WhatTheySaid: Enriching UK Parliament Debates with Semantic Web

Yunjia Li, Chaohai Ding, and Mike Wald

School of Electronics and Computer Science,
University of Southampton, UK
{yl2,cd8e10,mw}@ecs.soton.ac.uk

Abstract. To improve the transparency of politics, the UK Parliament Debate archives have been published online for a long time. However there is still a lack of efficient way to deeply analysis the debate data. WhatTheySaid is an initiative to solve this problem by applying natural language processing and semantic Web technologies to enrich UK Parliament Debate archives and publish them as linked data. It also provides various data visualisations for users to compare debates over years.

Keywords: linked data, parliamentary debate, semantic web

1 Introduction

The publicity of UK Parliament Debate, such as BBC Parliament[1] have exert tremendous influence on the transparency of politics in the UK. Political figures need to be responsible for what they have said in the debates as they are monitored by the public. However, it is still difficult currently to automatically analyse the debate archives to find the answer to questions such as: how the debates across months or even years are related to each other. For this purpose, we have developed WhatTheySaid[2] (WTS), which uses semantic Web and natural language processing (NLP) technologies to automatically enrich the UK Parliament debates and categorize them for searching, visualisation and comparison.

In UK, there are already applications, such as TheyWorkForYou[3], to provide extended functions for users to search debates and view the performances of each Member of Parliament (MP), such as the voting history and recent appearances. The semantic Web approach is also applied in Polimedia [1] as a way to model Dutch Parliament debates and enrich them with named entities and external links to different media. In this demo, we refer to the data sources and the methodologies provided by those previous work and build more advanced features to fullfil the following requirements: (R1) Calculate the similarities between debates so that users can easily navigate through similar debates; (R2) Categorise debates into different topics and extract the key statements, so that

[1] http://www.bbc.co.uk/democracylive/bbc_parliament/
[2] http://whattheysaid.org.uk
[3] http://theyworkforyou.com

users can easily spot the statements that are contradict to each other; (R3) Based on R2, link the debates to a fragment of debate video archive, so that users can watch the video fragment as the proof of the statement; (R4) Analyse the speeches of a particular MP and see how the sentiment is changing over time.

To demo the implementation of the requirements above, we have taken the UK House of Common debate data in 2013 from TheyWorkForYou as the sample dataset, and the following sections will go through the system.

2 Semantic Model of UK Parliament Debate

The WTS ontology[4] models UK Parliament debate structure and involved agents. This ontology reuses some vocabularies such as FOAF[5] and Ontology for Media Resource[6]. When designing this ontology, we have firstly referred to the data structure of TheyWorkForYou, where one debate is identified by a Heading and a Heading contains one or more Speeches. We have also added several attributes to Speech, such as sentimental score, primary topic, summarise text and related media fragment in order to save the data required to implement R2, R3 and R4 in Section 1.

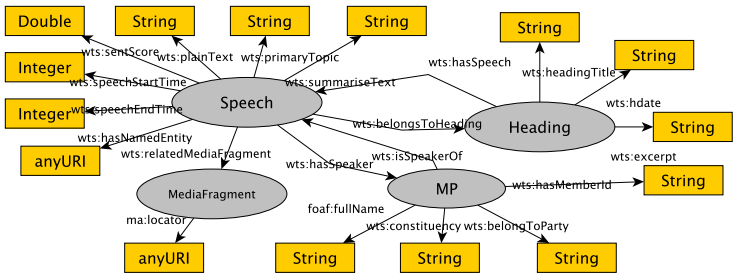

Fig. 1. WhatTheySaid Ontology

3 System Design and Walk-through

Figure 2 shows the architecture of WTS application. Our major data sources are the debate information from TheyWorkForYou, including debate date, speakers, headings, the text of speeches in each debate, etc., and the debate video with automatic transcripts provided by BBC Parliament archive. Then we use AlchemyAPI[7] to proceed sentimental analysis on each speech in the debates so that

[4] http://www.whattheysaid.org.uk/ontology/v1/whatheysaid.owl
[5] http://www.foaf-project.org
[6] http://www.w3.org/TR/mediaont-10/
[7] http://www.alchemyapi.com/

each speech made by a speaker will be allocated with a score between 1.0 (positive) and -1.0 (negative). For speeches with more than 1000 characters, we also carry out topic detection and text summarisation using AlchemyAPI.

To link the debates to each other, we apply TF-IDF [3] algorithm to calculate the similarity scores between each two debates. We firstly merge the plain text of all the speeches in a debate into one big debate document d. Then, given a debate document collection D and $d \in D$, a word w, we calculate the weighting of each document W_d:

$$W_d = f_{w,d} * log(|D|/f_{w,D}) \tag{1}$$

where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of corpus, and $f_{w,D}$ is the number of documents in which w appears in D [3]. In information retrieval, the Vector Space Model (VSM) represents each document in a collection as a point in a space and the semantically similarity of words is depended on the space distance of related points [4]. When the W_d is calculated for each document, we use cosine similarity[8] for the vector space to come up with the similarity score between any two debate documents. On the user interface, every time a debate document is viewed, we will list the top ten debates that similar to this debate, so that users can easily navigate through similar debates.

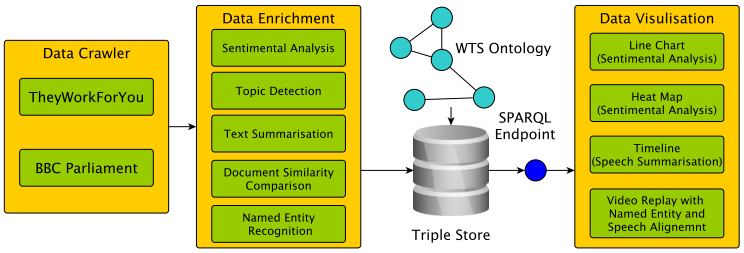

Fig. 2. WhatTheySaid Architecture Diagram

For named entity recognition, we use DBpedia Spotlight[9] to extract named entities and interlink those concepts to the speeches, where they are mentioned. All the enrichment information are saved in a triple store implemented by rdfstore-js[10], which also exposes a SPARQL Endpoint data querying and visualisation. For the whole 2013 year's debate, we have collected 68968 speeches and more than 400K named entities (with duplication) have been recognised. Using the model defined in Figure 1, we have generated more than 1.2 million triples.

[8] `http://en.wikipedia.org/wiki/Cosine_similarity`
[9] `https://github.com/dbpedia-spotlight`
[10] `https://github.com/antoniogarrote/rdfstore-js`

We visualise the enriched debate data in various ways. Firstly, we use both heat map and line chart to visualise the sentiment scores of speeches for each MP on yearly (see Figure 3(a)) and monthly basis respectively. We also provide a timeline visualisation (Figure 3(b)) for the statements in different topics made by a certain MP. To implement R3, we have referred to the previous work [2] and designed a replay page with the transcript and named entities aligned with the fragments of debate video[11]. The full demo is available online[12] and the RDF dataset is published for download[13]. We are planning to expand the application with more debates from early years, so that debates across years can be interlinked and enriched for analysis.

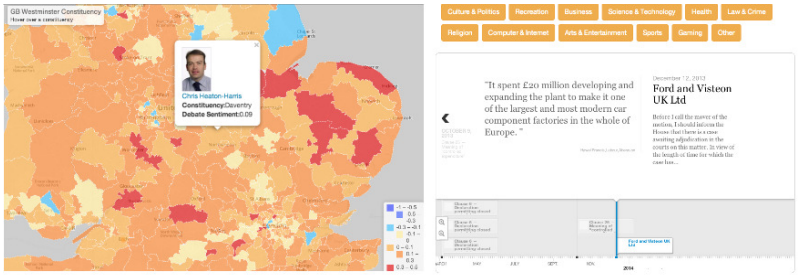

Fig. 3. WhatTheySaid Data Visualisation

4 Acknowledgement

This mini-project is funded by the EPSRC Semantic Media Network. We also would like to thank Yves Raimond from BBC and Sebastian Riedel from UCL for the support of this mini-project.

References

1. Juric, D., Hollink, L., Houben, G.J.: Bringing parliamentary debates to the semantic web. Detection, Representation, and Exploitation of Events in the Semantic Web (2012)
2. Li, Y., Rizzo, G., Troncy, R., Wald, M., Wills, G.: Creating enriched youtube media fragments with nerd using timed-text (2012)
3. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning (2003)
4. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1), 141–188 (2010)

[11] Due to copyright issues, we cannot make the debate video publicly available.
[12] `http://whattheysaid.org.uk`
[13] `http://whattheysaid.org.uk/download/wtstriple.ttl`