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

Mingyang Li (✉), Yao Shi, Zhigang Wang, and Yongbin Liu

Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
{meeya.yx,syontheway,wangzigo,yongbinliu03}@gmail.com

Abstract. Cross-Lingual Knowledge Bases are very important for global knowledge sharing. However, there are few Chinese-English knowledge bases due to the following reasons: 1) the scarcity of Chinese knowledge in existing cross-lingual knowledge bases; 2) the limited number of cross-lingual links; 3) the incorrect relationships in semantic taxonomy. In this paper, a large-scale Cross-Lingual Knowledge Base(named XLORE) is built to address the above problems. Particularly, XLORE 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 cross-lingual links, and introduces a pruning approach to refine taxonomy. Totally, XLORE harvests 663,740 classes, 56,449 properties, and 10,856,042 instances, among of which, 507,042 entities are cross-lingually linked. At last, we provide an online cross-lingual knowledge base system supporting two ways to access established XLORE, namely a search engine 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, knowledge sharing across different languages attracts 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-lingual knowledge base covering many domains and becomes the nucleus of Linked Open Data[2]. YAGO, MENTA and BabelNet are other famous large multi-lingual knowledge bases.

However, most non-English knowledge is pretty scarce. The knowledge distribution across different languages is highly unbalanced in Wikipedia-based knowledge bases. For instance, DBpedia contains 4.58 million English instances but no Simplified Chinese dataset published. On the other hand, the Chinese

[1] http://www.wikipedia.org
[2] http://linkeddata.org

© Springer International Publishing Switzerland 2015
J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 413–420, 2015.
DOI: 10.1007/978-3-319-25207-0_37

Hudong Baike[3] and Baidu Baike[4], both containing more than 11 million articles, are even larger than the English Wikipedia. If a knowledge base could be established based on both English Wikipedia and Chinese Hudong Baike, more Chinese-English knowledge can be generated.

We try to build a large-scale cross-lingual knowledge base generated from four heterogeneous online wikis, i.e. English Wikipedia, Chinese Wikipedia, Hudong Baike and Baidu Baike. This non-trivial task poses the challenges as follows: 1) Limited English-Chinese cross-lingual links within Wikipedia (i.e. 4.6%). How could we find enough Chinese-English `owl:sameAs` relations? 2) Noisy subsumption relations in the category systems, e.g. "Wikipedia-books-on-people", which is actually *subClassOf* "Books", is mistaken as sub-category of "People". How could we detect those incorrect semantic relations?

To tackle these issues, we propose a unified framework to build a Chinese-English knowledge base from four heterogeneous online wikis in three steps: 1) extract wiki dataset 2) extend cross-lingual link set 3) prune taxonomy. The generated KB, named **XLORE**[5], contains 663,740 classes, 56,449 properties and 10,856,042 instances. Specifically, we make the following contributions: (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.

2 Preliminaries

Online Wikis. Nowadays, Wikipedia is the largest data store of human knowledge. It has hold over 35 million articles in 288 languages by 2015. Baidu Baike and Hudong Baike are the most content-rich among the large-scale monolingual Chinese wikis currently. Hudong Baike contains more than 12 million articles until 2015 while Baidu Baike maintains over 11 million articles.

Wikis usually 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. A category system represents the relations between categories as a tree by *subCategoryOf*.

Wiki Pages. Articles from the wiki sources are similar in structure. An article a can be defined as follow: $a =< Ti(a), Ab(a), Li(a), Ib(a), C(a), U(a) >$ where $Ti(a), Ab(a), Li(a), Ib(a), C(a), U(a)$ denote 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

[3] http://www.baike.com
[4] http://baike.baidu.com
[5] http://xlore.org

similar entities, e.g. infobox of film 冰雪奇缘 (Frozen) is normalized by *Template:Infobox film*, and 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, an infobox-attribute is defined as $p =< tl, dl, v >$, where tl and dl denote the label in template and web page, and v is the attribute value.

Cross-lingual Links. In Wikipedia, cross-lingual links help readers switch to preferred languages. If an entity containing both Chinese article a_z and English article a_e, then a_z and a_e are *cross-linked*. Infobox templates may also be 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, and H^C represents the class hierarchy. Semantic relations include *subClassOf*, *instanceOf*, *relatedClassOf*, *relatedTopicOf*.

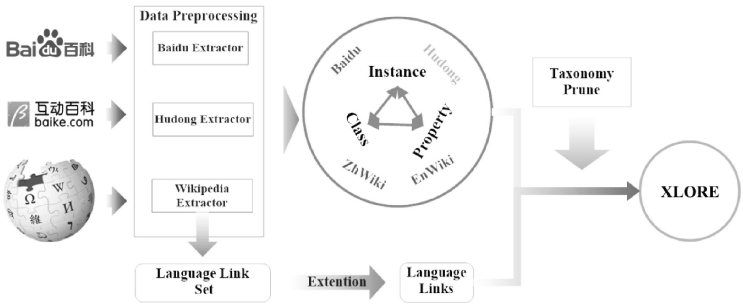

Fig. 1. Procedure of Building Our Cross-Lingual Knowledge Base XLORE

A Cross-Lingual Knowledge Base(CLKB) is a database conforming to a cross-lingual ontology, and is often integrated from various sources based on cross-lingual links. Thus CLKB is defined as: $CLKB =< KB_z, KB_e >$, where KB_z and KB_e denote the knowledge bases in Chinese and English.

Cross-Lingual Knowledge Base Building is to build a CLKB assembling knowledge from several English and Chinese wiki sources. Specifically, first build monolingual knowledge base from each W_i, then enrich the existing cross-lingual links, further refine the taxonomy, and finally integrate datasets in different languages based on the cross-lingual links as shown in Fig. 1.

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.

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 class hierarchy H^C 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.

To gain a more precise H^C in a wiki, we discard categories matching the following conditions: 1) There are auxiliary categories in Wikipedia, which help arrange specific articles or category pages. For example, *Lists of artists* or *Food templates*. 2) Some categories are related 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.

Property Extraction. A property is defined as an attribute of an entity. 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 describe general information of an entity. We define three Datatype properties as general-properties for a given article a: (1) label; (2) abstract; (3) URL.

Infobox-properties includes attributes acquired from infobox, such as 上映时间 (release date), 导演 (directed by) in a movie's infobox. As for the type of a property (i.e. Datatye or Object), a plain text value marks the property as Datatype while an entity reference determines the property as Object.

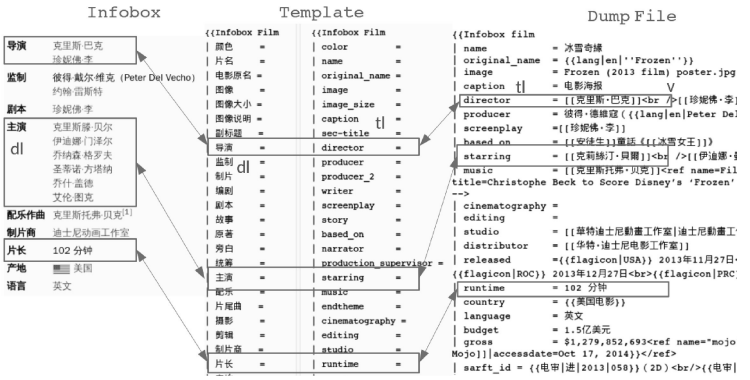

Fig. 2. Comparison of display label and template label in *Frozen* infobox

In Wikipedia, attribute labels displayed in webpage infoboxes are inconsistent with those in the dump file. Fig. 2 gives a case-control of display labels and template labels in 冰雪奇缘 (Frozen)'s infobox. The attribute label 主演 displayed on webpage is different from the label *starring* extracted from dump

file. However, readers generally consider the display label as an attribute label. Therefore, we employ the infobox template to bridge this gap: replace the template label in dump file by its matched display label. Then we sweep special characters in labels such as hyphen "-" or dot "•" in Wikipedia and ":" or "*" in Baidu. Furthermore, properties correspond to only one instance are discarded.

Instance Extraction. An article describes a unique entity in the world. Therefore we can extract an article as an instance. 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

We construct a CLKB with obtained structured data in the following three steps:

Cross-lingual Linking is to match the same entity (i.e., class, property and instance) in two languages. For class and instance, we utilize the linkage factor graph model in [6] to extend the original 227 thousand cross-lingual Chinese-English links in Wikipedia, and achieve 215 thousand links between English Wikipedia and Baidu Baike. For property, because Infobox-properties have no obvious cross-lingual links, we achieve the links using infobox templates:

1. Given two cross-linked templates, T_e and T_z, find the display labels mapping to the same template label. That is, if a template label tl_e in T_e is equal to a template label tl_z in T_z, $< dl_e, dl_z >$ are cross-lingual property labels;
2. Given two cross-linked articles, a_e and a_z, and their infobox templates, $T_e(a_e)$ and $T_z(a_z)$, for p_e in $T_e(a_e)$ and p_z in $T_z(a_z)$, if $p_e.tl$ is equal to $p_z.tl$, $< p_e.dl, p_z.dl >$ are cross-lingual property labels;
3. Given two cross-linked articles and their Infobox-properties P_e and P_z, for $p_e \in P_e$ and $p_z \in P_z$, $p_e.dl$ and $p_z.dl$ are cross-lingual when: (1)for datatype properties, $sim(p_e.v, p_z.v) > threshold$, where $sim(a, b)$ is a similarity function. (2)for object properties, $p_e.v$ and $p_z.v$ refer to the same entity.

Wikidata Integration. In order to integrate all wikis, we unify classes, instances or properties describing the same thing from four sources, and distribute a unique identifiers. For instance, we merge instances by the following steps: (1) Merge all instances extracted from wikis by title. (2) If a Chinese instance has a cross-linked English instance, that is, to an L_z, if there is $< L_z, L_e >$ in CL, make them as one instance. (3) Identify all instances, including both monolingual and cross-lingual, by IDs. The processes of unifying class and property are the same as instance.

Taxonomy Prune. There is inevitably noise in the taxonomy since we combine multi-source information without verification. Therefore, we introduce the

method from [8] to detect the correct *subClassOf* and *instanceOf* relations from *subCategoryOf* and *articleOf*. The ideal result after pruning is a tree, whose edges, nodes, and leaves respectively denote semantic 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 instance-class relations and *relatedTopicOf* for pruned class-class relations.

5 Result

Here we show statistic results of our CLKB, XLORE, and introduce the developed system based on XLORE dataset.

Table 1. Statistics of Elementary Extraction Result

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

Table 2. Statistics of XLORE

	Classes		Instances		Properties	
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%
Total	663,740	-	10,856,042	-	56,449	-

Knowledge Base Overview. 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, category system, specific articles, and attributes of articles. Table 1 shows the results we get after elementary extraction on 4 different wiki sources.

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.

Web Access to XLORE. We organize XLORE in Openlink Virtuoso, and provide a platform, which locates on http://xlore.org, to present an intuitive visualization in the forms of instance, class and property. URIs http://xlore.org/type/id (type could be *class*, *instance*, *property*) are created to identify each

Fig. 3. Sample Pages of Instance, Class and Property

entry . Fig. 3 shows sample pages of the integrated data. Language could be switched, which is convenient for both English speaking and Chinese speaking users. Besides these user-friendly pages, we provide two ways to access our knowledge base. For general users, they can send a query by inputing related text into searchbox to get probable related entries. To present practicable result, a fuzzy-query strategy is employed over all entries. We as well provide SPARQL interface for professional users to query our knowledge base. 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. Zhishi.me[5] is the first published Chinese large-scale Linking Open Data, which acquires structural information from Chinese Wikipedia, Baidu Baike and Hudong Baike. Similarly, Wang et.al learns an ontology based on category system and properties from Hudong Baike [7]. XLORE is an extension of CKB in multi-language. Utilizing the rich content of multiple online-wikis, XLORE gathers abundant valuable semantic information.

Cross-lingual Knowledge Bases. DBpedia [4] is one of the most widely-used [1,3] cross-lingual knowledge base in the world. It extracts various kinds of structured information from Wikipedia and employs the multi-lingual characteristic of Wikipedia to generate 97 language versions of content. Universal WordNet(UWN) [2] is a large multi-lingual lexical knowledge base built from

WordNet and Wikipedia through sophisticated knowledge extraction, link prediction, information integration, and taxonomy induction. XLORE enriches non-English things by employing other language-version wikis, which eliminates disadvantages of using Wikipedia only. It extracts more classes and properties automatically and validates precise semantic relations by a pruning approach.

7 Conclusion

This paper presents an approach of building a Chinese-English CLKB from multiple wiki sources. We extract structured information and unify data format. Then a cross-lingual 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 supporting keyword search and SPARQL query is provided to access the knowledge base.

Acknowledgement. The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), NSFC (No.61402220), Tsinghua University Initiative Scientific Research Program (No. 20131089256), Science and Technology Support Program (No. 2014BAK04B00), and THU-NUS NExT Co-Lab.

References

1. 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)
2. 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)
3. 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)
4. Mendes, P.N., Jakob, M., Bizer, C.: Dbpedia: a multilingual cross-domain knowledge base. In: LREC, pp. 1813–1817 (2012)
5. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi.me - weaving chinese linking open data. In: Aroyo, L., et al. (eds.) ISWC 2011, Part II. LNCS, vol. 7032, pp. 205–220. Springer, Heidelberg (2011)
6. 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)
7. Wang, Z., Wang, Z., Li, J., Pan, J.Z.: Building a large scale knowledge base from chinese wiki encyclopedia. In: Pan, J.Z., et al. (eds.) JIST 2011. LNCS, vol. 7185, pp. 80–95. Springer, Heidelberg (2012)
8. 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)