Large Scale Semantic Annotation, Indexing and Search at The National Archives

Diana Maynard and Mark A. Greenwood
University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
initial.surname@dcs.shef.ac.uk

4th October 2011

1 Introduction

This paper describes some of the work carried out as part of the UK Government Web Archive Project[1], which aims to help open up TNA's records of government websites (going back to 1997 and comprising some 700 million pages). Government funding has been allocated to publishing more and more material on government websites in open and accessible forms, but in many cases it is still very hard to find the information needed, because the search tools available only enable a primitive keyword-based search. Sophisticated and complex semantics can transform this archive, but our mission is to show how simple and straightforward mechanisms can add value and increase usage in the short and medium terms.

Our system, TNA-Search, thus aims to improve access to the enormous volume of data at TNA, both for the general public and for specialist researchers, by enabling a semantic-based search for categories of things, e.g. all Cabinet Ministers, all cities in the UK. Search results can include morphological variants and synonyms of search terms, specific phrases with some unknowns (e.g. an instance of a person and a monetary amount in the same sentence), ranges (e.g. all monetary amounts greater than a million pounds), restrictions to certain date periods, domains etc., and any combination of these. The search functionality allows queries that arbitrarily mix any combination of full-text, structural, linguistic and semantic queries, and can scale to gigabytes of text. Our experience is that faceted and conceptual search over spaces such as concept hierarchies, specialist terminologies, geography or time can substantially increase the access routes into textual data and increase usage accordingly.

2 System Architecture

TNA-Search aims to import, store and index structured data relevant for the web archive in a scalable semantic repository, in an easy to manipulate form, using linked data principles and in the range of tens of billions of facts. Links are

[1] http://www.nationalarchives.gov.uk/webarchive/

made from the web archive documents into the structured data, over hundreds of millions of documents and terabytes of plain text. The system allows browsing, search and navigation from the document space into the structured data space via semantic annotation and vice versa via a SPARQL endpoint, both as full text and as linguistic annotation structures. For example, it enables complex queries by different kinds of users, ranging from the medical student who wants to find all medical publications relating to Type 2 diabetes, to the public user who wants to search on a specific minister, e.g. for information about Gordon Brown's career in government.

TNA-Search is built on GATE [1] for semantic annotation and GATE Mímir [2] for indexing and search, and relies on a huge semantic repository combining FactForge and the SKB (Semantic Knowledge Base) ontology. Factforge[2] is a knowledge base developed by Ontotext[3] containing over 2.2 billion statements and containing datasets from DBpedia, Freebase, Geonames, UMBEL, WordNet, CIA World Factbook, Lingvoj and MusicBrainz. The SKB ontology was developed by Ontotext specifically for TNA-Search, based on the Central Government Ontology (CGO), and comprises official titles in the UK government (e.g. "Secretary of State for Health"), 8138 names of officials, and names of unambiguous UK government organizations. The CGO consists of a class hierarchy which describes governmental organizations, governmental roles, and concepts around the functioning of the government. The class hierarchy is supplied with 33 properties describing relationships between the government organizations and roles and the people who hold the roles, e.g. memberOfCabinet, or between different kinds of government organizations, e.g. hasCabinet. The knowledge base connects disparate elements in the UK Government Web Archive to create a more 'joined up' experience for the user, showing associations, linking between sites and allowing the archive to be segmented in different ways (e.g. everything related to DEFRA).

The basic methodology comprises the following set of steps:

1. Annotate the documents using GATE Embedded: this uses ontology-based information extraction (OBIE) tools to annotate entities in text and relate them to the ontology where appropriate.

2. Index the documents using GATE Mímir: this consists of creating an index based on the annotations produced in Step 1.

3. Search the documents using GATE Mímir: this entails querying the GATE Mímir index using full text, SPARQL or annotation-based queries.

4. Browse the results: the search results point back to the texts in the original archive.

Figure 1 depicts the overall architecture of TNA-Search. The source data is a set of WARC files representing the Government Web Archive. Semantic annotation is performed by GATE Embedded, with respect to the large semantic repository comprising the SKB and factual knowledge from Factforge. Once the GATE Mímir index has been created, any one of the three front-ends can be used to query the data (in this paper, we concentrate only on GATE Mímir).

[2] http://factforge.net
[3] http://www.ontotext.com

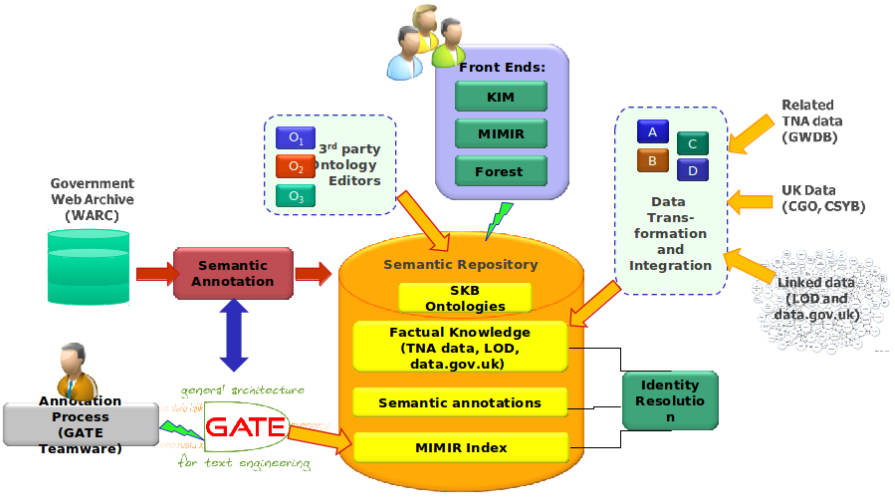

Figure 1: System Architecture

3 GATE Application

The GATE application consists of a set of processing resources (PRs) executed sequentially in a conditional pipeline over a corpus of documents. The conditional pipeline enables us to run some PRs only if certain conditions about the document are true: this is particularly useful when dealing with a heterogeneous dataset such as the government archives. The pipeline consists of 5 main parts:

- Linguistic pre-processing
- Gazetteer lookup
- Rule-based Annotation
- Semantic Annotation
- Co-reference
- Final output creation

The full paper will contain more details about these components.

4 Annotation Evaluation

We have performed two main evaluations on the annotation system: one for the first (prototype) system and one for the final (production) system. More details will be provided in the full paper, but the production system achieved results of 81% Precision, 85% Recall and 83% F-measure on a test set of documents selected from the archive, which represent a variety of document types and domains. Ongoing work is focusing on ways to improve these results further.

5 Indexing and Searching with GATE Mímir

In order to query the archive, we create a GATE Mímir index from the annotations produced. GATE Mímir [2] is a multi-paradigm information management index and repository which can be used to index and search over text, annotations, semantic schemas (ontologies), and semantic metadata (instance data). The multi-paradigm aspect refers to the accessing and linking together of multiple information sources, such as the textual content of the documents, the semantic metadata associated with the documents, and data in knowlege bases (such as the linked data cloud). Documents are annotated and indexed in GATE Mímir on the fly.

Accessing the data from a knowledge base allows GATE Mímir to understand generalisations, making it capable of answering more complex information needs, such as identifying documents that refer to "capital cities in Western Europe", without the documents explicitly mentioning this fact. At the same time, the semantics associated with the indexed documents make sure that references to any of the many places called London (other than the one in the UK) are not seen as relevant results to such a query. Additionally, support for sophisticated text-based queries allows GATE Mímir to filter the set of results to, for example, only those mentions that occur in the same sentence with the name of a particular government agency. Finally, GATE Mímir can use the semantic annotation of documents to perform some simple reasoning over the meaning of entities. For example, monetary amounts can be normalised according to the date of the document (to account for inflation), while measurements using different unit systems can be matched against each other.

6 Conclusions

In this paper, we have shown how the use of simple and straightforward mechanisms can add value and increase usage in the short and medium terms to a huge archive of information. Our solution, based on GATE (text mining), OWLIM (semantic repository) and GATE Mímir (multiparadigm search front-end components), provides search paradigms over semantic annotation that relates archival content to Linked Data and other structured sources. We import Linked Data into the semantic repository which provides a SPARQL endpoint (and also fulltext and annotation structure indices), and annotate the archive text relative to the repository. The knowledge base, which cross-references the UK Government Web Archive with a number of government-specific and more general ontologies, will continue to be progressively enhanced and developed, as will the complementary tools to mine both explicit knowledge within the archive and implicit knowledge which can be inferred from it. Text mining enables the extraction of facts from content and thus better interfaces, while the integration of the Semantic Knowledge Base makes it possible to combine a massive body of external knowledge with facts extracted from texts.

References

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for

Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[2] Hamish Cunningham, Valentin Tablan, Ian Roberts, Mark A. Greenwood, and Niraj Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*. Springer, 2011.