Intelligent Exploitation of Local Government Resources

Mike Rosner, Andrew Attard
University of Malta

LREC Workshop on Exploring and Exploiting Official Publications
Istanbul, May 2012
Background

- The Maltese Freedom of Information Act 2008 is a milestone in so far as citizens rights are concerned:
  - Enables disclosure of certain information which the public authorities might not have rendered public
  - Potential for more transparency on the workings of the public administration
- Although, in principle, governmental resources are legally available, in practice they are not that easy to obtain.
  - Cannot be browsed over the internet
  - Cannot be queried
  - Cannot be subjected to automated IE or annotation techniques
- This paper is presented as a result of the unexpected availability of a archive of language data originating from various Maltese Local Councils that is deemed to be public information.
Outline

1. Background
2. Contents and Structure of Corpus
3. Motivation
   - Aims
   - Objectives
4. Roadmap
   - Achievements
   - Future Work
5. Conclusion
Corpus Content

- Size: 6,753 files, 1.15GB
- Genre: government docs of different types having different styles.
  - Minutes
  - Memos
  - Data
- Subject matter: a very wide range of topics ranging from road repairs to social services.
- Language: data is mostly Maltese but English is also used. Generally inconsistent usage of Maltese characters.
- File formats: 13 different formats but predominantly MS Word and Excel files.
Minutes

- Comprises 93% of the corpus
- Organised by year (2007-2010) and further subdivided by locality
- Each subcollection (i.e. per locality) comprises
- Word Documents:
  - Actual minutes and accompanying documents concerning upcoming event details, maintenance issues, problems and solutions
  - Each Council adopts a consistent document structure
  - Not all Councils adopt the same document structure
- Excel Documents: listing financial data during specific periods of the year
Memos

- The memos collection is much smaller (7% of the whole corpus).
- Covering 2008 up to 2011
- Holding different Governmental memoranda which were made available during that particular year
Hypothesis

We feel that data of this kind is potentially of immense value in the following areas:

- LT Research and Development
- Research in the Humanities and Social Sciences
- Decision support within Local Councils. In particular, ability to profit from the fact that similar problems and issues tend to crop up in the different councils.
Data collections such as these are probably extremely common
Rarely accessed
Poor retrieval facilities

Our primary aim is therefore to provide progressively more sophisticated access to the information contained in the collection by exploiting existing technologies.

We propose a bottom up solution: from the basic data, through the contents, towards a coherent structure for the collection as a whole.
Objectives

1. **Automated Data Normalisation**: adoption of automated means to translated sources into standard representations for text and tabular representations.

2. **Automated Data Analysis**: the need to extract meaningful data from the collection using information extraction.
   - Topic analysis
   - Named Entity recognition

3. **Automated Data Organisation**: a representation scheme for bureaucratic documents. Balance between
   - generality (being able to accommodate a very wide range of document subject matter) and
   - specificity (document-type specific principles of organization that will improve retrieval of useful information).
Achievements - Data Normalisation

- The corpus is composed of documents in different 13 file formats
- 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 $\Rightarrow$ text
  - xls $\Rightarrow$ CSV
- A challenge for text files is the inconsistent usage of Maltese characters
Next Steps: Data Analysis

- Focus on meeting minutes
  - Automated identification of
    - 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)
Next Steps: Data Organisation

- Intelligent cataloguing system
- Exploit potential similarities between catalogues for bureaucratic document types and linguistic resources
- Solution based on
  - Definition of different document types using metadata categories.
  - A system for efficiently 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.
Conclusion

- The starting point for this paper was a collection of official documents whose content differs from e.g. pure legal documentation.
- Intuitively, is valuable and could serve different purposes: LT, HSS, decision support.
- We have described an approach aimed at transforming a passive document collection accessible to a few into an information rich resource available to many through the use of mainly existing technologies.
- However the purpose for which data could be used is unclear. This situation very common in LT for Humanities.
- It is therefore crucial to identify suitable user communities in order to cater for their needs and develop a system that is credible.