BreXearch: Exploring Brexit Data Using Cross-Lingual and Cross-Media Semantic Search

Lei Zhang[1], Maribel Acosta[2], Michael Färber[3], Steffen Thoma[2], and Achim Rettinger[2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
lei.zhang@fiz-karlsruhe.de
[2] Karlsruhe Institute of Technology (KIT), Germany
{acosta|steffen.thoma|rettinger}@kit.edu
[3] University of Freiburg, Germany
michael.faerber@cs.uni-freiburg.de

Abstract. BreXearch is a cross-lingual and cross-media semantic search system that focuses on the Brexit use case. This system has extracted the knowledge from various media sources (including online news sites, social media and live-TV) in three languages (i.e., English, German and Spanish) and integrated it with the additional background knowledge from DBpedia. Based on that, BreXearch allows us to search for different media channels using keyword queries or by means of entities and to ask complex questions regarding Brexit using SPARQL queries.

1 Introduction

The exit of the UK from the EU, known as *Brexit*, has already been subject to a number of sociological and economic studies, such as how topics and opinions spread in the public discussions (visible/transmitted via public media such as news articles, social media and TV shows) and how those topics and opinions relate to specific "items" such as persons in the public (*David Cameron*, etc.) and subjects of public concern (*Euroscepticism*, etc.). Recently, many approaches have been proposed for search and analysis concerning Brexit. The existing work ranges from using the mentioned words, topics, themes and sentiments, however, no explicit semantics in knowledge bases has been utilized for semantic processing. In this regard, we present BreXearch, a cross-lingual and cross-media semantic search system with a focus on the Brexit content, which allows us to retrieve media items from various sources in different languages using entities, e.g., the British politician `Boris_Johnson`, and to answer complex questions, e.g., *"Which politicians from the* Conservative Party of UK *were most present in different media channels and languages regarding a specific subject in the last two weeks before the Brexit referendum?"*, using SPARQL queries.

The main contributions of this work are: (1) we have collected a large dataset of media coverage on the Brexit referendum from various multilingual sources; (2) such media content has been then semantically enriched with annotations of both entities and categories from DBpedia; (3) in addition, we have introduced a

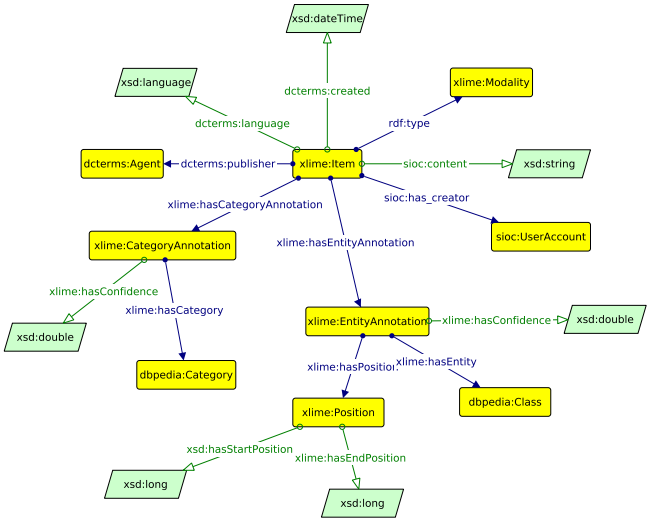

Fig. 1. The schema of the semantic data model.

semantic data model to describe and integrate knowledge extracted from media content in different modalities and languages; (4) in combination with additional background knowledge from DBpedia, BreXearch allows for cross-lingual and cross-media retrieval and analytics of the Brexit content. A screencast of the BreXearch demo is available at http://km.aifb.kit.edu/sites/BreXearch/.

2 BreXearch Architecture

The BreXearch architecture consists of a pipeline of several components for *data collection*, *annotation*, *modeling* and *storage*, which will be now briefly described.

Data Collection. To collect the Brexit data, we employ three media sources: (1) multilingual streams of news articles across the world; (2) social media data in multiple languages from social networks, forums, blogs and review sites; (3) live-TV streams consisting of video frames and audio for multilingual channels. From these data sources, we collect a dataset of media items related to the Brexit referendum held on June 23, 2016 by using a set of filters based on time and keywords. It results in around 240 thousand news articles, 12 million microposts and 900 TV programs. Afterwards, the textual content is extracted from the collected media items. More details about data collection can be found in [1].

Data Annotation. The extracted textual content is then semantically enriched with DBpedia annotations, where a very important part is *cross-lingual entity linking*, which detects not only named entities (e.g., `David_Cameron`) but also nominal entities (e.g., `Prime_Minister_of_the_United_Kingdom`) in the multilingual text and disambiguates them with DBpedia entities. For this, we employ our cross-lingual semantic annotation tool, called *X-LiSA* [2]. Another important part is *entity-based categorization*, which aims to derive the categories related to the media items. In most of the semantic knowledge bases, entities are

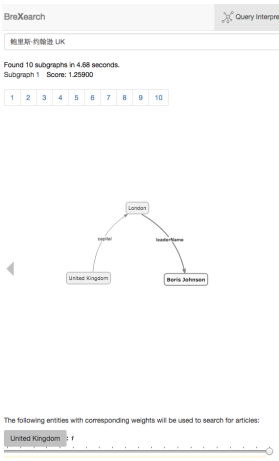

Fig. 2. Screenshot of query interpretation.

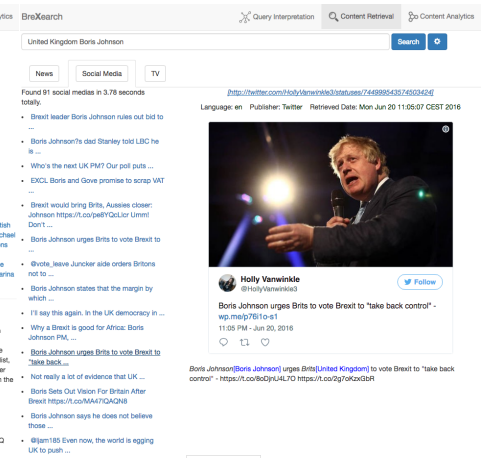

Fig. 3. Screenshot of content retrieval.

organized in a category hierarchy. For example, the DBpedia entity `Brexit` has its parent category `Category:Euroscepticism_in_the_United_Kingdom`, which in turn is a subcategory of `Category:Euroscepticism`. By utilizing this category hierarchy, each media item is further enriched with the related categories based on its mentioned entities (see more details in [1]).

Data Modeling. In order to enable semantic integration of and seamless access to media data in multiple modalities, languages and sources, we introduce a semantic data model. Its schema is depicted in Fig. 1, which enables to relate text and audio/video streams to entities and categories in DBpedia. For each entity annotation, the predicates that define the start and end positions of the entity mention are used in a flexible manner and may define character positions in the case of text, or milliseconds/frame numbers in the case of audio/video. Each category annotation captures one subject of the media content. In any case, each entity name mentioned in or each subject covered by any media item should relate to an entity or a category in DBpedia.

Data Storage. Based on the schema in Fig. 1, we model the annotated media data as both JSON documents and RDF triples. In addition, a NoSQL database (i.e., MongoDB) and a triple store (i.e., Virtuoso) for the data storage provide further query capabilities and data integration with restrictions and aggregates on multiple modalities, languages and sources as well as in a combination with additional background knowledge about entities and categories in DBpedia.

3 User Interface

Now we describe the features of the BreXearch user interface that can be used for cross-lingual and cross-media retrieval and analytics of the Brexit content.

Query Interpretation. BreXearch supports keyword search with a query in any language (even with keywords in multiple languages). Instead of retrieving media items directly by keywords, BreXearch first finds the query entity graphs

```
SELECT COUNT(DISTINCT ?media) as ?count ?politician WHERE {
   ?item dcterms:source ?media .
   ?item dcterms:created ?date .
   ?item dcterms:language "en"^^xsd:language .
   ?item rdf:type sioc:Microblog .
   ?item xlime:hasEntityAnnotation ?entityAnnotation .
   ?item xlime:hasCategoryAnnotation ?categoryAnnotation .
   ?entityAnnotation xlime:hasEntity ?politician .
   ?politician dbpedia-owl:party dbpedia:Conservative_Party_(UK) .
   ?categoryAnnotation xlime:hasCategory dbpedia:Category:Euroscepticism_in_the_United_Kingdom> .
   FILTER (?date > "2016-06-09"^^xsd:date && ?date < "2016-06-23"^^xsd:date)
} GROUP BY ?politician ORDER BY DESC (?count)
```

Fig. 4. Example of SPARQL query.

(QEGs) as the semantic interpretations of a query by exploring the knowledge graph of DBpedia with nodes representing entities and edges describing their relations. An example of QEG for the query "鲍里斯·约翰逊 UK" is shown in Fig. 2. More details about query interpretation can be found in [3].

Content Retrieval. Through the semantic annotation and integration on multiple modalities, languages and sources, BreXearch supports cross-lingual and cross-media retrieval of the Brexit content by means of entities, which can be either selected from the QEGs generated by query interpretation or directly entered into the search box by users. An example of the retrieved social media items using the entities `Boris_Johnson` and `United_Kingdom` is shown in Fig. 3.

Content Analytics. Using the knowledge extracted from different media and languages in combination with knowledge in DBepdia, BreXearch allows us to ask complex questions. For example, the question *"Which politicians from the* Conservative Party of UK *were most present in microblogs in English about the subject* Euroscepticism in the United Kingdom *in the last two weeks before the Brexit referendum?"* can be answered by the SPARQL query shown in Fig. 4.

4 Conclusions

In this paper, we present BreXearch, a semantic search system regarding Brexit supporting different media and languages. It provides a pipeline for cross-lingual and cross-media knowledge extraction from various multilingual media sources. Based on that, the user interface allows for cross-lingual and cross-media retrieval and analytics of the Brexit content. As for the future work, we would like to adapt BreXearch to other use cases, such as the US president election in 2016.

References

1. Zhang, L., Färber, M., Thoma, S., Acosta, M., Rettinger, A.: A semantically enriched brexit dataset using cross-lingual and cross-media knowledge extraction. Technical report, KIT, http://people.aifb.kit.edu/lzh/brexlime.pdf (2017)
2. Zhang, L., Rettinger, A.: X-LiSA: Cross-lingual Semantic Annotation. PVLDB **7**(13) (2014) 1693–1696
3. Zhang, L., Rettinger, A., Zhang, J.: A knowledge base approach to cross-lingual keyword query interpretation. In: ISWC. (2016) 615–631