Exploiting Knowledge Bases for Multilingual and Cross-lingual Semantic Annotation and Search

Lei Zhang*, Michael Färber, Andreas Thalhammer, and Achim Rettinger

Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany

Abstract. The amount of entities in large knowledge bases (KBs) has been increasing rapidly, making it possible to propose new ways of intelligent information access. In addition, there is an impending need for systems that can enable multilingual and cross-lingual information access. In this work, we firstly demonstrate *X-LiSA*, an infrastructure for *multilingual* and *cross-lingual semantic annotation*, which supports interfaces for annotating unstructured text in different languages using resources from KBs. Based on *X-LiSA*, we demonstrate *XKnowSearch!*, a novel system for *multilingual* and *cross-lingual semantic search*, which supports keyword search on textual data by exploiting its semantics.

1 Introduction

Within the context of globalization, *multilingual* and *cross-lingual* access to information has emerged as an issue of major interest. Nowadays, more and more people from different countries are connecting to the Internet and many Web users can understand more than one language. While the diversity of languages on the Web has been growing, for most people there is still very little content in their native language. As a consequence of the ability to understand more than one language, users are also interested in Web content in other languages. In order to achieve the goal that users from all countries have access to the same information, there is an impending need for systems that can help in overcoming language barriers and facilitate multilingual and cross-lingual information access.

In addition, the ever-increasing quantities of entities in large knowledge bases (KBs), such as Wikipedia, DBpedia, Freebase and YAGO, pose new challenges but at the same time open up new opportunities of intelligent information access. Recently, almost every major commercial Web search engine has announced its work on incorporating structured knowledge into their search results, including Google's Knowledge Graph, Yahoo!'s Web of Objects and Microsoft's Satori Graph/Bing Snapshots, where the large entity repositories have become valuable resources for bridging the gap between natural language text and knowledge.

In this work, we demonstrate *X-LiSA* [1], an infrastructure for *multilingual* and *cross-lingual semantic annotation*, which supports interfaces for annotating both text documents and Web pages with resources from KBs. It helps to bridge the ambiguity of natural language text and its formal semantics as well as to transform documents in different languages into a unified representation. Based

* Corresponding author. E-mail: l.zhang@kit.edu.

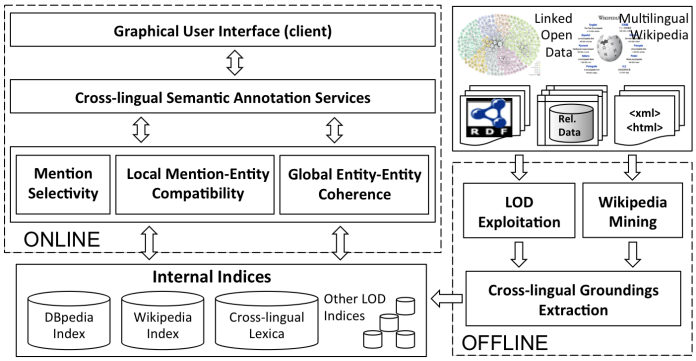

Fig. 1: The system architecture of *X-LiSA*.

on *X-LiSA*, we then demonstrate *XKnowSearch!*, a novel system for entity-based *multilingual* and *cross-lingual semantic search* by translating keyword queries in different languages to their semantic representation. With the help of *X-LiSA*, *XKnowSearch!* bridges the language barriers between queries and documents in different languages, and also facilitates query disambiguation and expansion.

The rest of the paper is structured as follows. We start with the description of *X-LiSA* in Sec. 2. Based on that, *XKnowSearch!* is then presented in Sec. 3. Finally, the appendix about addressing challenge criteria is provided in Sec. 4.

2 Description of X-LiSA

2.1 System Architecture

The system architecture of *X-LiSA* is illustrated in Fig. 1, where *cross-lingual groundings extraction* is performed offline to generate the needed indexes, which are then used by the online *cross-lingual semantic annotation* services.

Cross-lingual Groundings Extraction. For matching words and phrases in different languages against entities in KBs, both *X-LiSA* and *XKnowSearch!* utilize *xLiD-Lexica* [2], our recently established cross-lingual lexica by exploiting various kinds of structures in Wikipedia, such as anchor texts of hyperlinks and cross-language links, to extract the cross-lingual groundings of entities.

Mention Detection. The first challenge of semantic annotation lies in *mention selectivity* with the goal of detecting the boundaries of mentions in text documents that are likely to denote entities. In order to address the challenges of correctness, completeness and emergence of the detected mentions, we employ our recent work [3] that aims to detect both named entities and nominal entities. Such entity mentions serve as the input of entity disambiguation.

Entitiy Disambiguation. For each mention, its candidate entities are then extracted using *xLiD-Lexica*. While the feature of *mention-entity compatibility* captures the most likely entity behind the mention based on the cross-lingual groundings and the entity that best fits the context of the mention based on the cross-lingual semantic relatedness techniques, *entity-entity coherence* collectively captures the related entities as annotations of a document. These features are

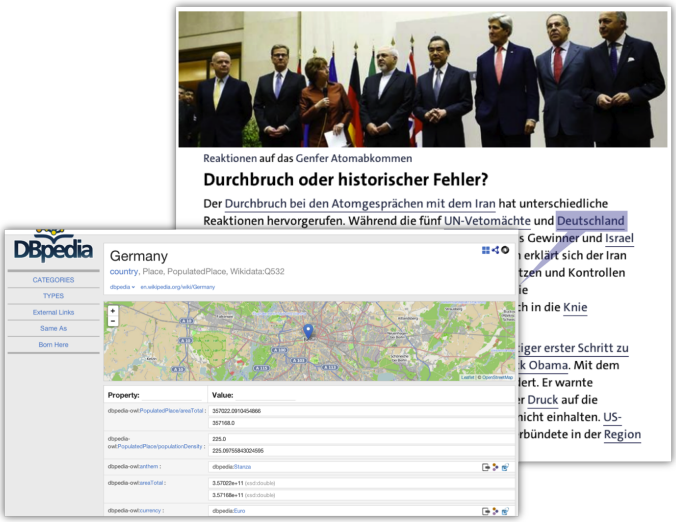

Fig. 2: Annotation service for web pages.

then employed by the *graph-based disambiguation* using a personalized PageRank algorithm to determine the final entity for each mention [1].

2.2 Demonstration

We demonstrate the framework of *X-LiSA* in terms of the *cross-lingual lexica*, the *online annotation service* and the use case of *media annotation and querying*.

Cross-lingual Lexica. Firstly, we would like to demonstrate the extracted cross-lingual lexica *xLiD-Lexica*[1]. The datasets are available as both RDF triples in N-Triples format and plain text files in JSON format. Based on these datasets, we built a SPARQL endpoint and Web interface such that users can easily access the information using SPARQL query language or through the Web interface.

Online Annotation Service. *X-LiSA* supports interfaces for annotating raw text and Web pages in different languages[2]. A screenshot of the cross-lingual semantic annotation service is shown Fig. 2, where the input is the URL of a German news article, the knowledge base is DBpedia and the output language is English. In order to allow not only users but also software agents to access the functionality of text annotation, we also provide the service, which takes raw text and web pages as input and yields the output of annotations in XML.

Media Annotation and Querying. Within the context of the XLike[3] and xLiMe[4] projects, *X-LiSA* has been widely used to annotate textual data from both mainstream media sites and social media, where the following partners have contributed large datasets, which are delivered as streams:

[1] `http://km.aifb.kit.edu/sites/xlid-lexica`
[2] `http://km.aifb.kit.edu/sites/xlisa`
[3] `http://www.xlike.org`
[4] `http://xlime.eu`

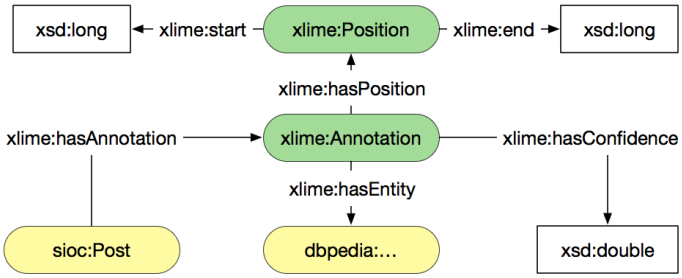

Fig. 3: Vocabulary for Media Annotation.

```
PREFIX xlime: <http://xlime-project.org/vocab/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT COUNT(DISTINCT ?media) as ?count ?entity WHERE {
   ?s xlime:hasAnnotation ?a .
   ?s dcterms:source ?media .
   ?s dcterms:created ?date .
   ?a xlime:hasEntity ?entity .
   ?entity dbpedia-owl:manufacturer dbpedia:Mercedes-Benz .
   FILTER (?date > xsd:date(now()-3600*24*14) && ?date < now()) .
}  GROUP BY ?entity ORDER BY DESC(?count)
```

Fig. 4: SPARQL Query Example.

- *Zattoo*[5]: textual information extracted from visual and audible TV data.
- *JSI NewsFeed*[6]: textual news data which is crawled from online news sites.
- *VICO Social Media*[7]: textual social media data which is crawled from forums, news, blogs, social networks, review sites and others.

For modeling the annotated media data as RDF triples, we define the xLiMe vocabulary, as shown in Fig. 3, using the SIOC ontology[8] and its extensions[9]. A SPARQL endpoint is provided for querying the data[10] For example, given the query "*Which cars produced by Mercedes-Benz were mentioned most in the last two weeks?*", the SPARQL query in Fig. 4 can be used to retrieve the answer.

3 Description of XKnowSearch!

3.1 System Architecture

X-LiSA offers opportunities for dealing with complex queries on the media data. However, the SPARQL query hinders casual users in expressing their information needs. By employing *X-LiSA* for offline text annotation, *XKnowSearch!* supports keyword search by capturing queries and documents at the semantic level and

[5] http://developer.zattoo.com
[6] http://newsfeed.ijs.si
[7] http://www.vico-research.com/en
[8] http://sioc-project.org/ontology
[9] http://kdo.render-project.eu
[10] http://km.aifb.kit.edu/sites/xlisa/queries.html

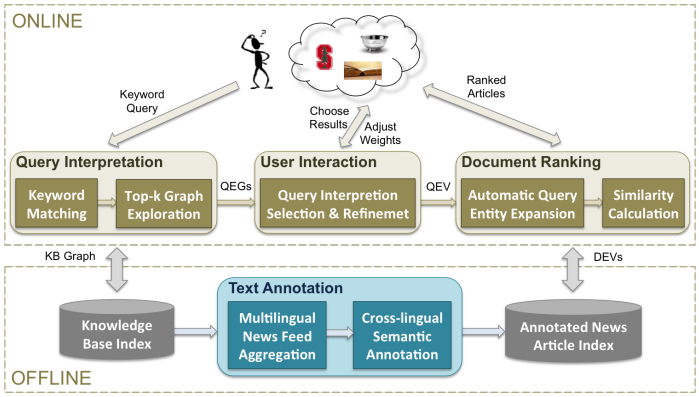

Fig. 5: The system architecture of *XKnowSearch!*.

also bridging the language barriers. The architecture of *XKnowSearch!* is shown in Fig. 5. In the following, we discuss the online processing components.

Query Interpretation. The search process starts with a keyword query in any language. Instead of retrieving documents by keywords, *XKnowSearch!* first finds the *query entity graphs* (*QEGs*) matching the keyword query by exploring the semantic graph of the KB. The resulting QEGs reflect different semantic interpretations of the keyword query and thus can help to refine the query and influence document ranking according to the search intents of users.

User Interaction. As an optional step, user interaction enables interactive query disambiguation and expansion according to users' search intents, where users can select and refine the QEGs by navigating the KB through semantic relations between entities. While user interaction provides users a more flexible way to influence the search process, users can also search the documents directly without interactive query refinement. In this case, the QEG with highest score will be selected as interpretation of the keyword query.

Document Retrieval. The entities in the selected QEG constitute the query entity vector (QEV), where each entry contains the corresponding entity weight, which is calculated by the exploration algorithm and can also be adjusted by users. For each document, we construct the *document entity vector* (*DEV*), where the entries contain the confidence scores of the annotations (i.e., the linked entities in the hub languages) of the document, which are generated by our semantic annotation system. Based on the entities in QEV and DEV, the score of each document can be calculated based on standard similarity measures, such as cosine similarity, which is then used for document ranking.

3.2 Demonstration

We would like to demonstrate four major features of *XKnowSearch!*[11] with the goal of addressing the challenges that traditional keyword search suffers from.

[11] `http://km.aifb.kit.edu/sites/XKnowSearch`

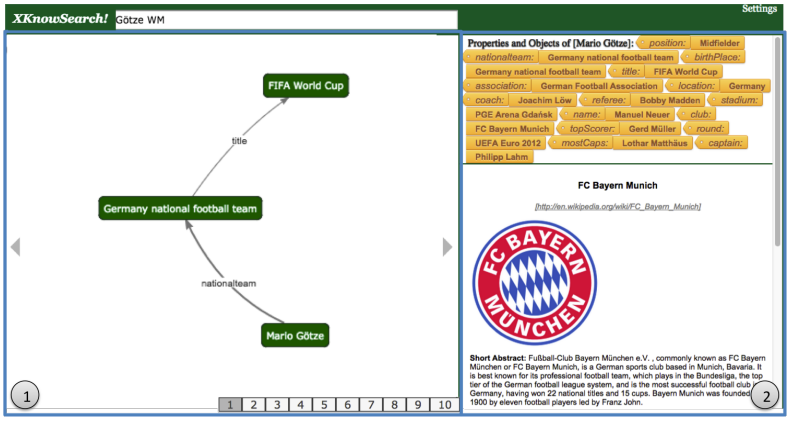

Fig. 6: Query entity graphs for keyword query *"Götze WM"*.

Query Flexibility. Traditional keyword search systems do not allow users to be involved in the search process to perform query refinement. *XKnowSearch!* supports two modes: *direct* and *indirect search*. Direct search takes a keyword query and retrieves the relevant documents directly. Indirect search provides the opportunity for users to understand the meaning of the query entities and the underlying semantic relations between them, such that users are able to refine and extend the information needs. While direct search enables users to search in a familiar and convenient manner, indirect search provides users a *more flexible way* to influence the search process according to their search intents.

Query Disambiguation. Keywords are naturally ambiguous due to the fact that they could refer to different entities in different contexts. In the multilingual and cross-lingual setting, this problem is more serious. For instance, the keyword *WM* can refer to Windows_Mobile in English or FIFA_World_Cup in German[12]. In *XKnowSearch!*, query interpretation component *automatically eliminates the ambiguity* of keywords by taking advantage of the context, i.e., other query entities, and by exploiting the semantic graph of KBs to generate the top-k QEGs. On the other hand, users can also *disambiguate the query manually* by selecting the most appropriate QEG and further refining it (cf. Frame 1 of Fig. 6).

Query Expansion. The query keywords are often incomplete in the sense that instead of the full entity name, only the aliases, acronyms and misspellings are usually given by users. *XKnowSearch!* supports query keywords matching entities in their *incomplete* forms. In addition, keyword queries might contain concept names representing a set of entities, e.g., "Internet companies of China". In *XKnowSearch!*, the matching concepts are automatically expanded into sets of individual entities. As query interpretation, QEG is more informative and expressive than keywords such that it can help users to *manually expand* the query by navigating KBs through semantic relations between entities and adding more intended entities that are used for document retrieval (cf. Frame 2 of Fig. 6).

[12] *WM* is the abbreviation of *Weltmeisterschaft* in German denoting *World Cup*.

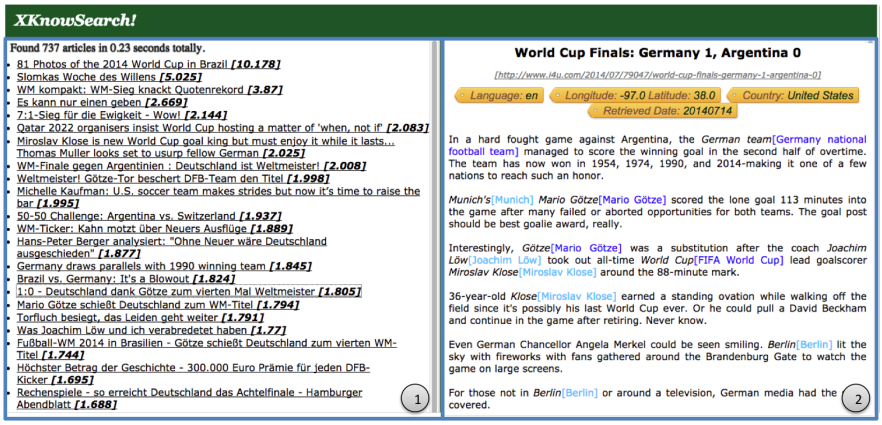

Fig. 7: Retrieved documents for keyword query *"Götze WM"*.

Cross-lingual Search. Term-based retrieval paradigm of keyword search suffers from the vocabulary mismatch problem, which is more challenging in the cross-lingual search setting. *XKnowSearch!* enables cross-lingual search in the sense that users can use keyword queries in any language to retrieve multilingual documents, especially in any other languages. The recent progress in cross-lingual technologies is largely due to the increased availability of multilingual resources on the Web. In this regard, we use the multilingual KB as an interlingua to connect keyword queries and documents across languages (cf. Frame 2 of Fig. 7).

4 Appendix: Addressed Challenge Criteria

4.1 Mandatory Criteria

The application has to be an end-user application. While *X-LiSA* has been widely used as software components in the XLike and xLiMe projects for enabling cross-lingual semantic annotation for publishers, media monitoring and new business intelligence applications, it also has a practical value to general Web users that can better understand text and Web pages though the description of the annotations in their preferred languages. In addition, *XKnowSearch!* was designed as an end-user Web application. It does not require any prior knowledge about the technical aspects and provides the intuitive keyword interface for multilingual and cross-lingual search such that it can be used by any user.

The information sources used. In *X-LiSA* and *XKnowSearch!*, we use multilingual data retrieved from heterogeneous sources under diverse ownership and control including large knowledge bases such as Wikipedia and DBpedia, and textual data collected from both mainstream media such as Bloomberg and New York Times, and social media such as Twitter and Facebook. These large datasets are provided by a consortium including both research institutes such as JSI NewsFeed, and business companies such as Zattoo and VICO Social Media.

The meaning of data has to play a central role. The semantic data plays a central role in both *X-LiSA* and *XKnowSearch!*, where all of the components

are based on Semantic Web technologies such as RDF and SPARQL, and are ultimately built to exploit the fusion of structured knowledge and unstructured text. In addition, the recent progress in cross-lingual technologies is largely due to the increased availability of multilingual KBs, such as Wikipedia and DBpedia. In this regard, the semantics of data captured by such KBs are employed as an interlingua to bridge the language barriers between different kinds of data.

4.2 Additional Desirable Features

The application provides an attractive and functional Web interface. *XKnowSearch!* is offered as a Web application for any user who wants to search Web documents in multiple languages using keyword query.

The application should be scalable. *X-LiSA* can annotate textual data from a set of up to 250 live TV channels, approximately 10 million social media posts and 250,000 news articles per day. *XKnowSearch!* is currently restricted to only news articles but will extended to other data in the near future.

Novelty, in applying semantic technology to a domain or task that have not been considered before. Although many efforts have been made to enable entity-aware Web search, there are still many limitations. For example, current search engines like Google cannot deal with entities expressed by concept name, e.g.,"Internet companies of China", and they do not support cross-lingual search. In this regard, *XKnowSearch!* is a novel system to multilingual and cross-lingual semantic search with the goal of addressing these limitations.

Functionality is different from or goes beyond pure information retrieval. Different from current search engines, *XKnowSearch!* provides a more flexible way to influence the search process by allowing users to understand the meaning of the query entities and the underlying semantic relations between them, such that users can easily refine and extend the information needs.

There is a use of dynamic data, perhaps in combination with static information. While the KBs used by *X-LiSA* is relatively static (updated every few weeks or months), the textual data to be annotated is dynamically retrieved from live TV channels, real-time social media and news streams.

There is support for multiple languages and accessibility on a range of devices. *X-LiSA* supports 13 languages including English, German, French, Italian, Portuguese, Spanish, Russian, Chinese and Catalan, Serbian, Slovenian, Croatian and Basque. *XKnowSearch!* currently supports only 3 languages, namely English, German and Chinese, which will be extended in the near future.

References

1. Zhang, L., Rettinger, A.: X-lisa: Cross-lingual semantic annotation. PVLDB **7**(13) (2014) 1693–1696
2. Zhang, L., Färber, M., Rettinger, A.: xlid-lexica: Cross-lingual linked data lexica. In: LREC. (2014) 2101–2105
3. Zhang, L., Dong, Y., Rettinger, A.: Towards entity correctness, completeness and emergence for entity recognition. In: WWW (Companion Volume). (2015) 143–144