Intelligent Exploitation of Local Government Resources

Mike Rosner, Andrew Attard
Faculty of ICT, Dept. Intelligent Computer Systems
Univerisity of Malta, Msida MSD2080 MAlta
E-mail: mike.rosner@um.edu.mt, andrew.attard@um.edu.mt

Abstract

Malta is divided into sixty-eight local councils each contributing to the most basic form of local government. Several meetings take place during which the councillors gather to discuss the maintenance and embellishment of the locality, each of which are noted down in Maltese. This paper concerns a corpus of local government documents. We suggest an approach to the problem of developing an intelligent browsing system that offers improved access to the information, for example to assist local councils in decision making, or to give members of the public more transparent way to browse local council documentation.

Keywords: Govenment corpus; intelligent catalogue system

Acknowledgement: The development of this paper has been partially supported by the Seventh Framework Programme and the ICT Policy Support Programme of the Euro- pean Commission under contract METANET4U (Grant Agreement 270893). Authors also gratefully acknowledge the contribution of the Valletta Local Council.

1 Introduction

The Freedom of Information Act 2008 is a milestone in so far as citizens' rights are concerned: it enables the public, directly or indirectly (through investigative journalists), to disclose that information which the public authorities have not rendered public. In this respect, a freedom of information law brings with it more transparency on the working of the public administration rendering it more accountable to different audiences, including not just the public at large, but also to other, more specialized levels of analysis. But although, in principle, governmental resources are freely available, in practice they are not that easy to obtain. Nor can they be browsed over the internet, be queried, or subjected to automated annotation techniques.

This paper is presented as a result of the unexpected availability of a archive of language data originating from various Maltese Local Council. Malta has 68 Local Councils – 54 in Malta and 14 in Gozo. Local Councils are regulated by the the Local Councils Act, modelled on the European Charter of Local Self-Government (Council of Europe), according to which a Local Council "shall be a statutory local government authority having a distinct legal personality and capable of entering into contracts, of suing and being sued, and of doing all such things and entering into such transactions as are incidental or conducive to the exercise and performance of its functions as are allowed under the Act."

As a consequence of this status, the documents flowing through a Local Council are quite diverse in several different senses, namely:

- **Genre.** The data consists of minutes, memos and data. Each of tese have their own special style.
- **Subject Matter.** The data indicates a wide range of topics ranging from road repairs to social services.
- **Language.** The data is mostly monolingual but is in two languages - both English and Maltese. Not all Maltese text is correctly written (e.g. omitting use of the normal characters instead of Maltese ones)
- **File Formats.** Files do not have uniform structure. That is, there is no overall priniciple of organisation. Particular kinds of document are not written in a uniform style. Minutes, for example, vary according to author. File formats are predominantly word and excel.

1.1 Motivation

A superficial perusal of the dataset suggests that official publications are important in the sense of offering an immense range of opportunities that could benefit three categories of usage:

- Research and development in language technology. The data contained provides examples of Maltese officialese. This is of interest from the stylistic and lexical point of view and could in principle exploited by various writing aids
- Research in the Humanities and Social Sciences. As we shall see, the files contain information that would be useful for the purposes of historical or sociological research concerning the different localities.
- Decision Support within Local Council. In Malta the Local Council system is such that information contained in archives has a tendency to be forgotten. Often, effort that has been expended discussing a particular problem is repeated when, years later, the same problem is rediscussed by a new set of Council members. Another issue is that the same problems tend to crop up in different Local Councils. Solutions that have been developed in one Council ought in principle to be available for scrutiny by another. Under the present arrangements, it is difficult to achieve this level of transparency. Although not explicitly addressed in this paper, we would like to see the introduction of a browsing system for official documents which would improve access to content within the Council system.

The remainder of this document explores the possibilities for improving access to these materials and is structured as follows. Section 2 gives a rough indication of the corpus content. Section 3 outlines the achievements so far and challenges ahead and section 4 provides a set of objectives. We conclude in section 5 .

2 Corpus Content

The corpus comprises around 6,753 files, containing 13 different types of file formats (including a type for the files saved in an undefined format). The size of the dataset is 1.15GB. We have not yet determined the number of words it contains.

This collection is made up of different governmental resources, organized in two main sections: Minutes and Memos. The next two sections give an overview of their content.

2.1 Minutes

The minutes collection, forming 93% of the obtained corpus, is further organised by year, covering the years from 2007 till 2010. Each year is then categorised by locality, thus having 68 sub-collections (with the exception of the year 2007-2008, containing only data of 59 localities) of information. Each sub-collection (representing a local-council) contains a number of:

- Word documents – listing all the events that took place during the meeting.
- Excel documents – listing all the financial data, during a specific period of the year.

The information embedded in the word documents, includes maintenance issues, upcoming event details, obstacles that the locality is facing, and more. Additionally, these documents also contain information about how such obstacles are overcome, maybe also based on previous solutions to similar problems.

After briefly evaluating the collection, we noted that there seems to be no uniform structure amongst the localities, resulting with a different document structure for each local council. Having said this, each local council seems to retain the same document structure over the years.

A more troublesome observation concerns the inconsistent usage of Maltese characters. Not all documents are written using the correct Maltese characters. Furthermore, many documents are mixed in the sense that they may also contain English text embedded with Maltese.

2.2 Memos

The memos collection is much smaller than the minutes, adding up to 7% of the whole corpus. The collection covers the years from 2008 up to 2011.

However, in contrast to the minutes' collection, the files are only categorized by year, holding different Governmental memoranda which were made available during that particular year.

3 Aims and Objectives

Document collections of this kind are probably extremely common, but at the moment only rarely

accessed. Even those who are allowed to access them have difficulty finding the information that they contatin. Our primary aim is therefore to provide progressively more sophisticated access to the information contained in the collection by adopting appropriate technical means. We propose the tackle the problem bottom up: from the basic data, through the contents, towards a coherent structure for the collection as a whole.

This aim leads us to the following objectives (in rough order of difficulty)

- Automated Data Normalization: the need to employ standard representations for text and tabular data and to employ automated methods to translate the sources into such representations. This process is not so very different to those employed for the preparation of other corpora, and we intend to reuse machinery that has already been employed in the development of the MLRS corpus (Borg-et-al, 2011) for this purpose, as reported further in section 4.1.
- Automated Data Analysis: the need to extract meaningful data from the collection. Techniques that are clearly relevant include named entity recognition and topic analysis although it is unclear at this stage how well currently available systems will cope with a collection of this kind.
- Automatic Data Organization: access to data would be greatly facilitated if there were some standard methods for structuring the data automatically. The variety of methods actually employed is bewildering, requiring special (i.e. manual) procedures to access information for practically every combination of locality, topic and document type. So a key issue is whether it is possible to devise a suitable classification scheme for bureaucratic documents. Any such scheme has to strike a delicate balance between generality (being able to accommodate a very wide range of document subject matter) and specificity (implementing principles of organization that will actually make a difference to the the retrieval of useful information.

4 Achievements and Challenges

4.1 Data Normalisation

Given that the corpus is composed of documents in different 13 file formats, there are a number of challenges as regards normalisation. Table 1 shows the distribution of file types. Luckily, the bulk of the corpus consists of pdf files (37%), doc files (34%) and xls files (22%) and these have been successfully converted, using mostly automated techniques, into formats that can be further manipulated: doc and pdf files were converted to txt files, while xls files were converted to csv files as shown in the right hand columns of the table.

Source Type	No. files	Target Type	No. Files
pdf	2559	txt	2101
doc	2308	txt	2295
xls	1520	csv	1275
jpg	196		
.rtf	43		
bmp	12		
xlsx	10		
unknown	9		
gif	5		
docx	4	txt	4
tif	3		
zip	2		
htm	2		

Table 1: structure of collection by file type

A challenge for the text files is that most of them do not use the standard set of Maltese characters, and we are experimenting with automatic spelling correction to overcome this problem.

4.2 Data Analysis

The information residing in these files is clearly valuable but not easily accessible. We are proposing to exploit existing well-established information extraction techniques (see Cunningham-et-al 2005) involving for instance named entity recognition and topic analysis to identify key elements of well known document types and to build gazetteers that include the names of people, organisations, places and quantities. For example, our collection contains a large number of meeting minutes and within documents of this type we would propose to identify

- Which councillors attended the meeting
- The agenda proposed, and decided upon for the meeting
- The different topics, or issues which were to be covered during the meeting
- Other relevant information (e.g. decisions reached) in concordance with the meeting.

We would expect to find other key properties for other kinds of documents.

4.3 Data Organisation

The data is currently organised in a very rudimentary way. Furthermore there are inconsistencies in the way data has been organised by the several Local Councils that have contributed to the collection. We are therefore proposing a kind of intelligent cataloguing system based on sound principles of organisation. We believe that there are certain similarities between on the one hand, the problem of structuring document collections of the kind described and on the other, the organisation of repositories for linguistic resources in general. Techniques which apply to the second problem might fruitfully be applied to first one. Consequently, we will base our solution to the organisation of data around three major components:

- Definition of different document types together with their respective structuring principles. Here we envisage to approach the problem of structuring bureaucratic documents by developing a system of metadata categories not entirely dissimilar to the system already developed for the description of linguistic resources with the METANET4U project (METANET4U deliverable D4.1). This system would then serve as a skeleton into which the actual documents could be fitted.
- A system for intelligently mapping document resources into the document catalogue.
- A system for browsing the contents of the collection according to different principles of organisation such as locality, date, topic.

This system might also have the potential to offer a suggestion facility whereby solutions to common problems and frequently asked questions might be pooled in order to improve local decision making

5 Conclusion

The starting point for this paper was a document collection of a kind which is extremely common, extremely diverse and whose contents could be better exploited. We have described an approach which has the potential to transform a passive document collection only accessible to a few into an information rich resource available to many through the use of mainly existing technologies.

6 References

Borg, C., Fabri, R., Gatt, A., Rosner, M., Maltese and the Digital Age: Developing Electronic Language Resources for the Maltese Language, Linguistics Circle Presentation, University of Malta, 2011

Cunningham, H. Information Extraction, Automatic, Encyclopaedia of Language and Linguistics, Elsevier, 2005