xLiD-Lexica: Cross-lingual Linked Data Lexica

Lei Zhang, Michael Färber, Achim Rettinger

Institute AIFB, Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
{l.zhang, michael.farber, rettinger}@kit.edu

Abstract

In this paper, we introduce our cross-lingual linked data lexica, called xLiD-Lexica, which are constructed by exploiting the multilingual Wikipedia and linked data resources from Linked Open Data (LOD). We provide the cross-lingual groundings of linked data resources from LOD as RDF data, which can be easily integrated into the LOD data sources. In addition, we build a SPARQL endpoint over our xLiD-Lexica to allow users to easily access them using SPARQL query language. Multilingual and cross-lingual information access can be facilitated by the availability of such lexica, e.g., allowing for an easy mapping of natural language expressions in different languages to linked data resources from LOD. Many tasks in natural language processing, such as natural language generation, cross-lingual entity linking, text annotation and question answering, can benefit from our xLiD-Lexica.

1. Introduction

The ever-increasing quantities of semantic data on the Web pose new challenges but at the same time open up new opportunities of publishing and accessing information on the Web. The Semantic Web brings structures to the content on the Web, creating an environment where software agents can carry out sophisticated tasks for users (Berners-Lee et al., 2001). Over the past years, there is a growing amount of research on interaction paradigms that allow end users to profit from the expressive power of Semantic Web standards while at the same time hiding their complexity behind an intuitive and easy-to-use interface.

Linked Open Data (LOD)[1] is such a way of publishing semantic data that allows related data to be connected and enriched, so that different representations of the same content can be found, and links between related resources can be made to lower the barriers to linking data linked using other methods (Bizer et al., 2009a; Heath and Bizer, 2011). Currently, it is the best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using Resource Description Framework (RDF)[2]. Then SPARQL query language[3] can be used to express queries across diverse data sources, whenever the data is stored or viewed as RDF.

In addition, multilinguality and cross-linguality have emerged as issues of major interest for the Semantic Web community. 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 by facilitating multilingual and cross-lingual access to semantic data originally produced for a different culture and language. While LOD data sources allow you to make sophisticated queries, and to link other data sources on the Web to them, the cross-lingual information contained in LOD is rather rare. Nevertheless, it is essential to allow users to express arbitrarily information needs in their own language.

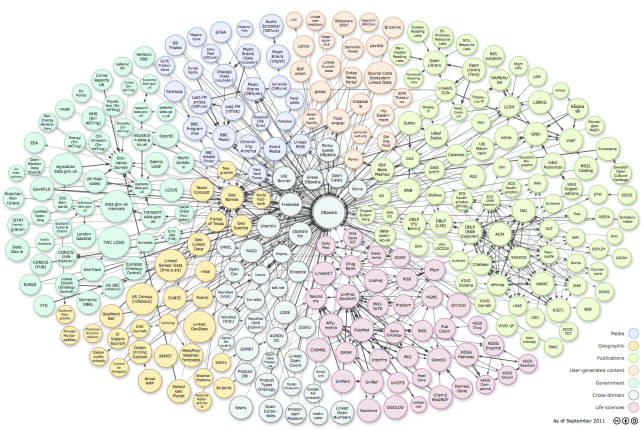

Figure 1: The Linked Open Data (LOD) Cloud. Each node stands for a single data source and each edge connecting two data sources represents the links between them.

As shown in Fig. 1, DBpedia[4], as a huge data source, stays in the center of the LOD cloud. It is a crowd-sourced community effort to extract structured information from Wikipedia in different languages and to make this information available on the Web (Auer et al., 2007; Bizer et al., 2009b). Although DBpedia is a large multilingual knowledge base (Mendes et al., 2012), the rich cross-lingual information contained in Wikipedia, which can be used in many tasks in natural language processing, are missing there. The goal of this paper is to bridge such gaps by extracting the cross-lingual groundings of linked data resources and integrating them into DBpedia and some other LOD data sources.

2. Cross-lingual Lexica Extraction

In this section, we describe the process for extracting the cross-lingual linked data lexica for DBpedia and some other LOD data sources based on the links between DBpedia and them. At first, we briefly introduce some useful information in Wikipedia and then discuss the extraction process as well as some examples of the extracted lexica.

[1] http://lod-cloud.net/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/rdf-sparql-query/
[4] http://dbpedia.org/

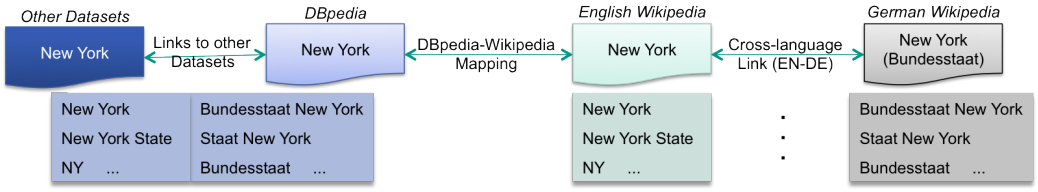

Figure 2: Cross-lingual linked data lexica extraction.

	English (EN)	German (DE)	Spanish (ES)	Chinese (ZH)
#Articles	4,014,643	1,438,325	896,691	509,197

(a) Number of articles

	EN-DE	EN-ES	DE-ES	EN-ZH	DE-ZH	ES-ZH
#Links (→)	721,878	568,210	295,415	224,056	139,135	126,734
#Links (←)	718,401	581,978	302,502	226,083	141,363	125,523
#Links (merged)	722,069	593,571	307,130	232,494	143,587	131,325

(b) Number of cross-language links

Table 1: Statistics about articles and cross-language links in Wikipedia

Wikipedia is the largest online encyclopedia up to date, which is an ever-growing source of manually defined resources and semantic relations between them contributed by millions of users over the Web. All of Wikipedia's content is presented on pages, such as articles and categories. Articles supply the bulk of Wikipedia's informative content, each of which describes a single resource. In addition, articles often contain links to equivalent articles in other language versions of Wikipedia. A wide range of applications can benefit from its multilingualism.

In addition, Wikipedia provides several elements that associate articles with terms, also called *surface forms*, that can be used to refer to the corresponding resources. Now we introduce these elements, which can be extracted using the Wikipedia-Miner toolkit (Milne and Witten, 2013):

- *Title of Wikipedia article*: The most obvious elements are article titles. Generally, the title of each article is the most common name for the resource described in this article, e.g., the article about the U.S. state New York has the title "New York".

- *Redirect page*: A redirect page exists for each alternative name which can be used to refer to a resource in Wikipedia. For example, the article titled "New York State", which is the full name of the U.S. state New York, is redirected to the article titled "New York". Redirect pages often indicate synonyms, abbreviations or other variations of the pointed resources.

- *Disambiguation page*: When multiple resources in Wikipedia could have the same name, a disambiguation page containing the references to those resources is usually created. For example, the disambiguation page for the name "New York" lists more than 30 associated resources that could have the same name of "New York" including the U.S. state New York as well as the film "New York, New York" and so on. These disambiguation pages are very useful in extracting abbreviations or other aliases of resources.

- *Anchor text of hyperlinks*: The article in Wikipedia often contains hyperlinks pointing to the pages of resources mentioned in this article. For example, there are anchor texts "NY" appearing more than 400 times in Wikipedia pointing to the article about the U.S. state New York. The anchor text of a link pointing to a page provides the most useful source of synonyms and other variations of the linked resource.

The process for extracting the cross-lingual groundings of linked data resources is shown in Fig. 2. We start from DBpedia resources, for each of which, we find the corresponding English Wikipedia article.

As mentioned, Wikipedia articles that provide information about the equivalent resources in different languages are connected through the cross-language links. For example, the English article "New York" links to "New York (Bundesstaat)" in the German Wikipedia, "Nueva York (estado)" in the Spanish Wikipedia and many others. In order to derive the cross-lingual lexica, we employ such cross-language links in Wikipedia to find the corresponding Wikipedia articles in different languages. Table 1 shows some statistics of the Wikipedia used in this work. We analysed cross-language links between Wikipedia articles for each pair of supported languages in both directions and keep only articles for which aligned versions exist at least in one direction. For instance, we have extracted $721,878$ cross-language links from English Wikipedia to German Wikipedia, and $718,401$ cross-language links from German to English. By merging them together, we obtain $722,069$ cross-language links, which are used to construct the Wikipedia comparable corpus of the English-German language pair.

Generally, article titles, redirect pages, disambiguation pages and link anchors are all considered as surface forms, i.e., terms (including words and phrases) that have been

```
<http://dbpedia.org/resource/New_York>  <http://xlid-lexica.org/block> _:bxyz .
_:bxyz <http://xlid-lexica.org/block#lang> "de" .
_:bxyz <http://xlid-lexica.org/res#linkDocCount> "8569"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/res#linkOccCount> "8936"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/res#probability> "0.006404"^^<http://www.w3.org/2001/XMLSchema#double> .
```

(a) Resource block.

```
<http://xlid-lexica.org/sf/Staatsregierung> <http://xlid-lexica.org/block> _:bxyz .
_:bxyz <http://xlid-lexica.org/block#lang> "de" .
_:bxyz <http://xlid-lexica.org/sf#label> "Staatsregierung"@de .
_:bxyz <http://xlid-lexica.org/sf#linkDocCount> "63"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/sf#linkOccCount> "65"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/sf#textDocCount> "1454"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/sf#textOccCount> "1816"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/sf#linkProbability> "0.043328"^^<http://www.w3.org/2001/XMLSchema#double> .
```

(b) Surface form block.

```
<http://dbpedia.org/resource/New_York>  <http://xlid-lexica.org/block> _:bxyz.
_:bxyz <http://xlid-lexica.org/block#lang> "de" .
_:bxyz <http://xlid-lexica.org/res#sf> <http://xlid-lexica.org/sf/Staatregierung> .
_:bxyz <http://xlid-lexica.org/res#senseLinkDocCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/res#senseLinkOccCount> "1"^^<http://www.w3.org/2001/XMLSchema#integer> .
_:bxyz <http://xlid-lexica.org/res#priorProbability> "0.015384"^^<http://www.w3.org/2001/XMLSchema#double> .
```

(c) Sense block.

Figure 3: Examples of RDF statements in N-Triples of our cross-lingual lexica.

used to refer to resources in some way. Based on the above sources, we extract surface forms of DBpedia resources in different languages from Wikipedia. Furthermore, we use the links between DBpedia and various other data sources, as shown in Fig. 1, to derive cross-lingual groundings of resources from other LOD data sources.

Besides the extracted surface forms in different languages, we also exploit statistics of the cross-lingual groundings to answer the following questions:

- How important is a resource in different languages?
- How important is a surface form in different languages?
- How strong is a surface form in different languages associated with a resource?

To address the first question w.r.t. a resource res, we investigate the number of links made to this resource, denoted as $res\#linkOccCount$, and the number of distinct articles containing links to this resource, denoted as $res\#linkDocCount$, in Wikipedia for a specific language. Based on that, we calculate the probability, denoted as $res\#probability$, that this resource res appears as links in Wikipedia either as

$$res\#probability = \frac{res\#linkOccCount}{totalLinkOccCount} \tag{1}$$

or as

$$res\#probability = \frac{res\#linkDocCount}{totalLinkDocCount} \tag{2}$$

where $totalLinkOccCount$ represents the total number of links appearing in Wikipedia and $totalLinkOccCount$ represents the total number of articles in Wikipedia. The difference between Eq. 1 and Eq. 2 lies in the granularity of link occurrence, namely at the hyperlink level or article level respectively. In our extracted lexica, we use the probability computed in Eq. 2, which can be considered as the importance indicator of a resource in different languages.

Regarding the second question w.r.t. a surface form sf, we investigate the number of links using this surface form as anchor text and the number of articles containing this surface form as anchor text, denoted as $sf\#linkOccCount$ and $sf\#linkDocCount$ respectively, and the number of times this surface form is mentioned and the number of articles mentioning this surface form (either as anchor text or in plain text), denoted as $sf\#textOccCount$ and $sf\#textDocCount$ respectively, in Wikipedia for a specific language. Similarly, we calculate the probability, denoted as $sf\#probability$, that this surface form sf used as anchor text either as

$$sf\#probability = \frac{sf\#linkOccCount}{sf\#textOccCount} \tag{3}$$

or as

$$sf\#probability = \frac{sf\#linkDocCount}{sf\#textDocCount} \tag{4}$$

We use the probability computed in Eq. 4 to indicate the importance of a surface form in different languages.

To answer the third question w.r.t. a resource res and a surface form sf, we investigate the number of links using this surface form as anchor text pointing to this resource as destination, denoted as $res\#senseLinkOccCount$, and the number of articles containing this surface form as anchor text pointing to this resource as destination, denoted as $res\#senseLinkDocCount$. Based on the above counts, we calculate the probability $res\#priorProbability$ that this surface form sf goes to this destination res either as

$$res\#priorProbability = \frac{res\#senseLinkOccCount}{sf\#linkOccCount} \tag{5}$$

or as

$$res\#priorProbability = \frac{res\#senseLinkDocCount}{sf\#linkDocCount} \tag{6}$$

In our lexica, we use the probability computed in Eq. 5 to measure the strength of the association between a surface form in different languages and a linked data resource.
In addition, we transform all the cross-lingual lexica described above into RDF triples[5] such that users can easily access such information using SPARQL query language, which will be discussed in Sec. 3. in detail. In the following, we use some examples shown in Fig. 3 to introduce the RDF schema used to encode the extracted cross-lingual lexica and their statistics. Firstly, the resource block describes the counts and probability of a resource appearing as a link in Wikipedia. An example of the DBpedia resource "New York" is shown in Fig. 3a. Secondly, the surface form block contains the counts and probability of a surface form in different languages appearing as anchor text in Wikipedia. An example of the German surface form "Staatsregierung" is shown in Fig. 3b. Finally, the sense block contains the counts and probability of a surface form in different languages referring to a resource. An example of the resource "New York" as the sense of the German surface form "Staatsregierung" is shown in Fig. 3c.

3. Querying Cross-lingual Lexica

In this section, we describe a SPARQL endpoint[6] we built over the RDF data described in Sec. 2., which provides cross-lingual grounds of resources from DBpedia and some other LOD data sources. The endpoint is provided based on OpenLink Virtuoso[7] 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 of July 2013 in English, German, Spanish, Catalan, Slovenian and Chinese, and based on the canonicalized datasets of DBpedia 3.8 containing triples extracted from the respective Wikipedia whose subject and object resource have an equivalent English article. In Fig. 4, we list some examples of SPARQL queries to show how to use the endpoint to retrieve different information.
Since many tasks can be facilitated by the availability of such cross-lingual lexica, we will introduce some general usages and the related work in the following. Given a linked data resource, we can retrieve its possible surface forms in different languages together with the corresponding confidence scores. This will help for natural language generation from RDF graph and SPARQL queries, also called RDF and SPARQL verbalization (Ell et al., 2012; Ngomo et al., 2013). In addition, given a surface form in any language, we provide the resources which this surface form refer to with the corresponding confidence scores. This will help for the tasks like cross-lingual entity linking (McNamee et al., 2011; Cassidy et al., 2012; Zhang et al., 2013b), text annotation (Zhang et al., 2013a) and question answering (Ko et al., 2010; Ferrández et al., 2011; Cimiano et al., 2013).
In particular, we have already used the cross-lingual lexica in some of our work. First, we used the lexica to build our cross-lingual semantic annotation system, where the lexica can be employed to detect mentions of the linked data resources in natural language texts in different languages (Zhang and Rettinger, 2014). Second, we used the lexica to address the challenge of matching keyword query in different languages against entities in the knowledge bases for the cross-lingual document retrieval problem (Zhang et al., 2014). Recently, we developed a gold standard resource, called RECSA, for evaluating cross-

```
Select ?resource, ?probability from <http://xlid-lexica.org>
where {
        ?resource <http://xlid-lexica.org/block> ?b1 .
        ?b1 <http://xlid-lexica.org/res#sf> ?sf .
        ?b1 <http://xlid-lexica.org/res#priorProbability> ?probability .
        ?sf <http://xlid-lexica.org/block> ?b2.
        ?b2 <http://xlid-lexica.org/sf#label> "Staatsregierung"@de .
}
order by DESC(?probability) limit 100
```

(a) Find the top-100 resources with the surface form "Staatsregierung" in German.

```
Select ?label, ?probability from <http://xlid-lexica.org>
where {
        <http://dbpedia.org/resource/New_York> <http://xlid-lexica.org/block> ?b1 .
        ?b1 <http://xlid-lexica.org/res#sf> ?sf .
        ?b1 <http://xlid-lexica.org/res#priorProbability> ?probability .
        ?sf <http://xlid-lexica.org/block> ?b2.
        ?b2 <http://xlid-lexica.org/sf#label> ?label .
        ?b2 <http://xlid-lexica.org/block#lang> "de" .
}
order by DESC(?probability) limit 100
```

(b) Find the top-100 surface forms for the resource "New York" in German.

```
Select ?sf, ?probability from <http://xlid-lexica.org>
where {
        ?sf <http://xlid-lexica.org/block> ?b.
        ?b <http://xlid-lexica.org/sf#label> "Staatsregierung"@de .
        ?b <http://xlid-lexica.org/sf#linkProbability> ?probability .
}
```

(c) Retrieve the link probability of the surface form "Staatsregierung" in German.

```
Select ?probability from <http://xlid-lexica.org>
where {
        <http://dbpedia.org/resource/New_York>  <http://xlid-lexica.org/block> ?b.
        ?b <http://xlid-lexica.org/res#probability> ?probability .
        ?b<http://xlid-lexica.org/block#lang> "de" .
}
```

(d) Retrieve the probability that the resource "New York" appears as links in German Wikipedia.

```
Select ?resource, ?label, ?prob from <http://xlid-lexica.org>
where {
        ?res <http://xlid-lexica.org/block> ?b1 .
        ?b1 <http://xlid-lexica.org/res#sf> ?sf .
        ?b1 <http://xlid-lexica.org/res#priorProbability> ?probability .
        ?sf <http://xlid-lexica.org/block> ?b2.
        ?b2 <http://xlid-lexica.org/sf#label> ?label .
        ?label bif:contains "MJ"
}
order by DESC(?probability) limit 100
```

(e) Find the top-100 resources with the surface forms containing the string "MJ".

Figure 4: Examples of SPARQL queries over the RDF data of the extracted cross-lingual lexica.

[5] http://km.aifb.kit.edu/resources/xlid-lexica.nt
[6] http://km.aifb.kit.edu/services/xlid-lexica/
[7] http://virtuoso.openlinksw.com/

lingual semantic annotation, where we compiled a hand-annotated parallel corpus of 300 news articles in three languages with cross-lingual semantic groundings to the English Wikipedia and DBPedia. For this, we first employed our cross-lingual lexica to provide the candidate annotations automatically, which were then manually verified and cleaned by human annotators (Rettinger et al., 2014).

4. Conclusion

With this paper we make access to our cross-lingual linked data lexica (xLiD-Lexica) freely available. The xLiD-Lexica are constructed from Wikipedia articles in different languages as well as linked data resources from LOD, such as DBpedia resources. Multilingual and cross-lingual information access can be facilitated by the availability of our xLiD-Lexica, e.g., allowing for an easy mapping of natural language expressions in different languages to linked data resources. In this regard, we provide the cross-lingual groundings of linked data resources as RDF data, which can be easily integrated into the LOD data sources. In addition, we build a SPARQL endpoint over our xLiD-Lexica to allow users to easily access them using SPARQL query language.

Acknowledgments.
The authors acknowledge the support of the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342) and of the German Federal Ministry of Education and Research (BMBF) under grant 02PJ1002 (SyncTech).

5. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, pages 722–735.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5):34–43, May.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009a. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009b. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165.

Taylor Cassidy, Heng Ji, Hongbo Deng, Jing Zheng, and Jiawei Han. 2012. Analysis and refinement of cross-lingual entity linking. In *CLEF*, pages 1–12.

Philipp Cimiano, Vanessa Lopez, Christina Unger, Elena Cabrio, Axel-Cyrille Ngonga Ngomo, and Sebastian Walter. 2013. Multilingual question answering over linked data (qald-3): Lab overview. In *CLEF*, pages 321–332.

Basil Ell, Denny Vrandecic, and Elena Simperl. 2012. Spartiqulation: Verbalizing sparql queries. In *Proceedings of the International Workshop on Interacting with Linked Data (ILD 2012), Extended Semantic Web Conference (ESWC)*. CEUR-WS.org, Mai.

Óscar Ferrández, Christian Spurk, Milen Kouylekov, Iustin Dornescu, Sergio Ferrández, Matteo Negri, Rubén Izquierdo, David Tomás, Constantin Orasan, Guenter Neumann, Bernardo Magnini, and José Luis Vicedo González. 2011. The qall-me framework: A specifiable-domain multilingual question answering architecture. *J. Web Sem.*, 9(2):137–145.

Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Jeongwoo Ko, Luo Si, Eric Nyberg, and Teruko Mitamura. 2010. Probabilistic models for answer-ranking in multilingual question-answering. *ACM Trans. Inf. Syst.*, 28(3).

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W. Oard, and David S. Doermann. 2011. Cross-language entity linking. In *IJCNLP*, pages 255–263.

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia: A multilingual cross-domain knowledge base. In *LREC*, pages 1813–1817.

David N. Milne and Ian H. Witten. 2013. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239.

Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. 2013. Sorry, i don't speak sparql: translating sparql queries into natural language. In *WWW*, pages 977–988.

Achim Rettinger, Lei Zhang, Daša Berović, Danijela Merkler, Matea Srebačić, and Marko Tadić. 2014. Recsa: Resource for evaluating cross-lingual semantic annotation. In *LREC*.

Lei Zhang and Achim Rettinger. 2014. Cross-lingual semantic annotation of text using linked open data. Technical report, Institut AIFB, KIT, http://people.aifb.kit.edu/lzh/papers/xlisa_tr.pdf.

Lei Zhang, Achim Rettinger, Michael Färber, and Marko Tadic. 2013a. A comparative evaluation of cross-lingual text annotation techniques. In *CLEF*, pages 124–135.

Tao Zhang, Kang Liu, and Jun Zhao. 2013b. Cross lingual entity linking with bilingual topic model. In *IJCAI*.

Lei Zhang, Michael Färber, and Achim Rettinger. 2014. Exploiting semantics throughout the cross-lingual document retrieval process. Technical report, Institut AIFB, KIT, http://people.aifb.kit.edu/lzh/papers/sempro_tr.pdf.