italy, May 2006.

Thompson, H., Tobin, R., McKelvie, D. and Brew, C. (1997). *LT XML - software API and toolkit for XML processing*. Downloadable from http://www.ltg.ed.ac.uk/software/.