Building Large-Scale Cross-lingual Knowledge Base from Multi-Encyclopedia

Mingyang Li[†], Yao Shi[†], and Zhigang Wang[†]

[†]Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
`{lmy13@}tsinghua.edu.cn`
`{fantasysy@}sina.com`

Abstract. Abstract Text

Keywords: Knowledge Base, Semantic Web, Ontology, Cross-lingual

1 Introduction

With Linked Open Data(LOD) developing recent years, an increasing number of knowledge bases are generated for information sharing. Large-scale knowledge bases in LOD project, such as DBpedia[10], YAGO[7] and Freebase[3] are often used for information extraction[4], entity linking[14], recommendation[13, 5, 6]and many other applications. These knowledge bases are not only cross-domain but also multilingual, which is benefit for global knowledge sharing.

The nucleus of LOD DBpedia extracts structured information from Wikipedia and has already contained approximately 38.3 million things. Now DBpedia provides 124 versions of non-English language including Chinese. However, there are only XXX instances and 11 classes in Chinese. Compared with the 4.58 million things in English, the quality of Chinese information is obviously insufficient for further research and application.

A cross-lingual knowledge base can effectively promote global knowledge sharing, understanding and expanding. However, to build an available knowledge base, some problems must be addressed: (1) The imbalanced size of different language sources leads to less entities. Comparing the number of articles in English Wikipedia with those in Chinese Wikipedia, which are 5 million and 800 thousand separately, it's obvious that to structure a knowledge base in Chinese with Wikipedia is more challenging. (2) The number of existing cross-lingual links in Wikipedia accounts for a low proportion, which effects the quality of a bilingual knowledge base. Especially when there scarcely exists evident links in properties. Statistically, only XX thousand cross-lingual links can be found between English and Chinese articles. (3) The large but not rigorous category system in Wikipedia causes incorrect semantic relations in taxonomy. For example,example

Currently, there are several massive knowledge encyclopedias in Chinese, including Hudong Baike and Baidu Baike. To solve the imbalance problem, we utilize such separated sources to enrich our Chinese information. In this paper, we propose an approach to integrate four resources, which are English Wikipedia, Chinese Wikipedia, Hudong Baike and Baidu Baike, into one cross-lingual knowledge base, which contains XXX concepts, XXX instances and XXX properties. To get a more available result, we also make judgement on the relations among concepts and instances. At last, a SPARQL query interface is provided to access to the knowledge base. Specifically, our work makes the following contributions:

- We propose a method to build a Chinese-English cross-lingual knowledge base combining multi-encyclopedias. Among them, two Chinese encyclopedias are utilized to help balance and enrich information in two languages.
- We extend the cross-lingual link set by employing a cross-lingual knowledge linking discovery approach for concept and instance, and analyzing templates in Wikipedia for property.
- We prune the original taxonomy, which is extracted from encyclopedia category system, to retrieve more precise subClassOf and instanceOf relations in ontology.
- A website is developed based on our ontology and also a SPARQL interface is provided for public query operations in our knowledge base.

The rest of the paper is organized as follows. Section 2 introduces the four involved encyclopedias. Besides, definitions are proposed in this section. Section 4 presents the extraction approach in concept, instance and property level. Section 5 describes the procedure of building a knowledge base using extracting result. Section 6 shows the results of established knowledge base. Section 7 induces related work about this paper. Section 8 gives the conclusion.

2 Preliminary

In this section, we firstly introduce the four encyclopedias used to build and enrich our knowledge base. Then we give some related definitions to formalize our procedure.

2.1 Encyclopedias

Wikipedia Nowadays, Wikipedia is the largest data store of human knowledge. It was launched in 2001 and has hold over 35 million articles in 288 language by 2015. Out of these, English articles contribute most. There are 4.8 million articles in English while only over 800 thousand articles in Chinese. It is obvious that the quantity of English articles is far more than Chinese. Such imbalance makes those ontologies based on Wikipedia-only behave badly in cross-lingual aspect. For example, DBpedia has 683 concepts in English while only 11 Chinese concepts, that shows little help when trying to do cross-lingual research depending on DBpedia. To avoid this imbalance problem, two other Chinese encyclopedias are utilized to enrich Chinese source.

Articles in Wikipedia are manually edited by various authors. To avoid distinct formats and styles, Wikipedia provides templates to format their editing content, especially infoboxes. For example, The infobox in film (Interstellar) is edited according to the Template *Infobox film*, which maintains a property set of film articles.

Other Chinese Encyclopedias There are several large-scale monolingual Chinese Encyclopedias currently. Among those, Baidu Baike and Hudong Baike are the most content-rich. Hudong Baike was founded in 2005 and contains more than 12 million articles with about 9 million experts' contribution until 2015. At the mean time, Baidu Baike maintains more than 11 million articles. These two resources are similar in article structure, even in content in some cases.

Encyclopedia Page All these four encyclopedias provide two important elements which contain potential semantic information, category taxonomy and articles. A taxonomy presents the relations between categories. Usually a category has its sub-categories and super-categories. Fig.2.1 shows a screenshot of Hudong Taxonomy. An article describes an entity with rich information created and modified by several verified editors. Besides, an article may belong to one or more categories. The article content as well as the relation between an article and its belonged categories both help enormously when constructing a knowledge base. In general, there are five elements can be exploited in each article page:

Fig. 1. Taxonomy in Hudong

– Title: A Title is the label of an entity, which is unique to each article so that it can be used to distinguish entities.

– Abstract: An abstract is a brief summarize of the entity. It's always the first paragraph of an article. Usually it can be taken as an important feature due to the summary content.
– Infobox: Most of articles contain infobox. An infobox maintains structured data which are frequently subject-attribute-value triples formalized as a table. Information in this table includes important properties of an entity.
– Links: Links are entries to other articles within the encyclopedia. They lead readers to reference articles. Actually, they represent the relations between the current article and other articles.
– Category: The categories that an article belongs to are usually listed at the bottom of article page, shown as tags. An article attaches to one or more categories.

Fig. 2.1 shows a snap of an article in Chinese Wikipedia.

Fig. 2. A snap of Interstellar(Film) article in Chinese Wikipedia

2.2 Cross-lingual Links

In order to build a cross-lingual knowledge base, we need the mapping relation between Chinese sources and English sources, which can be used to fuze information in different language. In Wikipedia, some article pages have language links so that the reader can switch languages within the same article. Fig. 2.1shows language links of *Interstellar* on the right column of the Wikipedia page. Taking advantage of manually established language links, we can generate an initial cross-lingual ontology based on Chinese and English Wikipedia, similar to DBpedia. Afterwards, to combine Hudong Baike and Baidu Baike, which are lack of cross-lingual information but rich in Chinese data, we employ the approach proposed in [16] to discover cross-lingual links between English Wikipedia and Baidu Baike.

2.3 Definitions

Here we list definitions involved in the later sections:

Definition 1: A Encyclopedia is considered as a collection of articles, category system, which can be defined as: $W =< A, C >$, where A denotes articles, C denotes categories in W.

Definition 2: According to Section 2.1, there are several elements of an article. Therefore, each article a can be defined as follow:

$$a =< Ti(a), AB(a), Li(a), I(a), C(a) > \tag{1}$$

where $Ti(a), AB(a), Li(a), I(a), C(a)$ denotes title, abstract, links, infobox, category tags of article a.

Definition 3: As to an article a containing multi-language content in Wikipedia, L_e and L_z denotes its article tags, usually titles, in English and Chinese separately. Thus $cl(a) =< L_e(a), L_z(a) >$

Definition 4: An Infobox $I(a)$ in an article contains a set of attribute-value pairs p_1, p_2,.... In Wikipedia, an infobox is usually edited based on an appropriate infobox template recommended by Wikipedia, which here we denote as $T(a)$. Templates specify certain attributes, which are usually different from those displayed on the web page. Thus, we define an attribute-value pair as a triple $p =< tl, dl, v >$, where tl is attribute label in template, dl is displayed label in web page and v is the corresponding value.

3 Procedure Overview

Fig. 3 presents the whole procedure of our knowledge base building. We extract

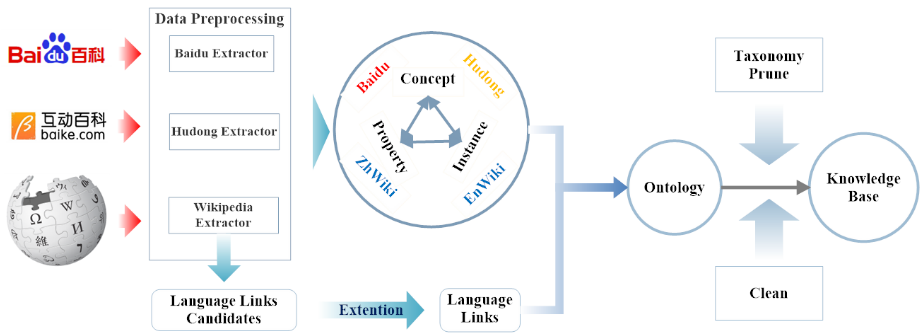

Fig. 3. Procedure of building our cross-lingual knowledge base

information from four sources, Baidu Baike, Hudong Baike, Chinese Wikipedia and English Wikipedia. Considering the different data format of each source, that is, html code page of Baidu and Hudong with different layout, and XML format of Wikipedia dump file, various extractors should be employed. After data parsing, we receive four datasets with concept lists, instance lists, property lists and their relations of each source. Meanwhile, an initial Chinese-English lan-link map is generated, which is then employed a link-discovery method to obtain a larger one. Using the extended language links, we combine the four datasets into one. At last, going through the procedure of taxonomy pruning and data-cleaning, we get a final knowledge base with higher accuracy.

4 Data Preprocessing

We prepare data for the knowledge base by a series of preprocessing operations. During this stage, our goal is to get a structured dataset including concepts, which are extracted from category taxonomy; instances, which are defined according to articles and properties based on both infobox and templates assistant. We will describe the extraction approach in detail below.

4.1 Concept Extraction

A concept is defined as a type of similar instances. For example, the concept of instance *Interstellar* is *Movie*. In general, a concept has super classes and sub classes, which means it has *subClassOf* relation with other classes. Concepts comprise a taxonomy which presents a backbone of an ontology.

In an encyclopedia, a category groups several articles and also has super-categories and sub-categories, just like concept does. Therefore we can extract concepts based on existing category system.

However, the whole taxonomy can not directly transform from category system because of the following problems:

- There are auxiliary categories in Wikipedia, which help arrange specific articles or category pages. For example, *Lists of artists* or *Food templates*.
- Some sub-category links in the category system maybe inconsistent. Some categories may contain itself as sub-category, or contain sub-category that also be the super-category of it. As Fig. **??** shows: In Hudong, the sub-category of (Head of State) contains itself as a child, which causes a circle in taxonomy tree. Meanwhile, in Wikipedia,
- Some categories relate to only one or two articles. According to the definition of concept, such categories are less representative to a group of instances, therefore it's unwise to retain it as concept.

To receive a cleaner and more precise concept taxonomy, we firstly do some refining works as follow:

- Delete list categories or template categories in Wikipedia.
- Delete inconsistent sub-categories and keep the super one.
- Delete categories that relate to less than two articles.

The cleaning work carried out in all encyclopedias are consistency when extracting. The remaining categories comprise an original concept taxonomy. A category and its sub-categories are correlated by the relation *SubClassOf*, and a category and its articles are correlated by the relation *InstanceOf*. However, there are still incorrect samples in the two relations. For example, *Tsinghua University* is not an entity of *Haidian District*, but relates to. Thus we will prune the taxonomy later.

4.2 Property Extraction

A property is defined as an attribute of an entity. It represents the relation between two instances or an instance and its value. We divided properties into two types: object property, whose value is an individual, such as *directed by*; datatype property, whose value is a literal text, such as *birth date*. Considering both content and infobox of an article, we extract two kinds of properties, general-properties and Infobox-properties.

General-properties Characteristics of an entity are seen as general-properties, including label, abstract, and url. Those properties describe specific information of an entity. The label property identifies a unique entity, whose value is the article title. The abstract property provides a brief description of an entity, whose value is the first paragraph of article. The url property saves the source of an entity, which is actually a url in each source. All of them are datatype properties.

Infobox-properties Attributes acquired from infobox are considered as Infobox-properties, such as (release date), (directed by) in a movie's infobox. Each attribute in the infobox is related to its own corresponding value. The value maybe a text or a reference to another entity. The type of a property, datatype or object, depends on the type of the value. Ordinarily, a plain text value marks the property as datatype while an entity reference determines the property as objecttype. For example, the attribute (release date) can be defined as a datatype property as its value is a datetime string. Meanwhile, (directed by) can be an object property because its value points to a person who directed the movie.

We are challenged when extracting properties from infoboxes:

– The attribute label displayed in the webpage infobox is inconsistent with it in the published dump file of Wikipedia. Fig.4.2 gives a mapping result of display label and dump label in *Interstellar*' infobox. The left is infobox while the other is a snap from dump file in Wikipedia. This movie's infobox follows template *Infobox film*, where dump labels come from. Obviously the two labels are similar but not the same. As a result we have to explore the display labels in Wikipedia rather than dump labels extracted from data file. As to Chinese sources, the property labels from Hudong Baike and Baidu Baike are equal literally to texts displayed on the webpage, so we don't need extra process as we have in Wikipedia.

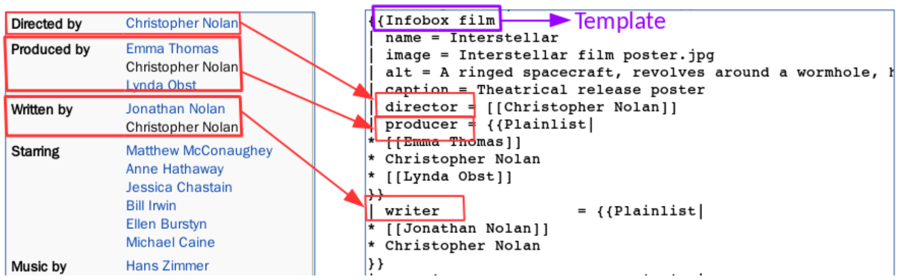

Fig. 4. Comparison of display label and dump label in *Interstellar* infobox

– There are special characters in labels. For example, Wikipedia usually use hyphen ”-” or dot ”” to mark sublabels. For example, population property have sub-properties ”-Density” and ”-Urban”. In addition, odd signs, such as colon or asterisk, may occur in Baidu or Hudong property labels by mistake.

To solve the problems above, we take advantage of template information. Specifically, Wikipedia institutes rules of rendering label in templates. Fig.**??** shows part of rules in *template:infobox film*. When a dump label in triple-bracket occurs in a dump page, we can replace it by its mapping display label. After collecting all the display labels, we make a filter to redress the label text.

4.3 Instance Extraction

In encyclopedia, an article describes an unique entity in the world. Therefore we can extract an article content as an instance. But we can't transform all the articles as instances because there are many illustrative or structure-related articles in Wikipedia, including category List pages and template documentations.

Each instance contains relations with concepts and properties. Take the movie (Interstellar) in Fig.2.1 as an example, concepts are assigned according to the category tags below the article page. (American science fiction films) is a concept of Interstellar. In the meantime, we obtain the label property from the article title, which is (Interstellar) and the abstract property from the first paragraph. The infobox-properties are also acquired via extracting from the infobox in the article. Besides, according to links in the content, we gain the reference between the current instance and others, such as (Warner Bros.).

After the preprocessing above, we harvest two types of information. One is the characteristics of instance, including instance-label, instance-abstract. The other includes relationships, containing concept-instance, instance-property-value and instance-instance.

5 Cross-lingual Knowledge Base Building

To construct a cross-lingual knowledge base with existing structured data, firstly we gather cross-lingual links which can help match the same entity in two languages and extend the link set. Secondly, we link this four encyclopedia, which is to say, respectively merging concepts, instances and properties from the four sources if they represent the same thing. Thirdly, we prune the taxonomy generated from concept relationships to retain a more accurate one. At last, we make instances and properties attached to the taxonomy to create a complete knowledge base.

5.1 Cross-lingual Links

Wikipedia has number cross-lingual links between English and Chinese. By extracting language links from Wikipedia, we can get an initial cross-lingual link set of concepts and instances. Moreover, we utilize the language-independent method in [16] to extend the language-link set. With the linkage factor graph model, we harvest a cross-lingual links extension as many as 20 thousands with an ideal precision 85.5% and a recall of 88.1%.

However, due to following templates, Infobox-properties have no obvious cross-lingual links. To acquire such links, we take the following steps: (1) Given a matched template, which has infobox rules in both English and Chinese, find the mapping pairs; (2) Given the English and Chinese infoboxes of a matched instance, compare their templates, find the matched label mapping to the same dump label. (3) Given the English and Chinese infoboxes of a matched instance, to datatype properties, compare the similarity of literal value; to object properties, check whether the value refer to the same entity. The whole process can be summarized as Algorithm **??**. Some of the symbols have already denoted in **??**

5.2 Cross-lingual Linking

In order to make all these encyclopedias linked to each other, we unify the same concept, instance and property from four sources, and give them unique identifiers. For example, we merge instances by the following method:

1) Given an instance extracted from Chinese Wikipedia, find the entity in Hudong and Baidu Baike with the same title.
2) Given an instance extracted from Hudong, find the entity which is not in Chinese Wikipedia but in Baidu Baike with the same title.
3) Other instances existing in only one encyclopedia are considered as independent instances. Up to now, we have combined all Chinese articles describing the same entity as one instance. Each instance may correspond to 1 to 3 sources.
4) To a $L_z(a)$, which is same as $Ti(a)$, find whether there is an English cross-lingual link in $CL < L_e, L_z >$. If exists, make the two as one instance and identify it using an ID.
5) If a L_z or a L_e has no matched cross-lingual link, number it with an new ID.

After the above steps, we acquire a list of instances and their unique IDs. Some of them contain cross-lingual information while some contain just monolingual information.

The process of unify concept and property is the same as instance. Meanwhile, all the relations in all sources are kept to prevent loss of information.

5.3 Taxonomy Prune

As a result of combining multi-source information without verifying, the taxonomy of concept system is mussy. For example, example. Therefore, we introduce the method from[18] to filter the correct subClassOf and instanceOf relations between two entities. Table. **??** shows some examples of correct relations. In particular, some language-dependent literal and language-independent structural features are defined to vectorize each concept or instance. The literal features include the head words singular/plural forms of labels of English entities, the prex/sufx relation of labels of Chinese entities. The structural features include article divergence for subClassOf, property divergence for subClassOf, category divergence for subClassOf, article divergence for instanceOf, property divergence for instanceOf, category divergence for instanceOf. By employing these features, a binary-classification model is trained based on Logistic Regression. The whole process is iterative by add assured result into training set and retraining the model to get higher precision. Confirming a right subClassOf or instanceOf not only depends on the prediction result of classification, but also cross-lingual knowledge validation, which ensures correctness if both the mapped English and Chinese relations are correct. Fig. **??** shows an example of the process.

The ideal result of pruning is a tree, whose edges denote relations, nodes denote concepts and leaves denote instances. However, since get rid of incorrect entity relations without consideration of integrity, getting a forest result is inevitable. We may omit some links but get more accurate relations instead.

5.4 Knowledge Base Building

To build a knowledge base, we first define specific URI to identify each element in the prepared dataset. URIs of concept, instance, and property are created by concatenating a namespace prefix to an ID. Table 1 lists the URI defined in the knowledge base.

Table 1. URIs for concept, instance, property in our knowledge base

Type	URI
Concept	http://xlore.org/concept/id
Instance	http://xlore.org/instance/id
Property	http://xlore.org/property/id

6 Result

6.1 DataSet

xmlhtmlarticleinstanceconceptproperty, wiki conceptproperty

6.2 Extracted Knowledge Base

We collect the resources from 4 wiki sites, Enwiki, Zhwiki, Hudong, and Baidu. Each of the wikis has three types of information, which can be utilized for constructing our ontology, namely, specific articles, classification system, and attribute of articles. We extract each raw data, then form the extracted information into well-structured data. Table2 the result we get after elementary extraction in 4 different wiki sources.

Table 2. Statistics of elementary extraction result

	Enwiki	Zhwiki	Hudong	Baidu
Instance	4304113	662650	5590751	5622404
Concept	982432	159705	31802	1300
Property	43976	18842	1187	139634

After fusing the heterogeneous resources, we harvest a cross-lingual graph with 1,093,855 classes, 200,223 properties, and 11,721,238 instances separately. With different methods of extraction and language linking discovery, these three kinds of entries show different results in languages. We give a breakdown of both Chinese knowledge and English knowledge in Table3.

Table 3. Statistics of our knowledge base

	Concepts	Instance	Property
English	982,982	4,311,719	54,802
English%	89.86%	37.79%	27.37%
Chinese	176,648	7,842,117	167,083
Chinese%	16.15%	66.9%	83.45%
Cross-lingual	65,775	432,598	21,662
Cross-lingual%	6.01%	3.7%	10.82%
Only English	917,207		33140
Only English%	83.9%		16.6%
Only Chinese	110,873		145,421
Only Chinese%	10.1%		72.6%
Sum	1,093,855	11,721,238	200,223

Our knowledge graph is organized in Openlink Virtuoso, which is a data management platform covering various server, including triple store.

6.3 Query Interface

We provide a website to demonstrate our knowledge in the forms of class, instance and property. The users of our system can choose language preference which is convenient for both English speaker and Chinese speaker. In the class webpage, we exhibit the label, super classes, subclasses, related topics, properties and instances of the specific class in bilingual way on condition that the corresponding English entries or Chinese entries exist. In the instance webpage we display bilingual label, super classes, related classes, abstract, property key value pairs, images and references. In the property webpage, we presents bilingual label, domains, ranges, and related instances of each property. Beside these user-friendly pages, we as well provide SPARQL interface for professional users to query our knowledge graph. Users can choose the language tags of their desired results by **"filter(langMatches(?label),"en"))"** or **"filter(langMatches (?label),"zh"))"**.

7 Related Work

In this section, we introduce some related knowledge bases and cross-lingual knowledge linking methods are referred to.

7.1 Chinese Knowledge Bases

Currently, several large-scale Chinese knowledge bases have been generated. Zhishi.me[11, 15] is the first Chinese large-scale Linking Open Data published. It acquires structural information from three original sources, Chinese Wikipedia, Baidu Baike and Hudong Baike and gains more than 5 million distinct entities. Zhishi.me helps generate knowledge base focused on relations in Junfeng Pans work[12]. Similar with Zhishi.me, CKB[17] is created from Hudong Baike. It first learns an ontology based on category system and properties, and then collects 19542 concepts, 2381 properties, 802593 instances. Besides using existing encyclopedias, CASIA-KB employs other types of sources(e.g. microblog posts, news pages, images) to enrich the structured knowledge.

7.2 Cross-lingual Knowledge Bases

DBpedia [2, 10] is one of the most used cross-lingual knowledge base in the world. It's extracts various kinds of structured information from Wikipedia and employ the multilingual characteristic of Wikipedia to generate 97 language versions of content. This knowledge base is widely applied in many domains, including media recommendation [13, 5, 6], entity linking[9] and information extraction [4]. Universal WordNet(UWN)[8] is a large multilingual lexical knowledge base which is build from WordNet and enriched its entities from Wikipedia. It is constructed using sophisticated knowledge extraction, link prediction, information integration, and taxonomy induction methods. The API is available to over 200 languages and more than 16 million words and names. UWN provides semantic relationship of list of word meanings for Aya's work on conceptual search [1]

8 Conclusion

This paper presents a procedure of building a Chinese-English cross-lingual knowledge base from four encyclopedia sources. At first, data preprocessing work is done to extract structured information and unify data format. Then a cross-lingual lan-link set is generated to help combine the bilingual sources. Besides, an extension method is applied to enrich the links. To refine our dataset, we also do pruning work on taxonomy. Finally, we acquire a knowledge base containing XXX concepts, XXX instances and XXX properties. Currently, a SPARQL query interface is provided to access the knowledge base.

References

1. Al-Zoghby, A.M., Shaalan, K.: Conceptual search for arabic web content. In: Computational Linguistics and Intelligent Text Processing, pp. 405–416. Springer (2015)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. Springer (2007)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM (2008)
4. Dutta, A., Meilicke, C., Niepert, M., Ponzetto, S.P.: Integrating open and closed information extraction: Challenges and first steps. In: NLP-DBPEDIA@ ISWC (2013)
5. Fernández-Tobías, I., Cantador, I., Kaminskas, M., Ricci, F.: A generic semantic-based framework for cross-domain recommendation. In: Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems. pp. 25–32. ACM (2011)
6. Kaminskas, M., Fernández-Tobías, I., Ricci, F., Cantador, I.: Knowledge-based music retrieval for places of interest. In: Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. pp. 19–24. ACM (2012)
7. Mahdisoltani, F., Biega, J., Suchanek, F.: Yago3: A knowledge base from multilingual wikipedias. In: 7th Biennial Conference on Innovative Data Systems Research. CIDR 2015 (2014)
8. de Melo, G., Weikum, G.: Uwn: A large multilingual lexical knowledge base. In: Proceedings of the ACL 2012 System Demonstrations. pp. 151–156. Association for Computational Linguistics (2012)
9. Mendes, P.N., Daiber, J., Jakob, M., Bizer, C.: Evaluating dbpedia spotlight for the tac-kbp entity linking task. In: Proceedings of the TACKBP 2011 Workshop. vol. 116, pp. 118–120 (2011)
10. Mendes, P.N., Jakob, M., Bizer, C.: Dbpedia: A multilingual cross-domain knowledge base. In: LREC. pp. 1813–1817 (2012)
11. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi. me-weaving chinese linking open data. In: The Semantic Web–ISWC 2011, pp. 205–220. Springer (2011)
12. Pan, J., Wang, H., Yu, Y.: Building large scale relation kb from text. In: 11th International Semantic Web Conference ISWC 2012. p. 93. Citeseer (2012)

13. Passant, A.: dbrecmusic recommendations using dbpedia. In: The Semantic Web–ISWC 2010, pp. 209–224. Springer (2010)
14. Shen, W., Wang, J., Luo, P., Wang, M.: Linden: linking named entities with knowledge base via semantic knowledge. In: Proceedings of the 21st international conference on World Wide Web. pp. 449–458. ACM (2012)
15. Wang, H., Wu, T., Qi, G., Ruan, T.: On publishing chinese linked open schema. In: The Semantic Web–ISWC 2014, pp. 293–308. Springer (2014)
16. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. In: Proceedings of the 21st international conference on World Wide Web. pp. 459–468. ACM (2012)
17. Wang, Z., Wang, Z., Li, J., Pan, J.Z.: Building a large scale knowledge base from chinese wiki encyclopedia. In: The Semantic Web, pp. 80–95. Springer (2012)
18. Wang, Z., Li, J., Li, S., Li, M., Tang, J., Zhang, K., Zhang, K.: Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)