X-LiSA: Cross-lingual Semantic Annotation

Lei Zhang
Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
l.zhang@kit.edu

Achim Rettinger
Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
rettinger@kit.edu

ABSTRACT

The ever-increasing quantities of structured knowledge on the Web and the impending need of multilinguality and cross-linguality for information access pose new challenges but at the same time open up new opportunities for knowledge extraction research. In this regard, cross-lingual semantic annotation has emerged as a topic of major interest and it is essential to build tools that can link words and phrases in unstructured text in one language to resources in structured knowledge bases in any other language. In this paper, we demonstrate X-LiSA, an infrastructure for cross-lingual semantic annotation, which supports both service-oriented and user-oriented interfaces for annotating text documents and web pages in different languages using resources from Wikipedia and Linked Open Data (LOD).

1. INTRODUCTION

In recent years, large repositories of structured knowledge, such as Wikipedia and Linked Open Data (LOD) sources including DBpedia, Freebase and YAGO etc., have become valuable resources for knowledge extraction technologies, especially for the automatic aggregation of knowledge from textual data. One essential component, which leverages such knowledge bases, is the linking of words or phrases in natural language text with elements from the knowledge bases, which we call *semantic annotation*. At the same time, in order to achieve the goal that speakers of different languages have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual and cross-lingual access to information originally produced for a different culture and language. This poses new challenges to semantic annotation tools which typically are language dependent and link textual data in one language to a knowledge base grounded in the same language. Ultimately, the goal is to construct *cross-lingual semantic annotation* tools that can link words or phrases in unstructured text in one language to resources in structured knowledge bases in any other language.

Proceedings of the VLDB Endowment, Vol. 7, No. 13

Cross-lingual semantic annotation is beneficial for many applications. For example, in cross-lingual information retrieval (CLIR), it can help to better understand the documents and queries in different languages by bridging entity mentions in them with the entities in knowledge bases. There is a clear advantage to do this: it will be possible for a user to identify and explore the background knowledge of the searched items for refining the information needs [5]. Besides, there is a growing amount of research on question answering relying on the structured data in knowledge bases published on the Web. As the data in different languages published on the Web and the need for accessing this data using different native languages are growing substantially, it is crucial to have systems that allow users to express arbitrarily complex information needs as questions in different languages [3]. These systems should firstly leverage the cross-lingual semantic annotation approaches to map the queried resources to their counterparts in knowledge bases and then based on that retrieve the answer of the question.

The semantic annotation task is challenging due to the *mention variation* and *entity ambiguity* problems [6, 7]. Mention variation means that a resource in knowledge bases may have multiple surface forms, i.e., terms (including words and phrases) that can be used to refer to this resource, such that it can be mentioned in different ways. This problem is more serious in the cross-lingual setting because the surface forms of a resource can be in different languages. On the other hand, the entity ambiguity problem is due to the fact that one mention in different contexts can also refer to several resources, which might be grounded in different languages for cross-lingual semantic annotation.

With the goal of addressing the mention variation and entity ambiguity problems as well as overcoming language barriers between mentions and resources, we would like to demonstrate X-LiSA, which supports both service-oriented and user-oriented interfaces for annotating text documents and web pages in different languages with resources from Wikipedia and Linked Open Data (LOD).

2. OVERVIEW AND RELATED WORK

In this section, we first formulate the cross-lingual semantic annotation problem, then briefly review the related work.

DEFINITION 1. *Given a knowledge base KB containing a set of entities* $E_{KB} = \{e_1, e_2, \ldots, e_n\}$ *and the relations between them, each entity* e *in KB is characterized by its textual description* $e.c$ *in language* L*, called* context *of* e*. For a document* D *in language* L'*, let* $M = \{m_1, m_2, \ldots, m_p\}$ *denote a collection of entity mentions in* D*. Each entity mention* m *is characterized by its name* $m.s'$ *in* L' *as a surface*

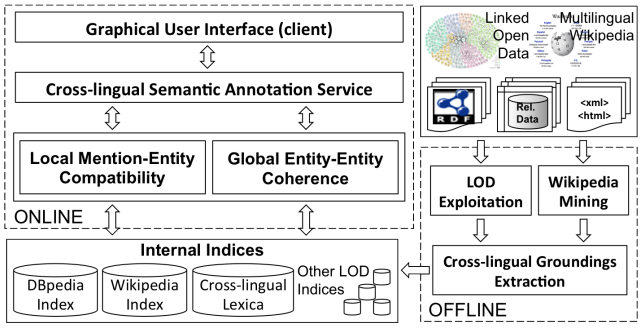


Figure 1: System Architecture.

form of the corresponding entity and its local surrounding sentences $m.c'$ in L', called context *of m. The objective of* cross-lingual semantic annotation *is to determine the referent entities in KB of the mentions in M.*

The task of semantic annotation is also called *entity linking* in the literature. Most systems in the monolingual setting employ the context similarity based approaches. The idea is to extract the discriminative features of an entity from its textual description, then link a mention in the document to the entity in the knowledge base with the largest similarity. The work in [4] proposes a Bag-of-Words (BOW) based method, which has been extended in recent work [6] by incorporating more features, such as popularity and name knowledge. However, in the cross-lingual setting, these approaches suffer from the vocabulary mismatch problem. To address this problem, we constructed a cross-lingual lexica for mention-entity matching and employed a concept-based approach for cross-lingual context similarity calculation to capture the *local mention-entity compatibility*.

There are also some recent approaches, which take the relations between candidate entities into account [8, 7], based on the assumption that the entities in the same document are related to each other. The drawback of such approaches is that either they don't take into account the other types of features extracted from entities or they are designed to leverage the relations as separate features, which results in the difficulty to incorporate the heterogeneous features. In this work, we propose to model the interdependence between different annotations as a disambiguation graph to capture the *global entity-entity coherence*, where the other evidences, such as the local mention-entity compatibility, can be easily incorporated into the model and collectively reinforced into high-probability annotation decisions based on the personalized PageRank algorithm, which has also been used for the word sense disambiguation problem [1].

3. DESCRIPTION OF X-LISA

In this section, we present the architecture of X-LiSA, as shown in Figure 1. While the *cross-lingual groundings extraction* is performed offline, the annotation is handled online based on the *local mention-entity compatibility* and the *global entity-entity coherence*.

3.1 Cross-lingual Groundings Extraction

Before we discuss the online steps, we first introduce the offline cross-lingual groundings extraction, where we constructed the cross-lingual linked data lexica, called xLiD-Lexica [10], by exploiting the multilingual Wikipedia to extract the cross-lingual groundings, namely the surface forms of entities in different languages.

As each Wikipedia article describes an entity in DBpedia, article titles, redirect pages and link anchors in Wikipedia can be used to refer to the entity. In addition, Wikipedia articles in different languages that provide information about the equivalent entity are often connected through the cross-language links. Based on the above sources, for each DBpedia entity grounded in one language, we extract its surface forms in other languages. Furthermore, we use the links between DBpedia and various other LOD data sources to derive cross-lingual groundings of entities from them.

We also exploit the statistics of the extracted cross-lingual groundings. Based on that, we define the probability $P(e|s')$ that surface form s' in language L' refers to entity e grounded in language L

$$P(e|s') = \frac{count_{link}(e, s')}{\sum_{e_i \in E_{s'}} count_{link}(e_i, s')} \quad (1)$$

where $count_{link}(e, s')$ denote the number of links using s' as anchor text pointing to e as destination and $E_{s'}$ is the set of entities that have the surface form s'.

In addition, we define the probability $P(s' \in M)$ that the surface form s' in a document is a mention name as

$$P(s' \in M) = \frac{count_{link}(s')}{count_{term}(s')} \quad (2)$$

where $count_{link}(s')$ denotes the number of articles that contain s' as anchor text and $count_{term}(s')$ denotes the number of articles where term s' appears.

3.2 Local Mention-Entity Compatibility

Now we discuss the online processing. Given a document in language L', we first gather all n-grams and match them against the surface forms to detect the possible entity mentions. For each mention m, we calculate the semantic similarity $SS(m, e)$ between m in language L' and its referent entity e grounded in language L as

$$SS(m, e) = \alpha \cdot LP(m, e) + \beta \cdot CS(m, e) \quad (3)$$

where $LP(m, e)$ is the link probability of e for m and $CS(m, e)$ is the context similarity between m and e, α and β are tunable parameters with $\alpha + \beta = 1$.

The link probability $LP(m, e)$ can be calculated using the probability $P(e|s')$ as discussed in Sec. 3.1. We have

$$LP(m, e) = P(e|m.s') = \frac{count_{link}(e, s')}{\sum_{e_i \in E_{s'}} count_{link}(e_i, s')} \quad (4)$$

Since the contexts of e and m, namely $e.c$ and $m.c'$, are in different languages, we cannot compute $CS(m, e)$ directly using Bag-of-Words (BOW) models due to the vocabulary mismatch problem. In this work, we employ a concept-based approach [9] by exploiting the interlingual concept space spanned by two sets of aligned concept vectors as

$$U = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \quad (5)$$

$$V = (\mathbf{v'}_1, \mathbf{v'}_2, \ldots, \mathbf{v'}_n) \quad (6)$$

where each pair of aligned vectors $\mathbf{u}_i$ for L and $\mathbf{v'}_i$ for L' represent the same concept. The term vectors $e.\mathbf{c}$ and $m.\mathbf{c'}$ for the contexts of e and m can be mapped to the concept vectors in the interlingual concept space as

$$U(e.\mathbf{c}) = (\langle \mathbf{u}_1, \mathbf{c} \rangle, \ldots, \langle \mathbf{u}_n, \mathbf{c} \rangle)^T \quad (7)$$

$$V(m.\mathbf{c'}) = (\langle \mathbf{v'}_1, \mathbf{c'} \rangle, \ldots, \langle \mathbf{v'}_n, \mathbf{c'} \rangle)^T \quad (8)$$

where each entry $\langle \mathbf{u}_i, \mathbf{c} \rangle$ in $U(e.\mathbf{c})$ ($\langle \mathbf{v'}_i, \mathbf{c'} \rangle$ in $V(m.\mathbf{c'})$) is the inner product of $\mathbf{u}_i$ and $\mathbf{c}$ ($\mathbf{v'}_i$ and $\mathbf{c'}$) representing the

association strength between them. Based on that, the context similarity $CS(m, e)$ between m and e can be calculated using the standard cosine similarity as

$$CS(m, e) = cos(U(e.\mathbf{c}), V(m.\mathbf{c}')) = \frac{\langle U(e.\mathbf{c}), V(m.\mathbf{c}')\rangle}{|U(e.\mathbf{c})| \cdot |V(m.\mathbf{c}')|} \tag{9}$$

In summary, the semantic similarity $SS(m, e)$ between m and e represents the *local mention-entity compatibility*, which captures the most likely entity behind the mention name and that best fits the context of the mention.

3.3 Global Entity-Entity Coherence

In this step, we construct a directed graph $G = \{N, E\}$, called *disambiguation graph*, where $N = N_M \uplus N_E$ is the disjoint union of *mention* nodes N_M and *entity* nodes N_E, and E is the set of directed edges. All detected mentions $M = \{m_1, ..., m_p\}$ and their candidate referent entities $E = \{e_1, ..., e_q\}$ are added into N_M and N_E, respectively. For each mention m and its candidate entity e, we add an edge from m to e into E. For each pair of entities e_i and e_j that are connected in KB, we add an edge between them into E, and the semantic relatedness $SR(e_i, e_j)$ between e_i and e_j is calculated based on Normalized Google Distance as

$$SR(e_i, e_j) = 1 - \frac{\log(\max(|E_i|, |E_j|)) - \log(|E_i| \cap |E_j|)}{\log(|N|) - \log(min(|E_i|, |E_j|))} \tag{10}$$

where E_i and E_j are the sets of entities that link to e_i and e_j in KB respectively, and N is the set of all entities in KB.

Once G is built, we apply the personalized PageRank algorithm over it. The calculation of the PageRank vector Pr over G is equivalent to resolving Eq. 11.

$$Pr = d \cdot T \cdot Pr + (1 - d) \cdot v \tag{11}$$

where T is a $(p+q) \times (p+q)$ transition probability matrix, v is $(p+q) \times 1$ vector and d is the so called damping factor. Each entry T_{ij} in T is the evidence propagation ratio from node i to node j, which is computed in Eq. 12.

$$T_{ij} = \begin{cases} \frac{SR(e_i, e_j)}{\sum_{k \in N_E(i)} SR(e_i, e_k)} & \text{if } i \in N_E,\ j \in N_E \\ \frac{SS(m_i, e_j)}{\sum_{k \in N_E(i)} SS(m_i, e_k)} & \text{if } i \in N_M,\ j \in N_E \end{cases} \tag{12}$$

where $N_E(i)$ is the set of entity nodes such that for each node $k \in N_E(i)$, there is an edge from i to k in G. The entry v_i in v is the initial evidence representing the prior importance of the mention m_i, which is calculated as

$$v_i = \begin{cases} \frac{P(m_i.s' \in M)}{\sum_{k \in N_M} P(m_k.s' \in M)} & \text{if } i \in N_M \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $P(m.s' \in M)$ represents the probability that the surface form $m.s'$ is a mention name, as discussed in Sec. 3.1. As a result of the personalized PageRank algorithm, each entity node receives a probability. Then, for each mention, we choose its candidate entity with the maximal probability.

To summarize, the intuition behind this step is that the nodes representing the correct referent entities of mentions in a document will be more relevant in the disambiguation graph (in the sense that they tend to be connected in KB) than the rest of the candidate entities, which should have less relations on average and be more isolated. As the mentions are connected to their candidate entities by directed edges, they act as source nodes injecting mass (probability) into the entities they are associated with (local mention-entity compatibility), which thus become relevant nodes, and spread their mass over the graph (global entity-entity coherence). Therefore, the resulting probability of entities can be seen as a measure that takes into account both local mention-entity compatibility and global entity-entity coherence.

4. DEMONSTRATION

X-LiSA is implemented using a client-server architecture with communication over HTTP using a XML schema defined in XLike project[1]. The server is a RESTful web service and the client user interface is implemented using Adobe Flex as both Desktop and Web Applications. The system can easily be extended or adapted to switch out the server or client. In this way, it supports both service-oriented and user-oriented interfaces for annotating multilingual text documents and web pages across the boundaries of languages.

First, we would like to demonstrate the extracted cross-lingual groundings described in Sec. 3.1. Using the Resrouce Description Framework (RDF)[2], we transform all the cross-lingual groundings into RDF triples each of which is composed of a sequence of (subject, predicate, object) terms. Based on that, we built a SPARQL endpoint over the resulting RDF data such that users can easily access the information using SPARQL query language[3]. The endpoint is provided based on OpenLink Virtuoso[4] as the back-end database engine. The RDF dataset used for this endpoint contains about 300 million triples of cross-lingual groundings. It is extracted from Wikipedia dumps[5] of July 2013 in English, German, Spanish, Catalan, Slovenian and Chinese, and based on the datasets of DBpedia 3.8[6].

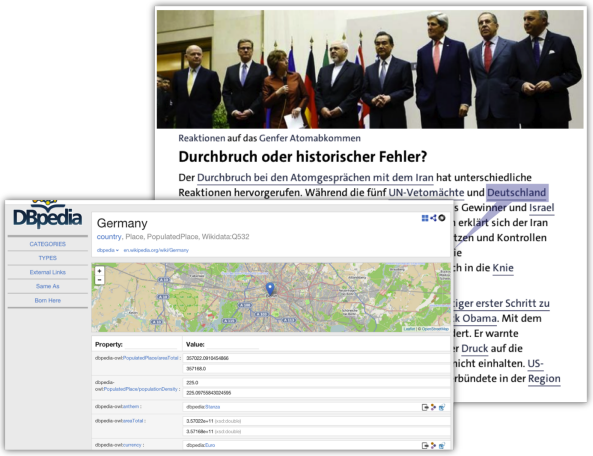
We now describe the functionality of our cross-lingual semantic annotation in X-LiSA, which supports interfaces for annotating text documents and web pages in different languages using resources from Wikipedia and LOD.

For annotating web pages, the input of the service includes the URL of the web page and the language of the content (input language) as well as the knowledge base and the language (output language) of its resources used for annotation, and the output is the web page with inserted annotations, which are linked to the corresponding resources in the output language. Figure 2a presents the screenshot of this service, where the input is the URL of a German news article, the used knowledge base is DBpedia and the output language is English.

For annotating text documents, the input of the service is the plain text contained in the document and the input language as well as the knowledge base and the output language used for annotation, and the output is the text document with highlighted annotations, which are linked to the corresponding KB resources in the output language. Figure 2b presents the screenshot of this service, where the input is a text document in English, the used knowledge base is Wikipedia and the output language is Chinese.

In order to allow not only users but also software agents to access the functionality of our cross-lingual semantic annotation, we also provide the service, which takes text documents and web pages as input and yields the output of annotations in XML as shown in Figure 2c.

[1] http://www.xlike.org/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/rdf-sparql-query/
[4] http://virtuoso.openlinksw.com/
[5] http://dumps.wikimedia.org/
[6] http://wiki.dbpedia.org/Downloads38

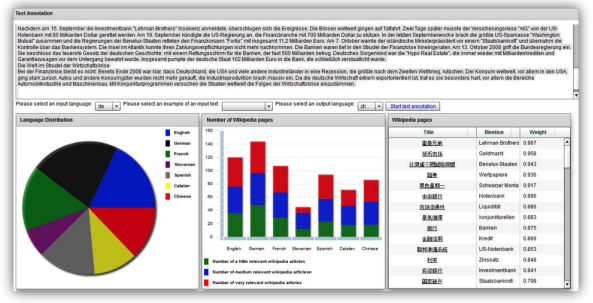


(a) Annotation service for web pages

(b) Annotation service for text documents

```
<?xml version="1.0" encoding="UTF-8"?>
<Response>
   <Source>Während die fünf UN-Vetomächte und Deutschland von einem wichtigen Erfolg
sprechen, sehen sich der Iran als Gewinner und Israel als Verlierer der Verhandlungen in
Genf.</Source>
   <Result>Während die fünf [[United_Nations_Security_Council_veto_power|UN-Vetomächte]]
und [[Germany|Deutschland]] von einem wichtigen Erfolg sprechen, sehen sich der Iran als
Gewinner und [[Israel]] als Verlierer der Verhandlungen in [[Geneva|Genf]].</Result>
   <Annotations>
      <Annotation URI="http://dbpedia.org/resource/Iran" lang="en" mention="Iran" />
      <Annotation
URI="http://dbpedia.org/resource/United_Nations_Security_Council_veto_power" lang="en"
mention="UN-Vetomächte" />
      <Annotation URI="http://dbpedia.org/resource/Germany" lang="en"
mention="Deutschland" />
      <Annotation URI="http://dbpedia.org/resource/Israel" lang="en" mention="Israel" />
      <Annotation URI="http://dbpedia.org/resource/Geneva" lang="en" mention="Genf" />
   </Annotations>
</Response>
```

(c) Annotation service output in XML



(d) Graphical user interface

Figure 2: X-LiSA Services and User Interface

A screenshot of the graphical user interface is shown in Figure 2d. It allows users to find the resources in knowledge bases mentioned in the input document. In the left pie chart, the users can see the percentage of resources in different languages as annotations of the input document. According to their weights, the resources in each language are organized in 3 relevance categories: high, medium and low. In the middle bar chart, the number of resources in each language and in each category is illustrated. The right data grid provides the links to the resources in the selected output language with their weights and the mentions in the input document. Clicking an individual link opens the corresponding resource in the knowledge base.

5. CONCLUSION

In this paper, we demonstrate X-LiSA, an infrastructure for cross-lingual semantic annotation developed within the XLike project. The services of X-LiSA have been widely used as components of the XLike pipeline architecture [2] for enabling cross-lingual processing for publishers, media monitoring and new business intelligence applications.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342) and FP7-ICT-2013-10 (XLiMe, Grant 611346).

7. REFERENCES

[1] E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *EACL*, pages 33–41, 2009.

[2] X. Carreras, L. Padró, L. Zhang, A. Rettinger, Z. Li, E. García-Cuesta, Ž. Agič, B. Bekavac, B. Fortuna, and T. Štajner. Xlike project language analysis services. In *EACL*, 2014.

[3] P. Cimiano, V. Lopez, C. Unger, E. Cabrio, A.-C. N. Ngomo, and S. Walter. Multilingual question answering over linked data (qald-3): Lab overview. In *CLEF*, pages 321–332, 2013.

[4] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716, 2007.

[5] M. Färber, L. Zhang, and A. Rettinger. Kuphi - an investigation tool for searching for and via semantic relations. In *ESWC*, 2014.

[6] X. Han and L. Sun. A generative entity-mention model for linking entities with knowledge base. In *ACL*, pages 945–954, 2011.

[7] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *SIGIR*, pages 765–774, 2011.

[8] D. N. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, pages 509–518, 2008.

[9] P. Sorg and P. Cimiano. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes of the Annual CLEF Meeting*, 2008.

[10] L. Zhang, M. Färber, and A. Rettinger. xlid-lexica: Cross-lingual linked data lexica. In *LREC*, pages 2101–2105, 2014.