Bridging the Gap between Cross-lingual NLP and DBpedia by Exploiting Wikipedia

Lei Zhang, Achim Rettinger, and Steffen Thoma

Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany
{l.zhang, rettinger, steffen.thoma}@kit.edu

Abstract. DBpedia, as a large repository of structured knowledge, has become a valuable resource for language technologies. On the other hand, multilinguality and cross-linguality have emerged as issues of major interest. However, the deficiency in unstructured resources and cross-lingual structures for DBpedia hinders its more wide-spread application in natural language processing (NLP) tasks, especially in the cross-lingual context. In this paper, we introduce our NLP resources extracted from Wikipedia, which aim at bridging the gap between cross-lingual NLP and DBpedia. In order to achieve this, we exploited various kinds of elements and structures in Wikipedia to derive different associations between NLP elements and DBpedia resources.

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. Natural Language Processing (NLP) technologies can both, benefit from and support linked data repositories. For instance, NLP and information extraction often involve various resources when processing text from different domains. The resources in the linked data sources, such as DBpedia [1], can be easily utilized to assist NLP modules in a variety of tasks. On the other hand, NLP and information extraction technologies can help to grow these structured data sources by automatic extraction of information from text.

DBpedia, as a large data source, is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web. In recent years, DBpedia has become a valuable resource for language technologies. However, the information in DBpedia is mostly extracted from Wikipedia infoboxes, resulting in rich structured data, while the natural language texts in Wikipedia are to a large extent not exploited. The deficiency in natural language resources for DBpedia hinders its more wide-spread application in NLP tasks. On the other hand, multilinguality and cross-linguality have emerged as issues of major interest nowadays. Although DBpedia is a large multilingual knowledge base [2], the rich cross-lingual structures contained in Wikipedia are missing there.

The goal of this work is to bridge the gap between cross-lingual NLP and DBpedia by exploiting multilingual Wikipedia. Besides infoboxes, extracting

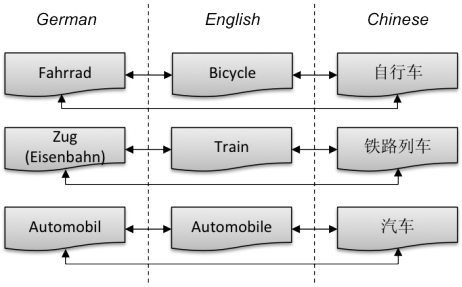

Fig. 1: Examples of interlingual resources in Wikipedia. The connecting arrows represent cross-language links.

additional information from the natural language text in Wikipedia and analyzing the semantics of its structures, such as redirect pages and anchor text of hyperlinks, would help to enrich DBpedia from the NLP perspective. Furthermore, Wikipedia currently supports approximately 280 languages and it also aligns articles in different languages that provide information about the same concept. Since a wide range of applications can benefit from its multilinguality and cross-linguality, it is essential to integrate the rich multilingual and cross-lingual information contained in Wikipedia into DBpedia, such that it is possible to leverage the huge amount of knowledge across languages.

The remainder of the paper is structured as follows: in Section 2, we present our extraction process based on various elements and structures in Wikipedia along with the dataset description in Section 3. Then we discuss related work in Section 4 before we conclude in Section 5.

2 Methodology

In this section, we firstly introduce some useful elements and structures in Wikipedia that can be employed to bridge the gap between cross-lingual NLP and DBpedia. Based on that, we then describe the extraction process for various associations between NLP elements in Wikipedia and DBpedia resources.

2.1 The Elements 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.

Interlingual Resources. Articles supply the bulk of informative content in Wikipedia. Each article describes a single resource and they often provide information about the equivalent resources in different language versions of Wikipedia. Based on the cross-lingual structure of Wikipedia, we define the *interlingual resource* using cross-language links, which are created by adding

references to corresponding articles in other languages to the source articles. Interlingual resources correspond to Wikipedia articles in different languages which are connected by the cross-language links. As shown in Figure 1, the interlingual resource $\langle Bicycle \rangle$ is defined by the English article *Bicycle* and all articles that are connected to this article via cross-language links, e.g., *Fahrrad* (German) and 自行车 (Chinese). In our example, for each interlingual resource, i.e., $\langle Bicycle \rangle$, $\langle Train \rangle$ and $\langle Bicycle \rangle$, there are three articles in English, German and Chinese that are fully connected across languages.

Labels. In addition, Wikipedia provides several structures that associate articles with terms (including words and phrases), also called *surface forms* or *labels* that can be used to refer to the corresponding resources. Now we introduce these elements, which can be extracted using the Wikipedia-Miner toolkit [3]:

- *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. For example, the English article about the resource $\langle Bicycle \rangle$ has the title "Bicycle", and the corresponding German and Chinese articles have the titles "Fahrrad" and "自行车", respectively.
- *Redirect page*: A redirect page exists for each alternative name, which can be used to refer to a resource in Wikipedia. For example, the articles titled "Pedal cycle" in English, "Stahlross" in German and "脚踏车" in Chinese are redirected to the articles titled "Bicycle", "Fahrrad" and "自行车", respectively. Thus, all these terms can be used to represent the resource $\langle Bicycle \rangle$. Redirect pages often indicate synonyms, abbreviations or other variations of the pointed resources.
- *Anchor text of hyperlinks*: The articles in Wikipedia often contain hyperlinks pointing to the pages of resources mentioned in the articles. For example, there are anchor texts "bike" appearing 50 times in English Wikipedia, "Rad" appearing 8 times in German Wikipedia and "单车" appearing 204 times in Chinese Wikipedia pointing to the articles about the resource $\langle Bicycle \rangle$. The anchor text of a link pointing to a page provides the most useful source of synonyms and other variations of the linked resource.

Words. Besides the structural elements, Wikipedia, as an extensive multilingual corpus, also provides plain text – that is, the full content of Wikipedia articles covering a wide range of topics, such as, but not limited to, arts, history, events, geography, mathematics and technology – in a vast amount with regard to the number of pages per language. Many NLP tasks can benefit from such unstructured resources, especially *words*.

2.2 Extracting NLP Data for DBpedia

In the following, we briefly describe the DBpedia resources and then discuss the extraction process for various associations between NLP elements in Wikipedia and DBpedia resources.

DBpedia Resources. The data in DBpedia is mainly extracted from structured information in Wikipedia editions in multiple languages. For each

Wikipedia page, there exists a Uniform Resource Identifier (URI) (henceforth also called *DBpedia resource*) describing the corresponding Wikipedia page in DBpedia. For example, the DBpedia resources `dbpedia:Bicycle`, `dbpedia-de:Fahrrad` and `dbpedia-zh:`自行车[1] correspond to the Wikipedia articles *Bicycle*, *Fahrrad* and 自行车, respectively. Moreover, each DBpedia resource can be aligned with an interlingual resource in Wikipedia and thus its corresponding articles in different languages[2]. In the above example, each of the mentioned DBpedia resources captured in one language can be mapped to the interlingual resource $\langle Bicycle \rangle$ and thus connected with all the three Wikipedia articles in English, German and Chinese. Similarly, through the interlingual resource as a hub, each Wikipedia article in one language can be mapped to all the DBpedia resources captured in different languages.

Label and Resource Reference Associations. We now discuss the reference association between labels and DBpedia resources. On the one hand, labels encode synonomy, because a resource could be represented by many labels, even in different languages. For example, the resource $\langle Bicycle \rangle$ can be denoted by the labels "bicycle" and "bike" in English, "Fahrrad" and "Rad" in German, "自行车" and"单车" in Chinese. On the other hand, labels also encode polysemy, because a label could refer to multiple resources. For example, the label "bike" can stand for both interlingual resources $\langle Bicycle \rangle$ and $\langle Motorcycle \rangle$ and thus can represent DBpedia resources in different languages, such as `dbpedia:Bicycle`, `dbpedia-de:Fahrrad`, `dbpedia-zh:`自行车, and `dbpedia:Motorcycle`, `dbpedia-de:Motorrad`, `dbpedia-zh:`摩托车.

Because all the labels are extracted from Wikipedia articles, the associated usage statistics can be mined for deriving the relationship between labels and resources. For example, the label "bike" refers to the resource $\langle Bicycle \rangle$ 50 times and to $\langle Motorcycle \rangle$ only 10 times such that "bike" is more likely to refer to $\langle Bicycle \rangle$. Based on the above observations , we define the probability $P(r|l)$ to model the likelihood that label l refers to resource r as

$$P(r|l) = \frac{count_{\text{link}}(r, l)}{\sum_{r_i \in R_l} count_{\text{link}}(r_i, l)} \tag{1}$$

where $count_{\text{link}}(r, l)$ denotes the number of links using l as label pointing to r as destination and R_l is the set of resources having label l. Then, we can semantically represent each term matching a label l as a weighted vector of its referent DBpedia resources r according to the weight $P(r|l)$.

In addition to $P(r|l)$, we also define the probability $P(l|r)$ to model the likelihood of observing l as label given resource r as

$$P(l|r) = \frac{count_{\text{link}}(r, l)}{\sum_{l_i \in L_r} count_{\text{link}}(r, l_i)} \tag{2}$$

[1] We use `dbpedia` for `http://dbpedia.org/resource/`, `dbpedia-de` for `http://de.dbpedia.org/resource/` and `dbpedia-zh` for `http://zh.dbpedia.org/resource/`.

[2] In this work, we use the terms *resource*, *interlingual resource* and *DBpedia resource* interchangeably, since they can be easily mapped to each other.

where L_r is the set of labels that can refer to resource r. Given a DBpedia resource r, we can therefore retrieve all terms, which are used as labels to refer to r in Wikipedia, together with the weights $P(l|r)$.

To calculate the strength w.r.t. the reference association of a pair of label l and resource r, the probability $P(r|l)$ and $P(l|r)$ are further processed to generate the point-wise mutual information (PMI) of l and r, defined as

$$PMI(l,r) = \log \frac{P(l,r)}{P(l)P(r)} = \log \frac{P(l|r)}{P(l)} = \log \frac{P(r|l)}{P(r)} \tag{3}$$

We define the probability $P(l)$ that label l appears as links in Wikipedia, no matter which resources it refers to, as

$$P(l) = \frac{\sum_{r_i \in R_l} count_{\text{link}}(r_i, l)}{N_{\text{link}}} \tag{4}$$

where N_{link} denotes the total number of links in Wikipedia. Similarly, we define the probability $P(r)$ that resource r is linked/referred to in Wikipedia regardless of the used label, as

$$P(r) = \frac{\sum_{l_i \in L_r} count_{\text{link}}(r, l_i)}{N_{\text{link}}} \tag{5}$$

According to Equation 3, 4 and 5, we derive the strength w.r.t. reference association of a pair of label l and resource r as follows

$$PMI(l,r) = \log \frac{count_{\text{link}}(r,l) \times N_{\text{link}}}{\sum_{r_i \in R_l} count_{\text{link}}(r_i, l) \times \sum_{l_i \in L_r} count_{\text{link}}(r, l_i)} \tag{6}$$

In terms of $P(r|l)$, $P(l|r)$ and $PMI(l,r)$, it is observed that the main difference between them lies in the normalization factor in the denominator of Equation 1, 2 and 6, respectively. Two terms, namely $\sum_{r_i \in R_l} count_{\text{link}}(r_i, l)$ standing for the frequency that label l appears as links and $\sum_{l_i \in L_r} count_{\text{link}}(r, l_i)$ denoting the frequency that resource r is linked/referred to in Wikipedia, are involved. The normalization factor of $P(r|l)$, i.e., $\sum_{r_i \in R_l} count_{\text{link}}(r_i, l)$, does not affect the ranking of r when l is specified (since the probabilities for different r are divided by the same term). Similarly, the normalization factor of $P(l|r)$, i.e., $\sum_{l_i \in L_r} count_{\text{link}}(r, l_i)$, does not affect the ranking of l when r is specified.

The normalization factor of $PMI(l,r)$ differs from $P(r|l)$ and $P(l|r)$ in involving both terms. It correlates to the inverse of frequency that the label l and the resource r used as links. This means that the labels and resources more rarely linked give higher contribution to the total association strength, which is similar to the inverse document frequency in IR. In this way, $PMI(l,r)$ attempts to make the association strength for different pairs of l and r comparable based on not only the correlation between l and r (represented by $count_{\text{link}}(r,l)$) but also their individual discriminability (represented by $\sum_{r_i \in R_l} count_{\text{link}}(r_i, l)$ and $\sum_{l_i \in L_r} count_{\text{link}}(r, l_i)$). Based on this guide, we can choose among $P(r|l)$, $P(r|l)$ and $PMI(l,r)$ for the particular tasks at hand.

Label and Resource Co-occurrence Associations. The reference association of a pair of label l and resource r captures the relationship in the sense that to which extent l refers to r and thus r is an intended sense of l. Besides that, we also utilize labels that frequently co-occur with a resource in its immediate context to derive co-occurrence association between labels and DBpedia resources, since such labels and resources are highly relevant.

The link structure in Wikipedia allows us to determine the labels within the context of a resource (defined by a sliding window of k sentences). To illustrate, let us consider the following paragraphs extracted from three Wikipedia articles in English, German and Chinese, which are all related to the resource $\langle Bicycle \rangle$.

*Example 1 (*Cycling*).* Cycling, also called bicycling or biking, is the use of bicycles for transport, recreation, or for sport. Persons engaged in cycling are referred to as "cyclists", "bikers", or less commonly, as "bicyclists". Apart from two-wheeled bicycles, "cycling" also includes the riding of unicycles, tricycles, quadracycles, and similar human-powered vehicles (HPVs).

*Example 2 (*Fahrradfahren*).* Der Ausdruck Fahrradfahren, auch Radfahren, bezeichnet die Fortbewegung auf einem Fahrrad. Er bezeichnet auch die Sportart Fahrradfahren, die als Freizeitbeschäftigung oder als sportlicher Wettkampf bis hin zum Leistungssport betrieben wird.

*Example 3 (*自行车运动*).* 自行车运动常指借助自行车（或称单车）开展的各种运动的总称，属于借助人力推动的半机械化运动，极少使用单轮车、三轮车、四轮车或其他用于运输、娱乐或运动的人力车辆开展此项运动。自行车运动在公路或小道上进行，根据不同的环境和要求开展此项活动，如自行车旅行、越野自行车运动、雪地自行车运动等等。

All the underlined words and phrases represent labels on the one hand, and represent links to the corresponding Wikipedia articles and thus the aligned resources on the other hand. In this way, each label can be semantically interpreted as a weighted vector of its neighboring linked resources and each resource can be treated as a weighted vector of its neighboring labels in different languages. For example, the label *human-powered vehicles* can be represented as a vector of the interlingual resources $\langle Bicycle \rangle$, $\langle Transport \rangle$, $\langle Recreation \rangle$, $\langle Sport \rangle$, $\langle Unicycle \rangle$, $\langle Tricycle \rangle$ and $\langle Quadracycle \rangle$ and thus also as a vector of corresponding DBpedia resources captured in any supported language. And the interlingual resource $\langle Bicycle \rangle$ and each corresponding DBpedia resource captured in one language can be represented as a vector of the labels *transport*, *recreation*, *sport*, *unicycles*, *tricycles*, *quadracycles* and *human-powered vehicles* in English, *Sportart Fahrradfahren* and *Leistungssport* in German, 运动, 单轮车, 三轮车 and 四轮车 in Chinese.

Next, we discuss the weights of the resources as interpretations of a label. For this, we define the probability $P_k(r|l)$ to model the likelihood that given a label l, the resource r co-occur with it in a window of k sentences as

$$P_k(r|l) = \frac{count_{\text{co-occur}}(r, l)}{\sum_{r_i \in R_l} count_{\text{co-occur}}(r_i, l)} \tag{7}$$

	English	German	Spanish	Catalan	Slovenian	Chinese
$\#Labels$	13M	4.6M	3.2M	0.9M	0.3M	1.3M
$\#Words$	2.6B	908M	666M	224M	48M	321M

Table 1: Statistics about words and labels in Wikipedia.

where $count_{\text{co-occur}}(r, l)$ denotes the frequency that l and r co-occur in a window of k sentences and R_l is the set of resources that co-occur with label l.

Then, we discuss the weights of the relevant labels given a resource. For this, we define the probability $P_k(l|r)$ to model the likelihood of l appearing in the context of resource r with size k as

$$P_k(l|r) = \frac{count_{\text{co-occur}}(r, l)}{\sum_{l_i \in L_r} count_{\text{co-occur}}(r, l_i)} \tag{8}$$

where L_r is the set of labels that co-occur with resource r.

Similarly, we calculate the strength w.r.t. the co-occurrence association of a pair of label l and resource r based on $P_k(r|l)$ and $P_k(l|r)$ as

$$PMI_k(l, r) = \log \frac{count_{\text{co-occur}}(r, l) \times N^{label}_{\text{co-occur}}}{\sum_{r_i \in R_l} count_{\text{co-occur}}(r_i, l) \times \sum_{l_i \in L_r} count_{\text{co-occur}}(r, l_i)} \tag{9}$$

where $N^{label}_{\text{co-occur}}$ is the sum of the frequency that label l and resource r co-occur in a window of k sentences for all pairs of l and r. The difference between $P_k(r|l)$, $P_k(r|l)$ and $PMI_k(l, r)$ w.r.t. co-occurrence association between labels and resources is similar to the reference association as discussed before.

Word and Resource Co-occurrence Associations. Apart from labels, there are many more words contained in Wikipedia for different languages, which can play an important role in NLP. The biggest advantage of word-based NLP approaches is the large amount of available data, such that they are not subject to data sparsity issues. As shown in Table 1, the number of words in Wikipedia significantly exceeds the number of extracted labels for the different languages. For example, the English Wikipedia alone contains over 2.6 billion words, over 100 times as many as the next largest English-language encyclopedia, Encyclopaedia Britannica, while it has only 13 million labels. Therefore, it is also crucial to derive the co-occurrence association between words and DBpedia resources for bridging the gap between NLP and DBpedia.

First, we define the probability $P_k(r|w)$ to model the likelihood that given a word w, the resource r co-occur with it in a window of k sentences as

$$P_k(r|w) = \frac{count_{\text{co-occur}}(r, w)}{\sum_{r_i \in R_w} count_{\text{co-occur}}(r_i, w)} \tag{10}$$

where $count_{\text{co-occur}}(r, w)$ denotes the frequency that word w and resource r co-occur in a window of k sentences and R_w is the set of resources that co-occur with word w. For each word w in one language, we can derive a vector of weighted co-occurred DBpedia resources r with the weight $P_k(r|w)$. In Example 1, the word *bicycling* can be represented as a weighted vector of the interlingual resources $\langle Bicycle \rangle$, $\langle Transport \rangle$, $\langle Recreation \rangle$, $\langle Sport \rangle$, $\langle Unicycle \rangle$, $\langle Tricycle \rangle$,

	Our Datasets			DBpedia NLP Datasets	
	#Label Resource Reference Associations	*#Label Resource Co-occurrence Associations*	*#Word Resource Co-occurrence Associations*	*#DBpedia Lexicalizations Entries*	*#DBpedia Topic Signatures Entries*
English	15,237,596	104,560,077	313,266,917	2,176,869	8,438,400
German	5,342,851	42,316,145	172,033,719	–	–
Spanish	3,563,379	34,404,641	106,951,335	–	–
Catalan	1,022,815	8,161,564	29,753,250	–	–
Slovenian	380,522	26,638,003	25,249,677	–	–
Chinese	1,425,827	16,286,187	19,851,666	–	–
Total	26,972,990	232,366,617	667,106,564	2,176,869	8,438,400

Table 2: Statistics of our datasets and DBpedia NLP Datasets.

$\langle Quadracycle \rangle$ and $\langle Human\text{-}powered\ transport \rangle$ and also the corresponding DBpedia resources captured in any supported language.

Next, we define the probability $P_k(w|r)$ to model the likelihood of word w appearing in the context of resource r with size k as

$$P_k(w|r) = \frac{count_{\text{co-occur}}(r, w)}{\sum_{w_i \in W_r} count_{\text{co-occur}}(r, w_i)} \tag{11}$$

where W_r is the set of words that co-occur with resource r. For each resource, a vector of words w appearing in the context of r with weights $P_k(w|r)$ can be generated. In the previous examples, regarding the resource $\langle Bicycle \rangle$ and all corresponding DBpedia resources captured in different languages, we can generate a vector of the words, such as *cycling* and *cyclists* in English, *Radfahren* and *Freizeitbeschäftigung,* in German, 半机械化 and 运输 in Chinese.

Finally, we calculate the strength w.r.t. the co-occurrence association of a pair of word w and resource r based on $P_k(r|w)$ and $P_k(w|r)$ as

$$PMI_k(w, r) = \log \frac{count_{\text{co-occur}}(r, w) \times N^{word}_{\text{co-occur}}}{\sum_{r_i \in R_l} count_{\text{co-occur}}(r_i, w) \times \sum_{w_i \in W_r} count_{\text{co-occur}}(r, w_i)} \tag{12}$$

where $N^{word}_{\text{co-occur}}$ is the sum of the frequency that word w and resource r co-occur in a window of k sentences for all pairs of w and r. We omit the discussion about the difference between $P_k(r|w)$, $P_k(r|w)$ and $PMI_k(w, r)$, because it is similar to that between $P(r|l)$, $P(r|l)$ and $PMI(l, r)$ as discussed before.

3 Datasets

In this section, we describe our datasets extracted based on the methodology presented in Section 2, where we used the Wikipedia dumps of July 2013 in English, German, Spanish, Catalan, Slovenian and Chinese.

Table 2 provides the main statistics of our datasets w.r.t. the three associations, namely label resource reference, label resource co-occurrence and word resource co-occurrence associations. In order to compare our datasets with the most related work, Table 2 also provides the statistics of DBpedia NLP Datasets[3], where the Lexicalization dataset contains the information similar to

[3] `http://wiki.dbpedia.org/Datasets`

Our Datasets	English	German	Chinese	DBpedia NLP Datasets	English
Label	Michael Jordan	Michael Jordan	迈克尔·乔丹		Michael Jordan
Resource	Jordan	Jordan	麥可·喬丹	*DBpedia*	Jordan
Reference	Air Jordan	Air Jordan	麥可·喬登	*Lexicalizations*	MJ
Association	His Airness	His Airness	米高·佐敦	*Dataset*	–
Dataset	MJ23	Jordan, Michael	邁克爾·喬丹		–
Label	Scottie Pippen	Chicago Bulls	波士特人		
Resource	Dennis Rodman	NBA	洛杉矶湖人		
Co-occurrence	Chicago Bulls	Basketball	城76人		
Association	United Center	Scottie Pippen	芝加哥公牛		
Dataset	NBA	San Antonio Spurs	圣安东尼奥马刺		
Word	nba	bulls	洛杉矶	*DBpedia*	game
Resource	basketball	chicago	凯尔特人	*Topic*	nba
Co-occurrence	bulls	spieler	波士顿	*Signatures*	team
Association	chicago	nba	芝加哥	*Dataset*	–
Dataset	game	basketballspieler	萨克拉门托		–

Table 3: Examples of top-5 results from our datasets and DBpedia NLP datasets for resource `dbpedia:Michael_Jordan`.

our label and resource reference associations. In the Topic Signatures dataset each DBpedia resource is represented by a term vector (of size 3 in most cases) extracted from Wikipedia article content using TF-IDF weights [2]. It is observed that our datasets contain more entries than the DBpedia NLP Datasets and provide information in more languages. Table 3 shows the top-5 results of different associations for the example resource `dbpedia:Michael_Jordan` from our datasets and the DBpedia NLP datasets. This conveys the impression that we achieve better results in terms of quantity and quality compared with the DBpedia NLP datasets.

Dataset Dumps. The first version of our datasets is available[4] as plain text files in JSON format. These files consist of a list of records, each identifying an association between an NLP element and a resource. An example of label resource reference association between the label *MJ23* and the resource *Michael Jordan* is shown as follows:

```
{
        "id": ObjectId("53f0cfdfe4b0e7085cf241a1"),
        "label": "MJ23",
        "resource": "Michael Jordan",
        "P(r|l)": "1",
        "P(l|r)": "0.0007199424046076314",
        "PMI(l,r)": "11.102683968056724"
}
```

Accessing API and GUI. In order to effectively access and automatically embed our datasets within applications, we developed a Java API, based on MongoDB[5] as backend, such that the dataset dumps in JSON format can be easily imported into MongoDB. The API provides a variety of methods to access different kinds of information, namely (1) the reference and co-occurrence associations with labels and words given a resource; (2) the reference and co-occurrence associations with resources given a label; (3) the co-occurrence

[4] `https://people.aifb.kit.edu/lzh/nlp/`

[5] `http://www.mongodb.org/`

associations with resources given a word. In addition, we ship the API with a graphical user interface (GUI) that allows the user to browse our datasets. The accessing API and the GUI are accessible as open source on GitHub[6].

4 Related Work

In this section, we review the related work and discuss our contributions from two perspectives, namely dictionary datasets and lexical knowledge bases.

4.1 Dictionary Datasets

Dictionaries contain associations that map labels to DBpedia resources as their senses, which can be applied to many applications, such as Named Entity Disambiguation [4]. Now we will discuss some dictionaries in the following.

The work closest to ours is the DBpedia NLP datasets [2], which describe a number of extended resources for DBpedia that specifically aim at supporting computational linguistics tasks, where the Lexicalizations dataset contains the information similar to that captured by our label and resource reference association. Overall, there are 2 million entries of English labels and resources in the dictionary, where for each label-resource pair, the probabilities $P(r|l)$, $P(l|r)$ and the pointwise mutual information $PMI(l, r)$ are given.

The Crosswikis dictionary [5] is a similar, but much larger dataset for English Wikipedia concepts. It has been built at web scale and includes 378 million entries. Similar to the DBpedia Lexicalizations dataset, the probabilities $P(r|l)$ and $P(l|r)$ have been calculated and is available in the dictionary.

The means relation of YAGO[7] has been used as dictionary by AIDA [6], a tool for disambiguation of named entities in text, to identify candidate entities for a (possible ambiguous) mention. The entries in the dictionary were extracted from link anchors, disambiguation pages and redirection links in Wikipedia.

Similar to the YAGO means relation, the Redirect Disambiguation Mapping (RDM) dictionary has been constructed by solving disambiguation pages and redirects and using these alternative labels additionally to the original labels of the DBpedia entities. This dictionary has been compared with other datasets in the context of Named Entity Disambiguation tasks in [4].

Recently, we built the cross-lingual linked data lexica, called xLiD-Lexica [7], to provide the cross-lingual groundings of DBpedia resources that have been used in our cross-lingual semantic annotation system [8]. We stored the dataset as RDF data in the N-Triples format and built a SPARQL endpoint for allowing easy user access by using the SPARQL query language.

While the use of Wikipedia for extracting reference associations between labels and DBpedia resources is not new, our work is different in that besides reference associations we also study the co-occurrence associations between

[6] `https://github.com/beyondlei/nlp-lexica`
[7] `http://www.yago-knowledge.org/`

different NLP elements, namely labels and words, and DBpedia resources. In addition, we provide both reference and co-occurrence associations in the cross-lingual setting by extracting labels and words from Wikipedia editions in multiple languages and exploiting cross-lingual structures of Wikipedia.

4.2 Lexical Knowledge Bases

In the past few years, there has been a growing interest in extracting knowledge from Wikipedia and other knowledge sources such as WordNet, for constructing multilingual lexical knowledge bases. In the following, we introduce several state-of-the-art lexical knowledge bases.

WikiNet [9] is a multilingual semantic network constructed from Wikipedia and includes semantic relations between Wikipedia entities, which are collected from the category structure, infoboxes and article contents.

UWN [10] is an automatically constructed multilingual lexical knowledge base, which is bootstrapped from WordNet and built by collecting evidence extracted from existing wordnets, translation dictionaries and parallel corpora. This results in over 800,000 words in over 200 languages in a semantic network with over 1.5 million links from words to word senses. Its extension MENTA [11] adds a large scale hierarchical taxonomy containing 5.4 million named entities and their classes, which is also built from WordNet and Wikipedia.

Similarly to UWN and MENTA, BabelNet [12] integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia into a unified, wide-coverage, multilingual lexical knowledge base through a novel mapping algorithm that can establish the mappings between a multilingual encyclopedic knowledge repository (Wikipedia) and a computational lexicon of English (WordNet) with high accuracy. In general, BabelNet is a multilingual semantic network, which connects concepts and named entities in a very large network of semantic relations, made up of more than 9 million entries, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages.

These lexical knowledge bases go one step beyond the dictionary datasets by integrating semi-structured information from Wikipedia with the relational structure of other knowledge sources into a semantic network to provide the meanings of words and phrases and to show how such meanings are semantically related based on their semantic relations. In addition to the senses, our co-occurrence associations provide complementary information about the relatedness of words and labels with DBpedia resources in a multilingual and cross-lingual setting. In this way, each word or label in any language can be represented as a vector of DBpedia resources and vice versa, which can be applied to many applications such as cross-lingual semantic relatedness [13].

5 Conclusions

In this paper, we presented our approach to extracting NLP resources from the multilingual Wikipedia to enrich DBpedia, which aim at bridging the gap

between cross-lingual NLP and DBpedia. In order to achieve this, we exploited various kinds of elements and structures, such as anchor text of hyperlinks and cross-language links, in Wikipedia to derive different associations between NLP elements extracted from Wikipedia editions in multiple languages and DBpedia resources. It is hard to verify the quality of the extracted datasets directly as we can not find a test collection. However, it is possible to validate the effectiveness of the proposed methods in a certain kind of application such as cross-lingual entity linking and cross-lingual semantic relatedness, which are considered as our future work.

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).

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. J. Web Sem. **7**(3) (2009) 154–165
2. Mendes, P.N., Jakob, M., Bizer, C.: Dbpedia: A multilingual cross-domain knowledge base. In: LREC. (2012) 1813–1817
3. Milne, D.N., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artif. Intell. **194** (2013) 222–239
4. Steinmetz, N., Knuth, M., Sack, H.: Statistical analyses of named entity disambiguation benchmarks. In: NLP-DBPEDIA@ISWC. (2013)
5. Spitkovsky, V.I., Chang, A.X.: A cross-lingual dictionary for english wikipedia concepts. In: LREC. (2012) 3168–3175
6. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP. (2011) 782–792
7. Zhang, L., Färber, M., Rettinger, A.: xlid-lexica: Cross-lingual linked data lexica. In: LREC. (2014) 2101–2105
8. Zhang, L., Rettinger, A.: X-lisa: Cross-lingual semantic annotation. PVLDB **7**(13) (2014) 1693–1696
9. Nastase, V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A.: Wikinet: A very large scale multi-lingual concept network. In: LREC. (2010)
10. de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: CIKM. (2009) 513–522
11. de Melo, G., Weikum, G.: Menta: inducing multilingual taxonomies from wikipedia. In: CIKM. (2010) 1099–1108
12. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193** (2012) 217–250
13. Hassan, S., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: AAAI. (2011)