Building a Large-Scale Cross-Lingual Knowledge Base from Heterogeneous Online Wikis

No Author Given

No Institute Given

Abstract. Cross-lingual Knowledge Bases are important for global knowledge sharing. However, there are few Chinese-English knowledge bases due to the following reasons: 1) the scarcity of Chinese knowledge; 2) the limited number of current cross-lingual language links; 3) the incorrect relation in semantic taxonomy. In this paper, a large-scale cross-lingual knowledge base(CLKB) is built to solve the above problems. Particularly, the CLKB integrates four online wikis including English Wikipedia, Chinese Wikipedia, Baidu Baike and Hudong Baike to balance the knowledge volume in different languages, employs a link-discovery method to augment the language links, and introduce a pruning approach to refine taxonomy. Totally, CLKB harvests 663,740 classes, 56,449 properties, and 10,856,042 instances, among of which, 507,042 entities are cross-lingually linked. To the best of our knowledge, our CLKB is the first large-scale Chinese-English knowledge base with balanced knowledge quantity. At last, we provide a knowledge-display system supporting two ways to access the established CLKB, namely a keyword search entrance and a SPARQL endpoint.

Keywords: Knowledge Base, Cross-lingual Linking, Taxonomy Pruning

1 Introduction

As the Web is evolving to a highly globalized information space, sharing knowledge across different languages is attracting increasing attentions. Multi-lingual knowledge bases have significant applications such as information retrieval, machine translation and deep question answering. DBpedia, by extracting structured information from Wikipedia[1], is a multi-domain knowledge base covering many domains and becomes the nucleus of Linked Open Data[2]. Obtained from the automatic integration of WordNet[3] and Wikipedia, YAGO, MENTA and BabelNet are other famous large multi-lingual knowledge bases.

However, most non-English knowledge in those knowledge bases is pretty scarce. Fig. 1 shows a simplified long tail distribution of the number of articles on six major Wikipedia language versions. Due to the imbalance of different Wikipedia language versions, the knowledge distribution across different

[1] `http://www.wikipedia.org/`
[2] `http://linkeddata.org/`
[3] `http://wordnet.princeton.edu/`

languages is highly unbalanced in those knowledge bases genuagenerated from Wikipedia. For instance, DBpedia contains 4.58 million English instances but even no Chinese dataset published. On the other hand, the Chinese Hudong Baike[4] and Baidu Baike[5], both containing more than 11 million articles, are even larger than the English Wikipedia. If a knowledge base could be established between English Wikipedia and Chinese Hudong Baike, an Chinese-English knowledge base with much higher coverage in Chinese can be constructed.

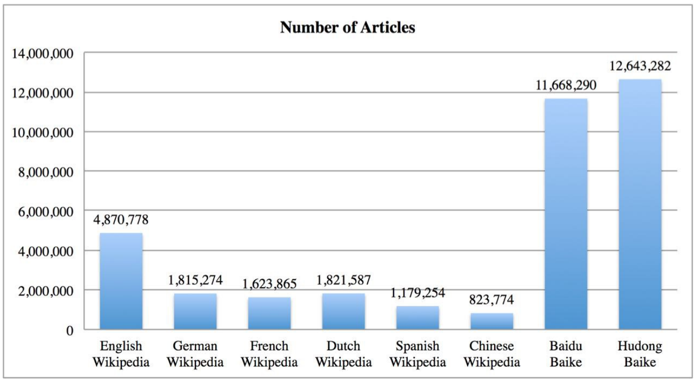


Fig. 1. Number of Articles on Major Wikipedias, Baidu Baike and Hudong Baike

To enrich the Chinese knowledge, we try to build a large-scale Chinese-English knowledge base by semantifying four heterogeneous online wikis, which are English Wikipedia, Chinese Wikipedia, Hudong Baike and Baidu Baike. This non-trivial task poses the challenges as follows:

1. The cross-lingual links are highly limited inside Wikipedia. For instance, there are only 9% Chinese-English matched articles in Wikipedia in all articles. How could we find enough Chinese-English `owl:sameAs` relations?
2. The subsumption relations of the online wikis' category systems contain lots of noise. For example, in English Wikipedia "Wikipedia-books-on-people", which is actually `subClassOf` "Books", is taken as the sub-category of "People" mistakenly. How could we detect those incorrect semantic relations?

Driven by these challenges, we propose a unified framework to build a Chinese-English knowledge base from four heterogeneous online wikis. The framework contains three steps: extracting dataset from online wikis, extending an initial cross-lingual link set, and pruning taxonomy for more precise semantic relations. The generated knowledge base contains 663,740 classes, 56,449 properties and 10,856,042 instances. Specifically, we make the following contributions:

[4] `http://www.baike.com/`
[5] `http://baike.baidu.com/`

1. We extend cross-lingual link set by employing a cross-lingual knowledge linking discovery approach for class and instance, and by analyzing templates in Wikipedia for property.
2. We prune the original taxonomy, which is extracted from wiki category system, to retrieve more precise *subClassOf* and *instanceOf* relations.
3. An online-system supporting keyword search and SPARQL endpoint is provided for public access to our knowledge base.

The rest of the paper is organized as follows. Section 2 presents some basic concepts and the problem formulation. In Section 3 we introduce detailed extraction approaches. Section 4 describes cross-lingual integration. The experimental results are reported in Section 5. Related works are given in Section 6. Finally we conclude our work in Section 7.

2 Preliminaries

In this section, we give definitions about our knowledge base and describe our task.

Online Wikis. Nowadays, Wikipedia is the largest data store of human knowledge. It was launched in 2001 and has hold over 35 million articles in 288 languages by 2015. Out of these, English articles contribute 13% while Chinese have 2%.

Baidu Baike and Hudong Baike are the most content-rich among the large-scale monolingual Chinese wikis currently. Hudong Baike was founded in 2005 and contains more than 12 million articles with about 9 million experts' contribution until 2015. Meanwhile, Baidu Baike maintains over 11 million articles.

Wiki Pages. Articles from the wiki sources are similar in structure. Usually they provide two important elements with potential semantic information, category system and articles. Here, we define an encyclopedia wiki as: $W =< C, A >$, where C denotes categories, A denotes articles in W. A category system represents the relations between categories as a tree by the relation *subCategoryOf*. An article describes an entity with rich information. In general, six elements can be exploited in each article page:

- Title: A Title is the label of an entity, whose uniqueness can be used to identify entities.
- Abstract: An abstract which is always the first paragraph in content, is a brief summary of the entity.
- Infobox: An infobox maintains structured data describing representative features of entity in subject-attribute-value triples.
- Link: Links are entries to other articles within the wiki. They represent the relations between the current article and others.
- Category: Categories that an article belongs to are usually listed at the bottom of article page. An article has *articleOf* relation with its categories.
- URL: Each article has an HTTP url to locate itself on web.

An article a can be defined as follow:

$$a =< Ti(a), Ab(a), Li(a), In(a), C(a), U(a) >$$

where $Ti(a), Ab(a), Li(a), In(a), C(a), U(a)$ denotes title, abstract, links, infobox, category tags, url of article a.

Notably, infoboxes in articles are generated based on certain templates recommended by Wikipedia. An infobox template collects attributes describing similar entities. For example, content in infobox of film 冰雪奇缘 (Frozen) is normalized by *Template:Infobox film*, which maintains a property set of films that should be given assignment. We denote the infobox template used in article a as $T(a)$. However, attribute labels in templates are usually different from those displayed on the webpage. Thus, we define an attribute as a triple $p =< tl, dl, v >$, where tl is attribute label in template, dl is display label in web page and v is the corresponding value. The value maybe a text or a reference to another entity.

Cross-lingual Links. In Wikipedia, some article pages have cross-lingual links which help readers switch to corresponding articles in other languages. As for an entity containing both Chinese article a_z and English article a_e in Wikipedia, we say a_z and a_e are *cross-linked*. Because templates are also articles in Wikipedia actually, they maybe cross-linked. For a pair $cl \in CL$, $cl =< L_z, L_e >$, where L_z and L_e denote the entity's cross-lingual links in Chinese and English. .

Knowledge Base. A knowledge base is a formal specification of a group of entities. Our knowledge base is described as a 4-tuple:

$$KB =< C, P, I, H^C >$$

where C, P, I are the sets of classes, properties, and instances respectively. H^C represents the hierarchical relationships of classes.

Our knowledge base includes four kinds of semantic relation, which are *subClassOf* of class-class, *instanceOf* of class-instance, *relatedClassOf* of instance and related class, *relatedTopicOf* of class and related topic.

A Cross-Lingual Knowledge Base(CLKB) is a database conforming to a cross-lingual ontology. Taking advantage of cross-lingual links, several monolingual knowledge bases generated from various sources can be merge into one. Thus our Chinese-English knowledge base is defined as:

$$CLKB =< KB_z, KB_e >$$

where KB_z and KB_e denote the monolingual knowledge base in Chinese and English.

Cross-Lingual Knowledge Base Building. In this paper, our task is to build a CLKB assembling knowledge from several English or Chinese wiki sources. Given an online-wiki W_i, we get dataset including class list C_i, instance list I_i, property list P_i, class hierarchical relationships H_i^C of W_i by extraction. We then enrich the existing cross-lingual link set CL using a link-discovery method. We also refine taxonomy by checking if an *articleOf* or *subCategoryOf* is really an *instanceOf* or *subClassOf* relationship. Our final output is a Chinese-English CLKB combining KB_z and KB_e by utilizing cross-lingual links in CL. The whole procedure is shown in Fig. 2.

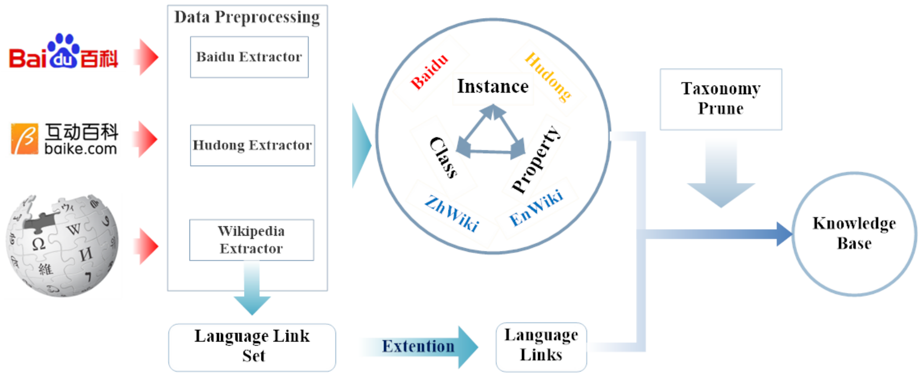


Fig. 2. Procedure of Building Our Cross-Lingual Knowledge Base

3 Semantic Data Extraction

Semantic data extraction aims to achieve a structured dataset from the input wikis. Specifically, we extract classes from category system, instances according to articles, and properties based on infoboxes.

3.1 Class Extraction

A class is defined as a type of similar instances. For example, the class of instance 冰雪奇缘 (Frozen) is 电影 (Film). In general, a class is semantic related with other classes by $subClassOf$ relation. Such relations comprise a taxonomy which presents the backbone of an ontology. In a wiki, a category groups several articles and also has $subCategoryOf$ relation with others. Therefore we can extract classes based on existing category system.

However, the whole class hierarchy can not directly transform from category system due to 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 categories relate to only one article. According to the definition of class, such categories are less representative types, therefore it's unwise to retain them as classes.

To gain a more precise H^C in a wiki, we discard categories matching the above conditions, then build the original class hierarchy H^C using the remaining categories.

3.2 Property Extraction

A property is defined as an attribute of an entity. It represents the relation between an entity and its value. We divide properties into two types: object property, whose value is an individual, such as 导演 (directed by); datatype property, whose value is a literal text, such as 出生日期 (birth date). Considering both content and infobox of an article, we extract two kinds of properties, General-properties and Infobox-properties.

General-properties. General-properties describe general information of an entity. We define three datatype properties as general-properties for a given article a: (1) label ~~property~~; (2) abstract ~~property;~~ (3) URL ~~property~~.

Infobox-properties. Attributes acquired from infobox are considered as infobox properties, such as 上映时间 (release date), 导演 (directed by) in a movie's infobox. 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 object. For example, the attribute 上映时间 (release date) can be defined as a datatype property as its value is a datetime string. On the other hand, 导演 (directed by) is an object property because its value points to a person who directed the movie.

We are challenged when extracting properties from infoboxes:

– In Wikipedia, the attribute label displayed in the webpage infobox is inconsistent with that in the published dump file. Fig. 3 gives a mapping result of display labels and dump labels in 冰雪奇缘 (Frozen)'s infobox. Webpage infobox is on the left. The middle is a glance from dump file in Wikipedia. The attribute label 主演 displayed on webpage is different from label *starring* extracted from dump data. However, readers generally consider the display label rather than extracted label as an attribute label.

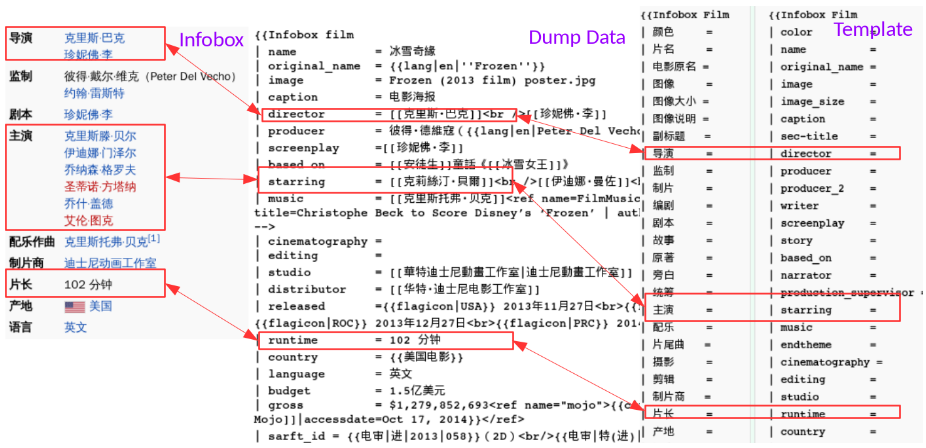


Fig. 3. Comparison of display label and dump label in *Frozen* infobox

– There are special characters in labels. Wikipedia usually uses hyphen "-" or dot "•" to mark sublabels. For example, attribute 人口 (population) has sub-attributes "-Density" and "-Urban". In addition, ":" or "*", may occur in Baidu or Hudong attribute labels by mistake.

To solve the problems above, we take advantage of template information. Specifically, Wikipedia institutes rules of rendering label in templates. Right of Fig. 3 shows an example of *Template:Infobox film*, where relations of display labels and dump labels come from. When a dump label occurs in dump file, we replace it by its mapping display label. Then we filter attribute labels to get correct ones.

3.3 Instance Extraction

An article describes a unique entity in the world. Therefore we can extract an article as an instance. During the extraction, illustrative or structure-related articles in Wikipedia are deleted, including category list pages and template documentations.

We harvest four types of information during this stage. (1) General-properties of instance, including title as label property value, first paragraph as abstract property value and HTTP URL as URL property value; (2) Infobox-properties which are acquired via extracting from the infobox in the article; (3) $articleOf$ relation with categories listed at the bottom of article page. For example, 冰雪奇缘 (Frozen) is an article of category 美国电影作品 (American films); (4) Reference relation with other instances according to links in the content, such as 冰雪女王 (The Snow Queen).

4 Cross-lingual Integration

To construct a CLKB with obtained structured data, firstly we argument cross-lingual links, which can help match the same entity in two languages. Secondly, we integrate this four wikis, that is, respectively merging classes, instances and properties from the four sources~~ if they represent the same thing~~. Thirdly, we prune the taxonomy generated from class relationships to make it more accurate.

4.1 Cross-lingual Linking

We follow two ways to get cross-lingual links for class, instance and property. One is extracting links according to Wikipedia language links for class and instance, the other is achieving links using infobox templates for property.

There are 227 thousand cross-lingual links between English and Chinese, which constitute the initial cross-lingual link set of classes 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 215 thousand with an ideal precision 85.5% and a recall of 88.1% between English Wikipedia and Baidu Baike.

However, due to using templates, Infobox-properties have no obvious cross-lingual links. Thus, we take the following steps to acquire property links:

1. Given two matched templates, T_e and T_z, find the display labels mapping the same template label. That is, for p_e in T_e and p_z in T_z, if tl_e is equal to tl_z, $< dl_e, dl_z >$ are cross-lingual properties;
2. Given the English and Chinese infoboxes of two matched articles a_e and a_z, compare their templates, English template $T_e(a_e)$ and Chinese template $T_z(a_z)$, find the matched display labels mapping to the same dump label;
3. Given the English and Chinese infoboxes of a matched instance, for datatype properties, compare the similarity of literal value; for object properties, check whether the value refer to the same entity.

In order to make all wikis link to each other, we unify the same class, instance and property from four sources, and give them unique identifiers. For instance, we merge instances by the following steps: (1) Merge all instances extracted from Chinese wikis by instance title. (2) To an L_z, find whether there is an English cross-lingual link L_e in $CL < L_e, L_z >$. If exists, make the two as one instance and identify it using an ID, otherwise number it with a new ID.

The process of unifying class and property is the same as instance. Meanwhile, all the relations are kept to prevent loss of information.

4.2 Taxonomy Prune

There is inevitably noise in the taxonomy since we combine multi-source information without verification. Therefore, we introduce the method from [18] to detect the correct *subClassOf* and *instanceOf* relations from *subCategoryOf* and *articleOf*. In particular, some language-dependent literal and language-independent structural features are defined to vectorize each class or instance. Employing these features, a binary-classification model is trained based on Logistic Regression. The whole process is iterative by retraining model with assured result to get higher precision. Moreover, the prediction results are validated by cross-lingual information.

The ideal result after pruning is a tree, whose edges, nodes, and leaves respectively denote relations, classes and instances. However, since getting rid of incorrect entity relations without consideration of integrity, a forest result is inevitable.

To retain integrity of semantic relation, we define two types of new relations: *relatedClassOf* for cut class-instance relations and *relatedTopicOf* for pruned class-class relations.

5 Result

Here we show statistic results of our CLKB and introduce the system based on the knowledge base.

5.1 Extracted Knowledge Base

We collect the resources from four online wikis, English and Chinese Wikipedia dump files in May, 2014, Hudong html pages until May, 2014, and Baidu html pages until September, 2014. Each of the wikis has three types of information, which can be utilized for constructing our knowledge base, namely, specific articles, category system, and attributes of articles. We extract each raw data, then form the extracted information into well-structured data. Table 1 shows the result we get after elementary extraction on 4 different wiki sources. Besides, we obtain 227 thousand class or instance links between English and Chinese in Wikipeidia, and increase the number by 215 thousand Enwiki-Baidu language links and 10 thousand property links.

Table 1. Statistics of Elementary Extraction Result

	Enwiki	Zhwiki	Hudong	Baidu
#Instance	4,304,113	662,650	5,590,751	5,622,404
#Class	982,432	159,705	31,802	1300
#Property	43,976	18,842	1187	139,634

We create URIs `http://clkb/type/id` (type could be *class*, *instance*, *property*) to identify each entry and provide corresponding information if users look up knowledge over HTTP protocol to achieve the knowledge base.

After fusing the heterogeneous sources, we harvest a cross-lingual knowledge base with 663,740 classes, 56,449 properties, and 10,856,042 instances respectively. With different methods of extraction and language link discovery, these three kinds of entries show different results in languages. We give a breakdown of both Chinese knowledge and English knowledge in Table 2.

Table 2. Statistics of Our Knowledge Base

	Classes		Instance		Property	
English	639,020	96.26%	3,879,121	38.79%	15,380	27.24%
Chinese	88,615	13.35%	7,409,519	68.25%	51,618	91.44%
Cross-lingual	63,895	9.63%	432,598	3.98%	10,549	18.69%
Only English	575,125	86.65%	3,446,523	31.75%	4,831	8.56%
Only Chinese	24,720	3.72%	6,976,921	64,27%	41,069	72.75%
Total	663,740	-	10,856,042	-	56,449	-

Our knowledge base is organized in Openlink Virtuoso, which is a data management platform rendering various services, including triple store.

5.2 Web Access to Knowledge Base

We provide a platform to present an intuitive visualization of our knowledge in the forms of class, instance and property. Fig. 4 shows sample pages of the integrated data. Language could be switched, which is convenient for both English speaking and Chinese speaking user. In the class webpage, we exhibit the label, super classes, subclasses, related topics, properties and instances of the

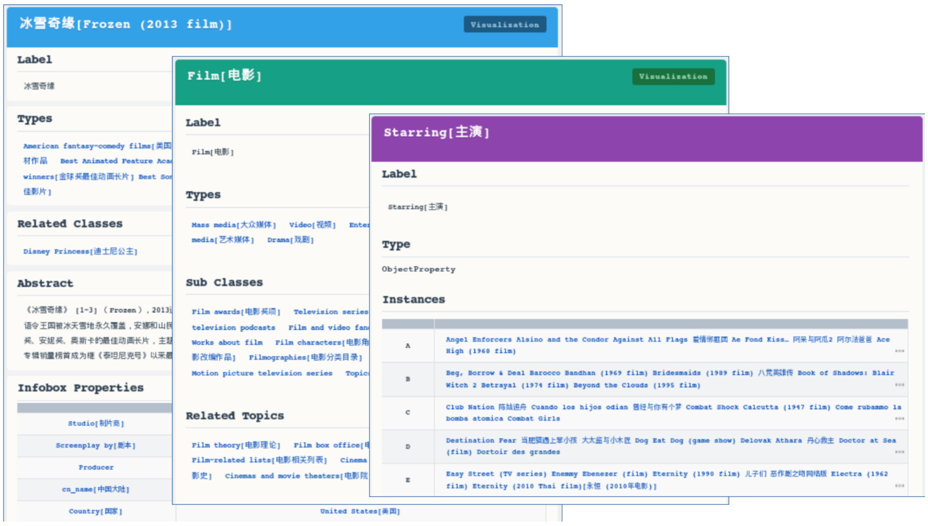
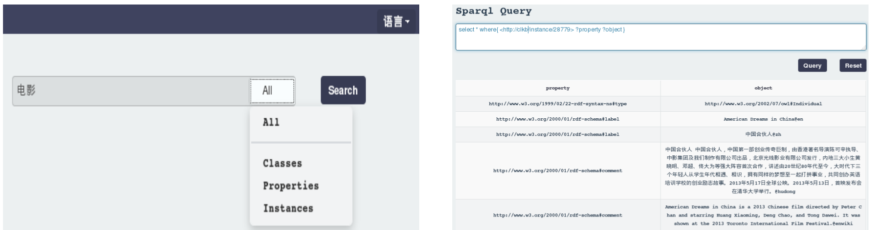


Fig. 4. Sample Pages of Instance, Class and Property

specific class in bilingual way. In the instance webpage we display bilingual label, super classes, related classes, abstracts, infobox-properties, images and URL references. In the property webpage, we present bilingual label, type, and related instances of each property.



(a) A Sample of Search Box (b) A Sample of SPARQL Query

Fig. 5. Two Ways Accessing to Our Knowledge Base

Beside these user-friendly pages, we provide two ways to access our knowledge base shown in Fig. 5: via the search engine or via SPARQL endpoint. For general users, they can make a query by inputing related text into searchbox to get probable related entities. To present practicable result, An index is generated over all entities. 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"))"**.

6 Related Work

In this section, we introduce some related knowledge bases.

Chinese Knowledge Bases. Currently, several large-scale Chinese knowledge bases have been generated. Zhishi.me[11, 15] is the first published Chinese large-scale Linking Open Data. 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 Pan's 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 19,542 classes, 2,381 properties, 802,593 instances. Besides using existing online-wikis, CASIA-KB employs other types of sources(e.g. microblog posts, news pages, images) to enrich the structured knowledge.

Cross-lingual Knowledge Bases. DBpedia [2, 10] is one of the most used cross-lingual knowledge base in the world. It extracts various kinds of structured information from Wikipedia and employ the multi-lingual characteristic of Wikipedia to generate 97 language versions of content. This knowledge base is widely applied in many domains, including media recommendation [5, 6], entity linking[9] and information extraction [4]. Universal WordNet(UWN)[8] is a large multi-lingual 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 classual search [1]

7 Conclusion

This paper presents a procedure of building a Chinese-English CLKB from four wiki sources. At first, we extract structured information and unify data format. Then a cross-lingual language link set is generated and expanded to help combine the bilingual sources. To refine our dataset, we also conduct pruning work on taxonomy. Finally, we acquire a CLKB containing 663,740 classes, 56,449 properties, and 10,856,042 instances. Currently, an online-system and 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.: dbrec—music 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)