Semantic Media Network

a new paradigm for navigable content for the 21st Century

SemanticNews: Enriching publishing of news stories

Final Report

9th January 2014

Jonathon Hare jsh2@ecs.soton.ac.uk
David Newman drn@ecs.soton.ac.uk
University of Southampton

Wim Peters w.peters@dcs.shef.ac.uk
Mark A. Greenwood m.greenwood@dcs.shef.ac.uk
University of Sheffield

Jana Eggink Jana.Eggink@bbc.co.uk
BBC R&D

University of Southampton

The University Of Sheffield.

BBC

Executive Summary

A central goal for the EPSRC funded Semantic Media Network project is to support interesting collaboration opportunities between researchers in order to foster relationships and encourage working together (EPSRC priority 'Working Together'). **SemanticNews** was one of the four projects funded in the first round of Semantic Media Network mini-projects, and was collaboration between the Universities of Southampton and Sheffield, together with the BBC.

The SemanticNews project aimed to promote people's comprehension and assimilation of news by augmenting broadcast news discussion and debate with information from the semantic web in the form of linked open data (LOD). The project has laid the foundations for a toolkit for (semi-) automatic provision of semantic analysis and contextualization of the discussion of current events, encompassing state of the art semantic web technologies including text mining, consolidation against Linked Open Data, and advanced visualisation.

SemanticNews was bootstrapped using episodes of the BBC Question Time programme that already had transcripts and manually curated metadata, which included a list of the topical questions being debated. This information was used to create a workflow that a) extracts relevant entities using established named entity recognition techniques to identify the types of information to contextualise for a news article; b) provides associations with concepts from LOD resources; and, c) visualises the context using information derived from the LOD cloud.

This document forms the final report of the SemanticNews project, and describes in detail the processes and techniques explored for the enrichment of Question Time episodes. The final section of the report discusses how this work could be expanded in the future, and also makes a few recommendations for additional data that could be could be captured during the production process that would make the automatic generation of the contextualisation easier.

Table of Contents

Glossary

ANNIE A Nearly-New Information Extraction system.

DBPedia A crowd-sourced effort to extract structured data from Wikipedia, which can then be interrogated with sophisticated queries.

FOAF Friend-Of-A-Friend. A schema for defining people and the links between them in RDF.

GATE General Architecture for Text Engineering. An open-source platform for building workflows for almost any text-processing problem.

GeoNames A geographical database that has it's own ontology for defining locations using RDF.

IE Information Extraction. The automated extraction structured information from unstructured or only partially structured machine-readable text.

LOD Linked Open Data. Data that is both accessible to all and structured through definition of its interrelations. Commonly this data is available in RDF.

LODIE Linked Open Data Information Extraction. Enrichment of entities retrieved via information extraction with a Linked Data URI and a confidence of accuracy.

NE Named Entity. A part of written text used to identify a person, place, organization or an expression of a time, quantity, monetary value, percentage, etc.

NER Named Entity Recognition. Location and classification of Named Entities within a piece of written text.

OWL Web Ontology Language. Allows concepts and their properties to be defined and specifies how these should be described using RDF.

SIOC Semantically Interlinked Online Communities. An OWL ontology for describing content generated by users in online communities, such as forums, blogs and social networks.

SPARQL SPARQL Protocol and Resource Query Language. Widely used language for writing queries against RDF data.

TwitIE An open-source Information Extraction (IE) pipeline for microblog text such as tweets on Twitter.

YAGO A semantic knowledge base that contains structured data derived from Wikipedia, WordNet and GeoNames

1. Introduction

The goal of the SemanticNews project has been to enrich news video media produced by the BBC with contextual information in a visual way. After discussions with the BBC it was decided that the Question Time programme would provide the most helpful resource to build a case study exemplar for this goal.

Question Time has a consistent and well-structured format. It has around 4 or 5 questions per episode, each on a different subject. Each episode has manually curated metadata available in various formats from the BBC programmes website. The BBC could also provide the subtitle stream for episodes in a machine-readable format. Beyond this, user-generated content was also available in the form of tweets (from Twitter) referring to Question Time.

The subtitles and tweets although rich in information required significant work to extract this information and represent it in a Linked Open Data (LOD). Various text-processing scripts, GATE applications and SemanticNews's own OWL ontology were required to achieve this task and the way this was accomplished is described in section 2 of this report.

After representing the information in as LOD, it was then possible to make use of it to build a demo application, as described in section 3 of this report. The purpose of the demo application was to show that various visualisations could be built around the information extracted and this could be integrated together and with the original media (the video of the Question Time episode) and interact synchronously in a user-friendly way.

As the demo application developed, it became more apparent as to how further visualizations could be added to it that makes use of external LOD. One example successfully implemented and described in section 2 and 3 of this report, is the election results visualization. Other possible visualizations, improvements to the demo application and extracted LOD and potential generalization of SemanticNews's approach are described in section 4 of this report. This section also includes recommendations to the BBC on simple ways their media and metadata for programmes could be enhanced, making it easy to extract and generate a wider range of LOD from it. The ultimate aim is to make the task of providing contextual information to the user in a visual way more straightforward.

2. Technical Approach

Collating and processing the multiple data sources required to build the demo application described in section 3, involves integrated work from the technical partners. Figure 1 shows an overview of the many steps required to provide the rich and interlinked information needed by the demo application.

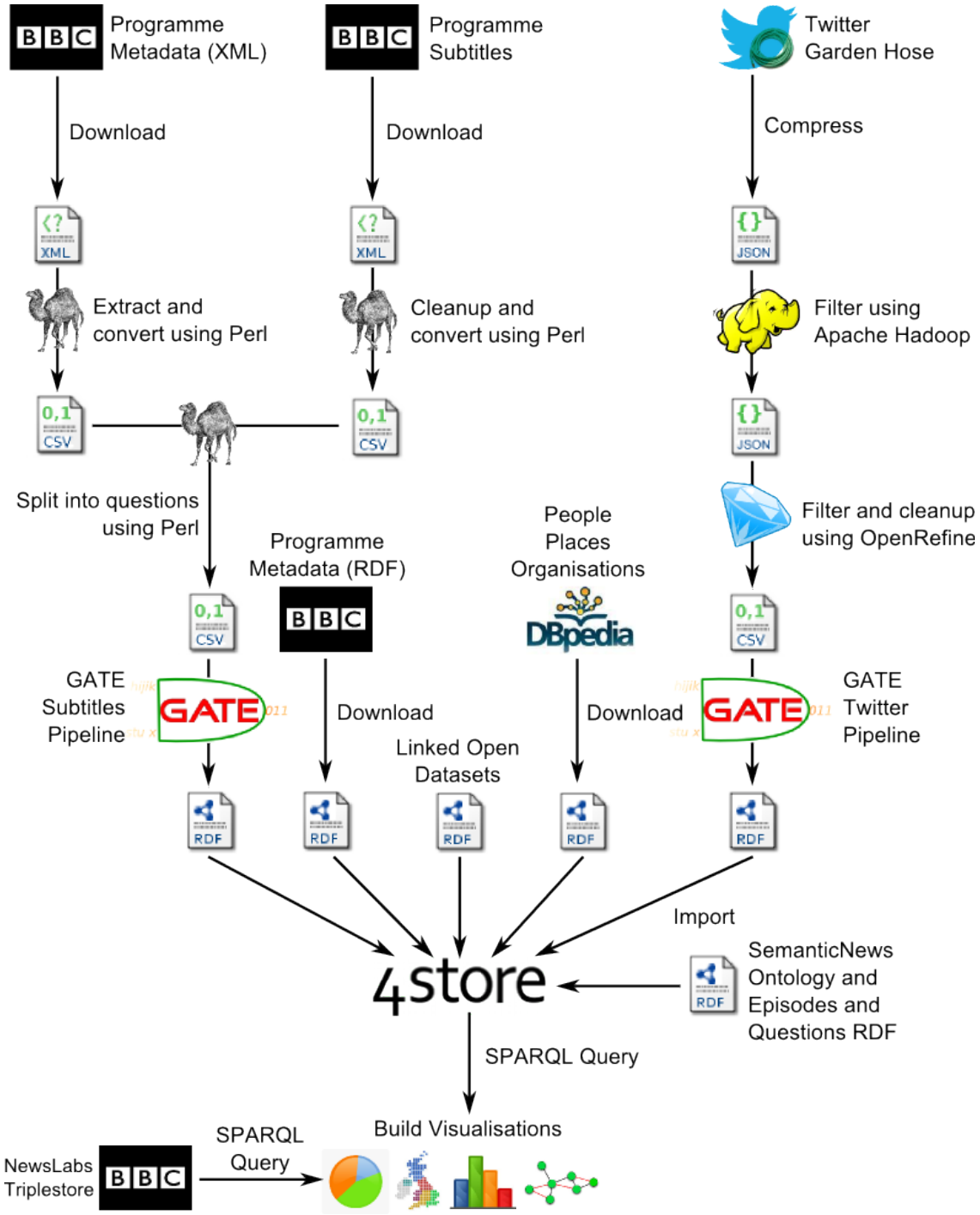

Figure 1: SemanticNews pipeline for collating and processing data

2.1. Data pre-processing

The basis of the whole SemanticNews pipeline is the identification of Named Entities (NEs). There are two main media sources from which these can be extracted:

1. Subtitles from TV programme.
2. Tweets from people whilst watching the TV programme.

Whilst both these source are rich in information, a significant of effort was required to make them suitable to be evaluated by the GATE pipeline for Named Entity Recognition (NER) and Linked Open Data (LOD) enrichment.

The subtitles provided by the BBC for a particular TV programme come as an XML file. This file is in fact broadcast as the subtitles stream along with the video and audio for the programme. This means this file contains information such as font colouring and justification for how the subtitles should appear on a viewer's TV screen. This presented a problem when trying to extract the text of the subtitles; typically, the stream would define each word individually as it would appear on screen but at the end of the sentence, the whole sentence would be justified. This meant that every subtitle would appear twice. Therefore, a script was written to remove this duplication and provide a much more simple formatting of the subtitles that just included the time offset (from the start of the programme) and the sentence that was spoken.

The time offset of a sentence within a programme is critical because each sentence needs to be designated to a particular segment of the programme and the ordering of the sentences needs to be maintained. The task of designating sentences to programme segments requires additional information about the offsets of these segments within the programme. For the originally chosen programme set, (Question Time episodes in 2010), it was possible to make use of metadata from the BBC Programmes site[1], which provided the time offsets of the questions within an episode, to split up the subtitles. Subtitles files for each question could then be presented to the GATE pipeline individually, ensuring the NEs extracted were associated with the relevant question.

The source for tweets used for the originally chosen programme set was a dataset collated from the Twitter Garden Hose[2] in 2010. This contains a 10% random sample of all tweets. In 2010, the *@bbcquestiontime* user was not as heavily advertised during Question Time episodes, so a number of hashtags, users and keywords were provided to Apache Hadoop[3], (a distributed processing framework that uses MapReduce [Lin2010]), to filter out tweets that were relevant to Question Time. These were then further filtered using OpenRefine[4], as this allowed false positives to be more easily identified and target filters to avoid removing relevant tweets. OpenRefine also allowed for extraneous fields to be removed from the metadata for a tweet, so the GATE pipeline need only be provided with relevant information.

[1] http://www.bbc.co.uk/programmes/

[2] http://allthingsd.com/20101110/twitter-firehose-too-intense-take-a-sip-from-the-garden-hose-or-sample-the-spritzer/

[3] http://hadoop.apache.org/

[4] http://openrefine.org/

2.2. The GATE pipeline: Named Entity Recognition and Enrichment

The pipeline that executes this task is illustrated in Figure 2. It should be seen as an exploded view of the GATE emblem in Figure 1.

Identification of Named Entities (NEs) such as people, organisations and locations is fundamental to the process of knowledge extraction and is the starting point of more advanced text mining algorithms.

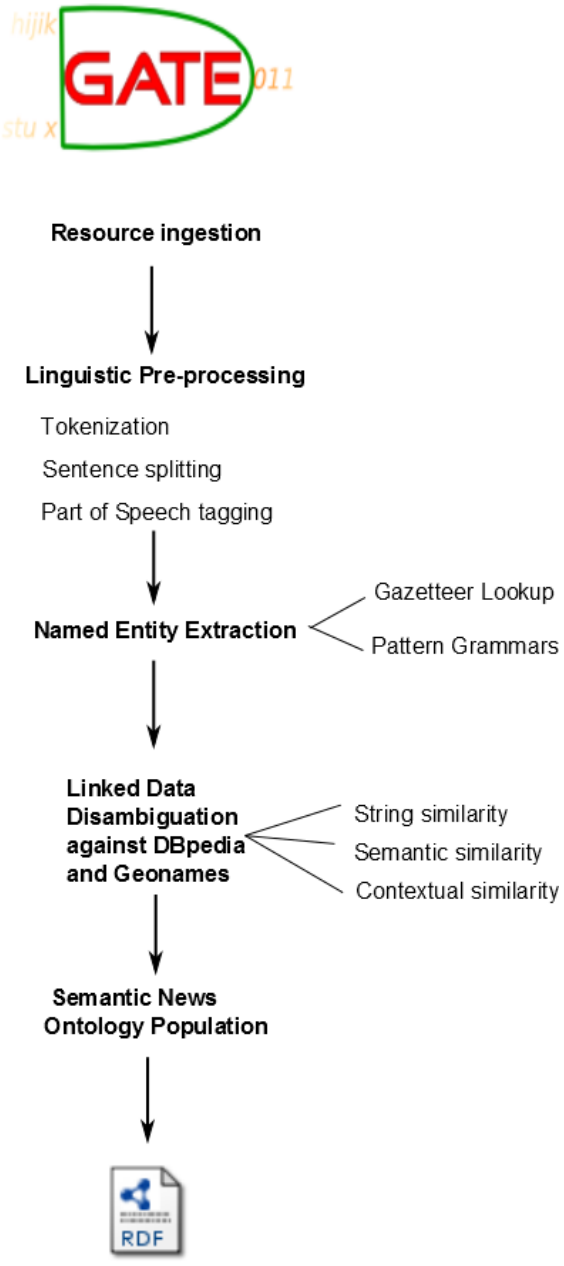

Figure 2: The GATE pipeline

The GATE architecture [Cunningham2011] applies the two modules below for this process in SemanticNews.

In the first stage, GATE's ANNIE Information Extraction [Cunningham2002] system identifies named entities in text and provides annotations as metadata. ANNIE combines gazetteer lookup lists and rule-based grammars and produces NE types such as Organization, Location and Person. ANNIE also resolves co-reference so that entities with the same meaning are linked. For example, General Motors and GM would be identified as referring to the same entity.
ANNIE can process text of generally good quality. Information extraction from social media content, as applied in SemanticNews, has only recently become an active research topic, following early experiments which showed this genre to be extremely challenging for state-of-the-art algorithms [Derczynski2013]. Simple domain adaptation techniques (e.g. [Daume2007]) are not so useful on this genre, in part due to its unusual structure and representation of discourse, which can switch between one-to-one conversation, multi-party conversation and broadcast messages. For instance, named entity recognition methods typically have 85-90% accuracy on longer texts, but 30-50% on tweets [Ritter2011, Liu2012].

In order to handle social media specific text, we have used TwitIE, GATE's open-source NER pipeline, which is specifically adapted to microblog content [Derczynski2013]. It is available as a GATE plugin available to download from https://gate.ac.uk/wiki/twitie.html, and is usable via both the GATE Developer user interface and via the GATE API. The main GATE pipeline chooses between ANNIE and TwitIE on the basis of the type of document it ingests. Both pipelines apply linguistic pre-processing in the form of tokenization, sentence splitting and part of speech tagging.

The second stage, ontology-based entity recognition, is typically broken down into two main phases: candidate selection (entity annotation) and entity linking (also called reference disambiguation or entity resolution) [Rao2013].

Entity annotation is concerned with identifying in a given text all candidate mentions of instances from a knowledge base such as DBPedia[5]. The entity-linking step then uses contextual information from the text, coupled with knowledge from the ontology, to choose the correct instance URI if there are multiple candidates. If there is no such corresponding instance, then a NIL value needs to be returned (an open domain assumption). In particular, entity linking needs to handle name variations (entities are referred to in many different ways) and entity ambiguity (the same string can refer to more than one entity) [Ji2011, Rao2013].

Linking Open Data (LOD) resources, in particular DBpedia and YAGO[6], have emerged as key sources of large-scale ontological knowledge, and are used as anchor entity knowledge bases for entity linking. These offer:

[5] http://dbpedia.org/
[6] http://www.mpi-inf.mpg.de/yago-naga/yago/

1. Cross-referenced domain-independent hierarchies with thousands of classes and relations and millions of instances.
2. An inter-linked and complementary set of resources with synonymous lexicalisations.
3. Grounding of their concepts and instances in Wikipedia entries and other external data.

The rich class hierarchies are used for fine-grained classification of named entities, while the knowledge about millions of instances and their links to Wikipedia entries are used as features in the entity linking algorithms. However, as noted by [Gruhl2009], the large-scale nature of LOD resources also makes entity linking particularly challenging, due to the ambiguity introduced by the presence of so many instances.

For SemanticNews, GATE's LODIE application targets information extraction using Linked Open Data. It has been developed within the TrendMiner project[7]. An online demo is available at http://demos.gate.ac.uk/trendminer/obie/.

LODIE incorporates DBpedia and GeoNames[8] into the extraction process and performs disambiguation in order to unambiguously associate annotations with URIs. The focus is on open-domain entity linking, similar to that carried out by DBPedia Spotlight, as opposed to domain-specific approaches (e.g. [Gruhl2009]).

LODIE applies the following methodology:

1. Identify NEs (Location, Organisation and Person) using ANNIE.
2. For each NE add URIs of matching instances from DBpedia.
3. For each ambiguous NE calculate disambiguation scores.
4. Remove all matches except the highest scoring one.

The disambiguation algorithm uses context in which the particular entity appears and a weighted sum of the following three similarity metrics:

A. String similarity: refers to the Levenshtein distance between the text string (such as Paris), and the labels describing the entity URIs (for example, Paris Hilton, Paris and Paris, Ontario).

B. Structural similarity is calculated based on whether the ambiguous NE has a relation with any other NE from the same sentence or document. For example, if the document mentions both Paris and France, then structural similarity indicates that Paris refers to the capital of France. All other entity URIs can be disregarded, based on the existing relationship between *dbpedia:Paris* and *dbpedia:France*.

[7] http://www.trendminer-project.eu/
[8] http://www.geonames.org/

C. Contextual similarity is calculated based on the probability that two words have a similar meaning as in a large corpus (DBpedia abstracts in our case) they appear with a similar set of other words. To implement that we use the Random Indexing method [Sahlgren2005] and calculate similarity using the cosine function.

2.3. Domain Conceptualization

In order to conceptualise and exploit the SemanticNews information we created an OWL model of the project's domain. The extent of this ontology is illustrated in Figure 3. It is the product of the following design choices, a number of which had also been made for modelling the information space within the web archiving domain in the Arcomem project[9].

1. The adoption of main de facto standard concepts expressed by foundational ontologies including Dolce Ultra Light[10] [Gangemi02]. According to this ontology, there is a conceptual distinction between *InformationObject* and *InformationRealization*. The first covers the knowledge types and instances extracted from the material such as named entities. The second captures the concrete instantiations of these information objects in the form of multimedia web objects such as texts and images concrete data files that contain instances from these information objects. The *realizedBy* relation between *InformationObject* and *InformationRealization* bridges the notions of information and container, and expresses this high level conceptual distinction.

2. Links to existing Linked Data ontologies, such as FOAF[11] for people and organisations, DBpedia for people, organizations and places, SIOC[12] for social media (Twitter in our case), and GeoNames[13] for places. This is of course not an exhaustive list of possible links, but for now it covers a representative section from the LOD cloud. GATE's LODIE component used for this semantic enrichment and consolidation work, also provides a confidence score for each link.

3. Apart from concepts and properties for information content, realization and LOD linking, the ontology must contain elements from social media in the form of Tweets and Users.

4. Furthermore, it covers domain specific concepts for the detailed structure of the BBC data, i.e. Question Time episodes with their elements such as questions and subtitles.

The data model is downloadable from http://semanticnews.org.uk/ontology/semanticnews.owl.

[9] http://www.www.arcomem.eu/
[10] http://www.loa-cnr.it/ontologies/DUL.owl
[11] http://www.foaf-project.org/
[12] http://www.sioc-project.org/
[13] http://www.geonames.org/ontology/documentation.html

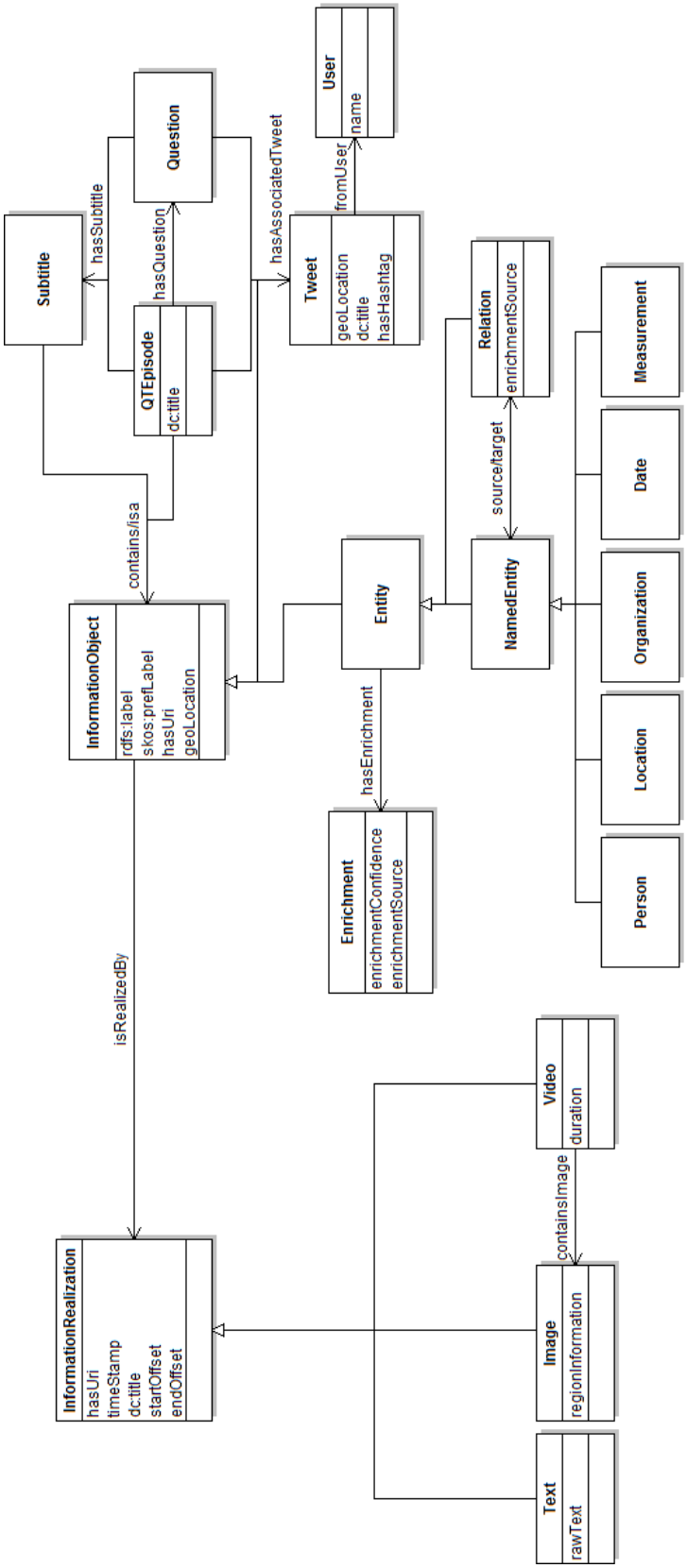

Figure 3: The SemanticNews Data Model

2.4. Collating and Querying

RDF generated by the GATE pipeline for subtitles and tweets can be directly imported into a RDF database. For the SemanticNews project the 4Store[14] application was used. In addition to this RDF a number of RDF graphs from other sources were either imported or remotely queried:

1. The SemanticNews ontology provides the concept and relationships between the entities described in the other RDF data sources.

2. The SemanticNews episodes and questions RDF graphs provide contextual information for these entities. In some cases, this information may be available from the BBC programmes metadata RDF. However, summarising the data in these graphs and referencing the appropriate BBC programmes graphs (using *rdfs:seeAlso*) simplifies the SPARQL queries required later.

3. BBC programmes metadata RDF provides a wide range of information about a particular episode of a programme. This includes broadcast times, channels and versions, episode descriptions, panellists, programme segments (questions), etc.

4. DBPedia people, places and organisations RDF graphs are representations of Wikipedia pages and provide information about the entities these pages describe, as well as relationships to other entities.

5. BBC News Labs triple store allows remote SPARQL queries to find news stories about particular entities.

6. Linked Open Datasets provide structured data in a wide array of areas: political, economic and social. In particular SemanticNews imports a dataset of 2010 election results to provide context on the location of a Question Time episode.

These sources and the results from GATE's NER and LODIE, interlink with each other to allow the following SPARQL queries to be written that can aggregate together data from these multiple sources:

1. Metadata (such as broadcast time, panellists, location, etc.) for a particular episode.
2. The questions and their time offsets within a particular episode.
3. The subtitles and entities associated with a particular question.
4. The tweets and entities associated with a particular episode.
5. The entities discussed in a particular question and additional entities that interlink these entities, with information about these links and a descriptive excerpt from Wikipedia.
6. The most recent news stories about a particular entity discussed in a question.
7. The previous (2010) election results for the constituency hosting the episode.

[14] http://4store.org/

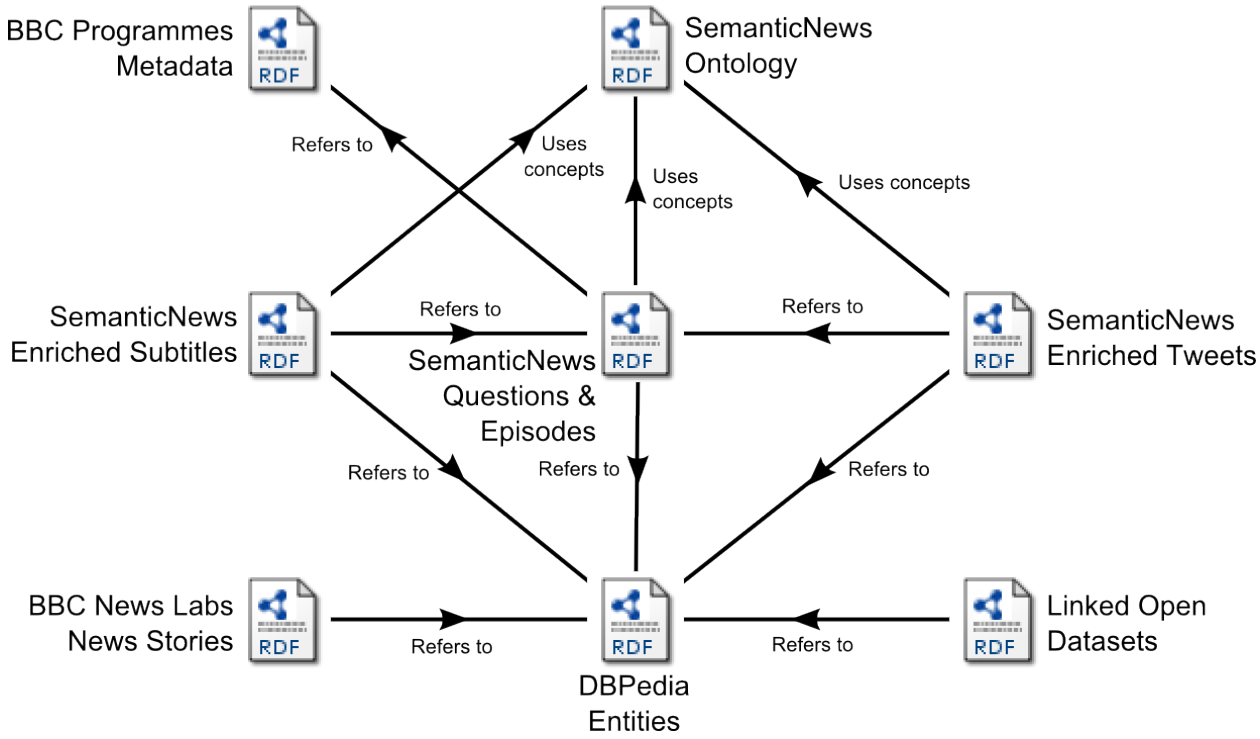

Figure 4: Inter-relations between SemanticNews RDF data sources

In most cases, these data sources were already in a usable form. However, in some cases some automated and even manual manipulation was required. The people, places and organization entities from DBPedia are often made up of huge numbers of triples, many of which are not required for SemanticNews's queries. To save importing all these triples, the NTriples format of the entity was downloaded because a Perl script could be used at a textual level to filter out irrelevant triples before importing only the relevant ones into 4Store [Grant2004].

The dataset available for elections results was retrieved from the Electoral Commission[15]. Although this source is Open data it is not in any real way Linked data. However, it was well structured. This made it possible to use OpenRefine to join the DBPedia resource URIs for the 2010 UK constituencies (retrieved through a direct DBPedia SPARQL query) to the constituency names in the Electoral Commission dataset. This did also require manual amendments to some of the constituency names, such as transforming "Antrim East" to "East Antrim". Wikipedia also provided a table of parties involved in the 2010 UK General Election with links to their Wikipedia pages[16]. Using an automated script it was possible to take the CSV generated from this table and the CSV exported from OpenRefine to produce an election results LOD dataset in RDF/XML format[17], which is conceptualised in a simple OWL ontology[18].

[15] http://www.electoralcommission.org.uk/__data/assets/excel_doc/0003/105726/GE2010-results-flatfile-website.xls
[16] http://en.wikipedia.org/wiki/2010_UK_General_Election#Results
[17] http://www.semanticnews.org.uk/data/elections/uk/general/2010.rdf
[18] http://www.semanticnews.org.uk/ontology/semanticnewselections.owl

3. The SemanticNews Demo Application

The purpose of the demo application is to demonstrate that contextual information about the people, places and organizations discussed in Question Time can be collated from multiple Linked Data sources and presented accompanying the video in a synchronized manner.

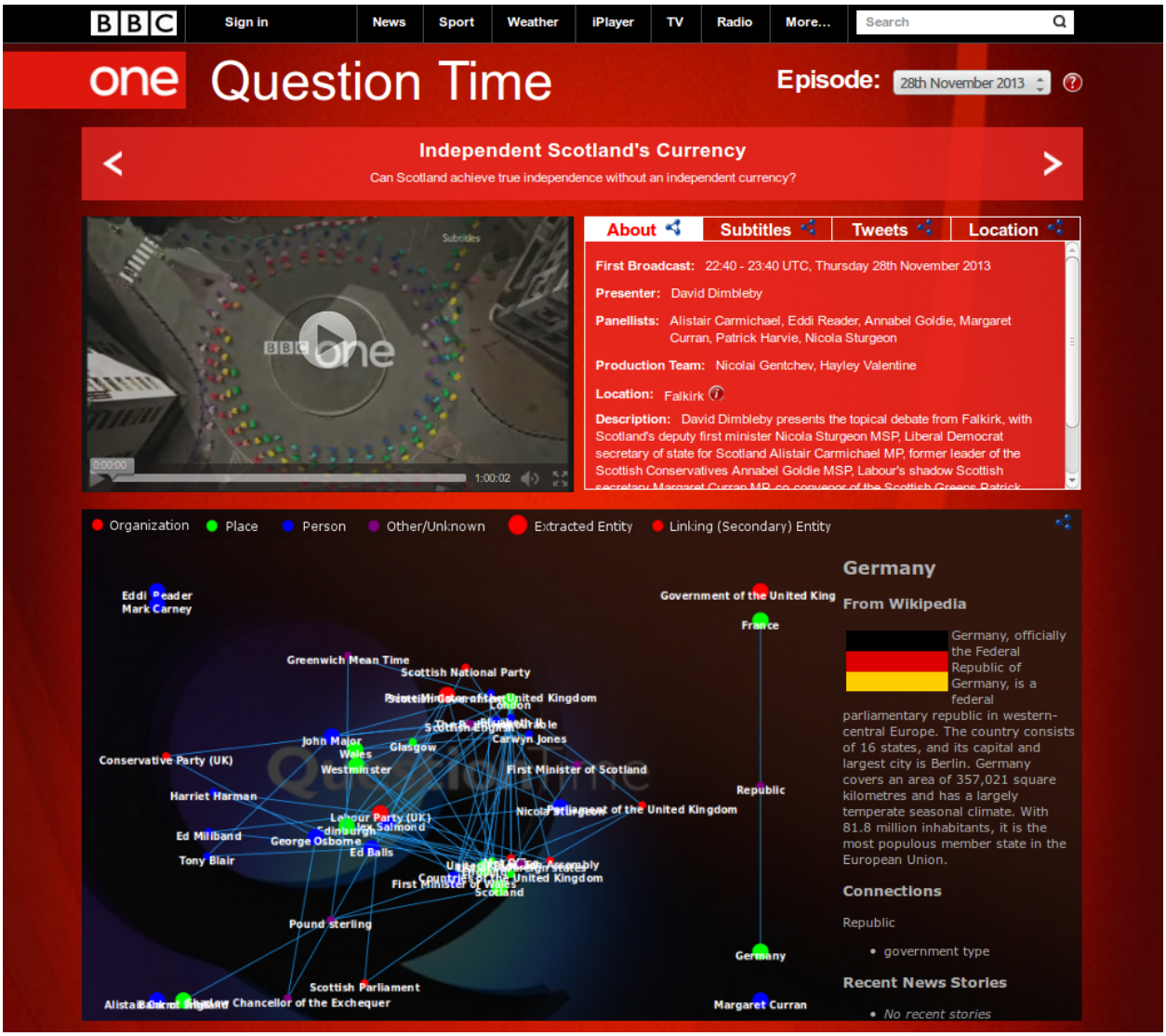

Figure 5: Full screenshot of SemanticNews demo application

The demo application illustrated in Figure 5 can be viewed at http://demo.semanticnews.org.uk/. The application allows the user to select the episode of their choice. However, this is currently limited to 2010 episodes and 2013 episodes from October 31st to November 28th.

Once the demo application has loaded the bar at the top of the page will display the first question of the episode. Clicking on the arrows at the right and left hand sides of this bar will move to the next or previous question respectively.

Beneath the question bar is the episode video in an HTML5 video tag. To the right of this there is a set of four tabs:

1. **About tab:** Displays information about the programme such as the first broadcast time; the presenter, panellists and production team; the location and the episode description.

2. **Subtitles tab**: Displays the subtitles for the current question in chronological order. Where a named entity has been identified, it is displayed in a larger and bolder font.

3. **Tweets tab**: Displays the tweets for the whole episode in chronological order. Like the subtitles tab, the identified named entities are displayed in a larger and bolder font.

4. **Location tab**: Displays information about the location that hosted the episode selected. In particular the constituency of the location and the previous election's voting results compared with the national average.

Each of these four tabs have blue triangular icons clicking on these will load a page with Linked Data resources, (including JSON, RDF and SPARQL queries), relating to the particular tab. Some of the tabs include red 'i' icons which link to addition information about an element.

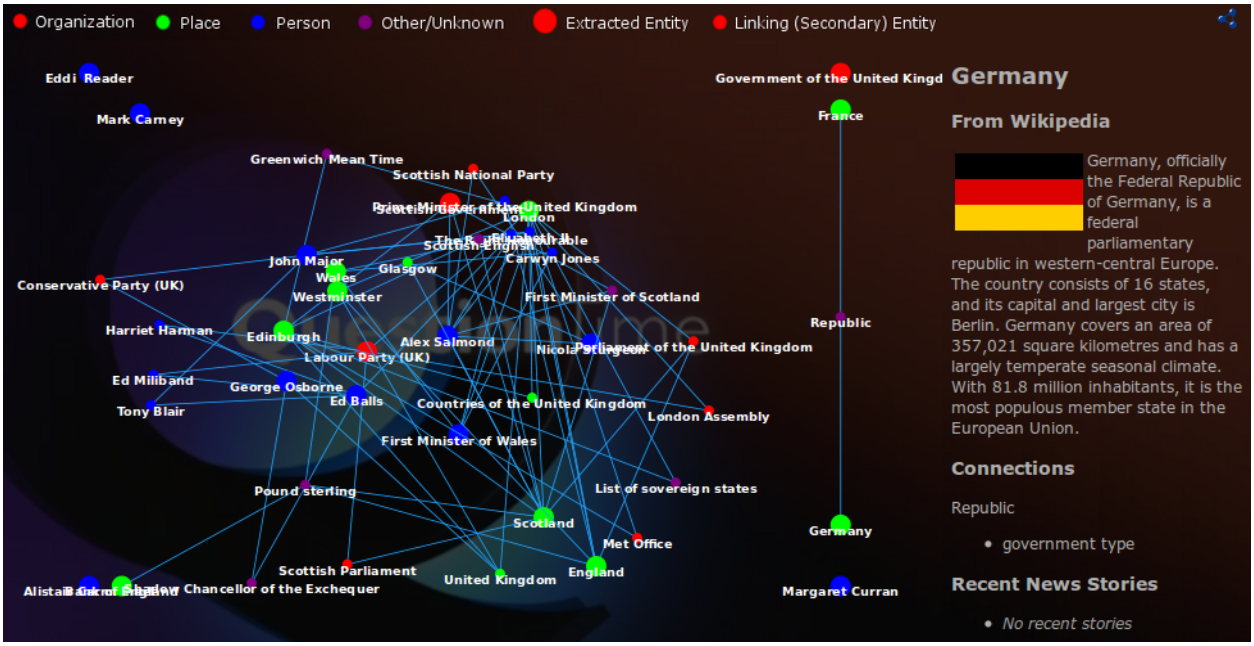

Figure 6: Screenshot of Question Visualization

Beneath the video and tabs is the main visualization as illustrated in Figure 6. This displays a visualization of the current question. This displays all the named entities identified from the

subtitles as nodes in an automatically generated Force Directed graph [Kaufmann2001]. People are displayed as blue nodes, places as green nodes and organisations as red nodes. It also uses the Linked Data DBPedia URIs of these named entities to identify other entities on DBPedia that are linked to by two or more of the named entities. These are also displayed on the visualization but as smaller nodes and where the node cannot be identified as a person, place or organisation it is displayed in purple. Like the four tabs about the question visualization, there is a blue triangular logo in the top right, which links to Linked Data resources used in building the question visualization.

Clicking on a node in the visualization graph will display information about that entity on the right hand side of the visualization. This information includes an extract from the Wikipedia article for entity, the connection of that entity node to other nodes in the graph and a list of recent news stories about that entity returned by a SPARQL query to the BBC News Labs triplestore.

Various aspects of the demo application interact with each other. In particular as the video plays, the subtitles and tweets in the tabs will scroll to be in sync with the time of the video. When the time in the video reaches that of the start of the next question the question bar, subtitles tab and question visualization will update to show the relevant question. Similarly, if the next or previous question is clicked for from the question bar, the time offset of the video, the subtitles tab and the question visualization will change to the new question. Clicking on the NE in the subtitles and tweets tabs also causes the question visualization to display information about this entity in its right hand bar.

4. Future Plans and Exploitation

The scope of the SemanticNews project has been fairly restricted, due to the short duration of the project. Beyond its technological goal of providing visual contextual information to accompany a particular BBC programme, the purpose of the project has been to investigate the feasibility of achieving this goal on a grander scale.

The demo application described in section 3 shows that there is enough data available to produce a multi-faceted visualization of the media and metadata provided by the BBC. However, in section 2 it has been shown that transforming this data into a useful format can be complex and often difficult to automate. In some cases these transformations have been successfully automated but in others manual effort is still required.

One example of this automation is the scripts used to pre-parse subtitles; these can also be reused on other programmes (they have already been tested on subtitles from a House of Commons debate on BBC Parliament). Equally, GATE's NER and LODIE pipeline for subtitles could be applied to these subtitles with little if any adaptation. The post-processing script for finding linking entities in DBPedia could be reused on other tasks, even beyond the wider scope of the SemanticNews project.

It is not just the automated scripts that could find a wider use, the collation of election data and transforming it into Linked Data RDF, is also an output that could be reused even beyond the wider scope of the SemanticNews project.

Although the demo application is multi-faceted there is still a lot more Linked Data that could have been incorporated. The project focussed on the Named Entities extracted from the subtitles and how to visualize them with the use of Linked Data. Equally interesting is looking at the programme metadata and how that can be used with Linked Data, as well as extending the extraction of semantic types under consideration from named entities to domain specific terminology and its Linked Data resolution.

At present there is just a basic example of using the manually defined episode constituency to link to election results and how they compare to the national average. Other data about the location and how it compares to national averages could be produced for things such as crime, health, welfare, and other socio-economic data. Also, discovering Linked Data associated with the panellists would be another avenue to investigate. E.g. if they were MPs, displaying statistics about their voting records might be interesting. Also evaluating their Wikipedia page (using DBPedia) to find associated entities could help place their appearance in context.

The BBC is continually working on various APIs to interact with their data. We have used a couple of them within the project but there are a few further things we can do. The BBC has been working on aggregating news stories under an *event* and even arranging them into a *storyline*. When this work is complete, being able to link individual questions out to an event or storyline will give a much wider range of news stories that should be more accurate that just linking an entity to its five most recent news stories. A short-term alternative could be to

investigate whether the BBC NewsLabs API could be used to find news stories that relate to two or more named entities that have been extracted from the subtitles. However, evaluating the pairs to query would not be a straightforward task.

Sheffield has developed a product called Mímir that could also benefit future extensions of SemanticNews. Mímir[19] 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 meta-data (instance data). It allows queries that arbitrarily mix full-text, structural, linguistic and semantic queries and that can scale to gigabytes of text. A typical semantic annotation project deals with large quantities of data of different kinds. Mímir provides a framework for implementing indexing and search functionality across all these data types.
In other words, the goal of Mímir is to allow for a richer search experience as it will allow unrestricted combinations of text, entities and Linked Data. A demo application using BBC news stories[20] was developed at the BBC #newsHACK event. An example query might be find all people who were born in Sheffield that were found in a BBC news story saying something (e.g. John Smith said or John Smith says):

```
{Person sparql = "SELECT ?inst WHERE { ?inst :birthPlace
<http://dbpedia.org/resource/Sheffield>}"} [0..4] root:say
```

Southampton has also been working on techniques that could be of use to SemanticNews, specifically in the area of video hyperlinking, whereby automatic methods are developed to link together relevant sections of video [Preston2013]. Southampton researchers took part in the 2013 MediaEval search and hyperlinking evaluation [Eskevich2013], which used video and transcript data provided by the BBC.

In summary there are many directions, in which SemanticNews could be extended, whether it be into a greater range of BBC broadcast media, using other Linked Data sources to enhance the current application or refining methods to provide more accurate and automated versions of the current visualizations.

4.1. Recommendations to the BBC

One critical factor for the success of any future work is to continue our close relationship with the BBC, and in particular to work collaboratively to improve the metadata and features of existing BBC APIs. Some specific recommendations follow.

The metadata for Question Time episodes in 2010 had programme segments, splitting each episode up into its individual questions. This made it possible to go straight from the subtitles XML stream to something that could be directly read in by GATE for an individual question. For these 2013 episodes this metadata was no longer available and the splitting of subtitles and the question metadata had to be generated manually. Reinstating this service would be of great benefit. There is also the opportunity for specific research in automating this process.

[19] https://gate.ac.uk/mimir/
[20] https://annomarket-demo.services.gate.ac.uk/newshack/

The BBC Programme metadata provides a description, which includes information about panellists in plain text; although it should be possible to use NER and LODIE to determine the panellists, if this information was semantically encoded as LOD URIs in the metadata it would avoid potentially introducing any errors through the LODIE process.

Often the BBC Programme metadata provides information about the location. Of particular use to the SemanticNews project would be an explicit property to define the constituency the Question Time episode was in. Otherwise, accurate data about the location would be useful; for example, not just London but The Borough of London.

It was observed that the BBC NewsLabs news stories API would find the closest matching identifier if it could not find a perfect match. This would often lead to some rather unexpected related news stories being found[21]. If this could be improved it would ensure the that related news stories provided in the question visualization actually refer to the entity clicked on.

The current subtitles stream produced by the BBC does is very presentation centric. It does highlight text different colours depending on the speaker. However, there is no information to define who is represented by each colour. Semantic enhancement of the subtitles so that each element was marked with the actually speaker, represented as a URI, would make it possible to make use of the information already partially encoded in the subtitle streams. This should be fairly easy for fully automatic transcription. It would then for example be possible to compare and contrast named entities mentioned by each panellist. If the semantic enhancement could not be incorporated into the existing format, it should only take a simple script to convert a semantic representation into this format.

[21] http://www.bbc.co.uk/blogs/internet/posts/BBC-News-Lab

References

[Cunningham2002] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan (2002) *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.* In Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL02).

[Cunningham2011] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters (2011) *Text Processing with GATE (Version 6).*

[Daume2007] H. Daume and D. Marcu (2007) *Frustratingly easy domain adaptation.* In Proc. of the Annual meeting of the Association for Computational Linguistics.

[Derczynski2013] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. (2013) *Microblog-Genre Noise and Impact on Semantic Annotation Accuracy.* In Proc. of the 24th ACM Conference on Hypertext and Social Media, ACM.

[Eskevich2013] M. Eskevich, G. J. Jones, S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In MediaEval 2013 Workshop, Barcelona, Spain, October 18-19 2013.

[Gangemi2002] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider (2002) *Sweetening Ontologies with DOLCE.* In Proc. EKAW, Springer.

[Grant2004] J. Grant and D. Beckett, eds. (2004) *RDF Test Cases: NTriples.* W3C Recommendation.

[Gruhl2009] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth (2009) *Context and Domain Knowledge Enhanced Entity Spotting in Informal Text.* In Proc. of the 8th International Semantic Web Conference (ISWC2009).

[Ji2011] H. Ji and R. Grishman (2011) *Knowledge base population: Successful approaches and challenges.* In Proc. of ACL2011, pages 1148-1158.

[Kaufmann2001] M. Kaufmann and D. Wagner, eds. (2001) *Drawing graphs: methods and models*, Lecture Notes in Computer Science 2025, Springer, pages 241-242.

[Lin2010] J. Lin and C. Dyer (2010) *Data-Intensive Text Processing with MapReduce.* Morgan & Claypool Publishers.

[Liu2012] X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou. (2012) *Joint inference of named entity recognition and normalization for tweets*. In Proc. of the Association for Computational Linguistics, pages 526-535.

[Preston2013] J. Preston, J. Hare, S. Samangooei, J. Davies, N. Jain, D. Dupplaw, and P. H. Lewis. A unified, modular and multimodal approach to search and hyperlinking video. In MediaEval 2013 / Search and Hyperlinking of Television Content, October 2013.

[Rao2013] D. Rao, P. McNamee, and M. Dredze (2013) *Entity linking: Finding extracted entities in a knowledge base*. In Multi-source, Multilingual Information Extraction and Summarization. Springer.

[Ritter2011] A. Ritter, S. Clark, Mausam, and O. Etzioni (2011) *Named entity recognition in tweets: An experimental study.* In Proc. of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK.

[Sahlgren2005] M. Sahlgren (2005) *An introduction to random indexing*. In Proc. of the Methods and Applications of Semantic Indexing Workshop, Copenhagen, Denmark.