Building Large-Scale Bilingual 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
{lijuanzi@}tsinghua.edu.cn

Abstract. Abstract Text

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

1 Introduction

Introduction Text

Current situation of KB Current situation of cross-lingual KB

Knowledge bases such as DBpedia, YAGO, and BabelNet are mainly built upon the multilingual Wikipedia. ..

Some problems are to be addressed:

- The imbalanced size of different Wikipedia languages leads to less data of various languages, which makes it more difficult to build a multi-language knowledge base.
- The number of existed cross-lingual links in Wikipedia is so small that affects the quality of a bilingual knowledge base. Especially there is no obvious links in properties.
- The large but irrigorous category system in Wikipedia causes incorrect semantic relations in taxonomy. For example,

Our main contributions include:

- We propose a method to build a Chinese-English ontology combining four encyclopedias, which are English Wikipedia, Chinese Wikipedia, Baidu and Hudong. The latter three contain large amount of Chinese knowledge, which helps balance information of two languages and enrich the knowledge base.
- We extend the cross-lingual link set by employing a cross-lingual knowledge linking discovery approach for concept and instance, and by analysing templates in Wikipedia for property.
- We prune the original taxonomy, which is extracted from encyclopedia category system, to retrieve more precise SubClassOf and InstanceOf relation.
- 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 encyclopedias which are English Wikipedia, Chinese Wikipedia, Baidu and Hudong. Besides, definitions are proposed in this section. Section 3 presents the extraction approach in concept, instance and property level. Method of initiating cross-lingual link set will also be mentioned. Section 4 propose an extension approach for discovering more cross-lingual relations. Section 6 describes how to prune the taxonomy. Section 7 shows the results of established knowledge base. Section 9 gives the conclusion and future work.

2 Preliminary

In this section, we firstly introduce the four encyclopedias which are used to build and enrich our knowledge base. They are English and Chinese Wikipedia, Baidu and Hudong Encyclopedia. Then we gives some related definitions to formalize our model.

2.1 Encyclopedias

Wikipedia Nowadays, Wikipedia is the largest store of human knowledge. It was launched in 2001 and has hold over 35 million articles in 288 language by 2015. Among these articles, English articles contribute most. There are 4,867,719 articles in English while only over 800,000 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 bad in cross-lingual aspect. For example, DBpedia has 683 concepts in English while only 11 Chinese concepts, that helps little when other poeple want to do cross-lingual research depending on DBpedia. To avoid the imbalance problem, two other Chinese encyclopedias are utilized to enrich Chinese source. Articles in Wikipedia are manually editted by various authors. Wikipedia provides many templates to guide authors composing their context in a regular way. For example, China use template XXX. Moreover, an infobox in one article also follows an infobox template, which maintains a property set of similar articles and display labels on website page.

Other Chinese Encyclopedias There are several large-scale monolingual Chinese Encyclopedias currently. Among those, Baidu Baike and Hudong are the most well-known. Hudong was founded in 2005 and contains more than 12 millions articles with about 9 millions experts' contribution util 2015, April. At the mean time, Baidu Baike maintains more than 11 millions articles. These two resource are similar in article structure, somethimes even in content.

Encyclopedia Page All these four encyclopedias have two important elements, articles and category taxonomy. A taxonomy presents the relation between categories. Usually a category has its sub classes and super classes. Besides, an article belongs to one or more categories. Fig. 2.1 An article describes an topic with

rich information created and modified by several verified editors. The content and relation of an article provide a lot when building a global? knowledge base. In general, there are five elements can be exploited:

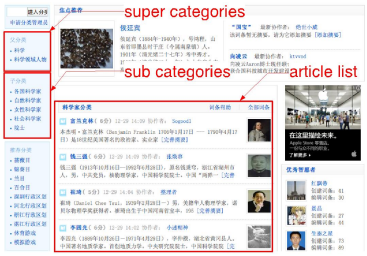

Fig. 1. Taxonomy in Hudong.

Fig. 2. A snap of an article in Baidu Baike

- Title: A Title is the label of topic. Every article has a unique title, which can be used to distinguish each topic.
- Abstract: An abstract is a brief summarize of the article topic. It's always the first paragraph of an article. Usually it can be taken as an important feature to a topic.
- Infobox: Most of articles contain infobox. An infobox maintains structured data which are in general subject-attribute-value triples formalized as a table. Information in this table includes important properties of the topic.
- Links: Links are entries directing 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 below the article page, shown as tags. An article attaches to one or more categories.

Fig. 2.1 shows a snap of an article in Baidu Baike. The five elements mentioned above are annotated.

2.2 Cross-lingual Links

In order to build a bilingual knowledge base, we need the relation between Chinese source and English source, which helps merge different language information into their unique and common topic. In Wikipedia, most articles have language links which guide the reader to the article in specific language under the same

topic. Fig. **??** shows an example of language links in Wikipedia. Taking advantage of such existed links in Wikipedia, we can generate an initial bilingual ontology based on Chinese and English Wikipedia. However, the result is just an ontology based on Wikipedia, like DBpedia. To combine Hudong and Baidu Baike ,which are lack of cross-lingual information but more Chinese source, we employ the approach from [] to discover bilingual links between English Wikipedia and Hudong in Section 5.1

2.3 Definitions

3 Extraction

In order to build a billinual knowledge base, we first set up an ontology to schema information from the four encyclopedia. The ontology includes concepts, which are extracted from category taxonomy, instances, which are defined according to articles and properties, which are based on infobox and also templates assistant. We will describe our building approach as follow.

3.1 Concept Extraction

A concept in ontology is defined as a type of similar instances. For example, the concept of instance *Tsinghua University* is *University* or *Organization*. In general, a concept has super classes and sub classes, which means it has *subClassOf* relation with other classes. Those concepts comprise a taxonomy which presents a backbone of the ontology. In an encyclopedia, a category groups several articles and also has super-categories and sub-categories, just like concept doing. Therefore we can extract concepts based on existed 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 that are typical of Wikipedia. For example, list of ... or tempalte:infobox
- 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-categories 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 representive to a group of instance, therefore it's unwise to retain it as concept.

To retrieve a cleaner and preciser concept taxonomy, we firstly do some refine works as follow:

- Delete specific categories in Wikipedia.

– Delete inconsistent sub-categories and keep the super one.
– Delete categories that relate to less than two articles.

The cleaning works are carried out in all encyclopedias for rule 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 inaccurate samples in the two relations. For example, example. We will prune the taxonomy later.

3.2 Property Extraction

A property is defined as an attribute of instance. It represents the relation between two instances or an instance and its value. We divided properties into two types: datatype properties, which ...; object properties, which .... Considering both content and infobox of an article, we extract two groups of properties, general-properties and Infobox-properties.

General-properties Characteristics of an instance are seen as general-properties, including label, abstract, and url. Those properties describe specific information of an instance. The label property identifies a unique instance, whose value is article title. The abstract property provides a brief description of an instance, whose value is the first paragraph of article. The url property saves the resource of an instance, which is actually a url in Wikipedia or Hudong or Baidu Baike. All of them are datetype properies.

Infobox-properties Attributes acquired from infobox data are considered as Infobox-properties, such as (release date), (directed by) in a movie's infobox. All attributes are relating with a typical value in the infobox, the value maybe a text or a reference, usually a url links to anothor instance. The type of a property, datatype or object, depends on the value. Ordinarily, a plain text value marks the property as datetype while an instance reference determines the property as object. 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 occur some chanllenges when extracting properties from infoboxes:

– There are some
–
–

3.3 Instance Extraction

In encyclopedia, an article describes an unique entity in the world. Therefore we can extract 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, Template documentations and.... Each instance contains relations with concepts and properties. Take the movie (Interstellar) as an example in Fig.**??**. concepts are assigned according to the category tags below the article page. (American science fiction films) is a concept to Interstellar. In the meantime, we obtain label property from article title, which is (Interstellar) and abstract property from the first paragraph. Infobox-properties are also acquired via extracting from infobox in the article. Besides, according to links placing in content, we gain the reference between the current instance to 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 is relationships, containing concept-instance, instance-property-value and instance-instance.

3.4 Cross-lingual Links Extraction

Wikipedia has number cross-lingual links between articles of English and Chinese. By extracting language links from Wikipedia, we can get an initial cross-lingual link set of concepts and instances. However, in property aspect, there is no obvious infobox links between English and its mapping Chinese article. According to the edit mechanism of Wikipedia, we use infobox template to get display label and cross-lingual links of an infobox-property. Fig. **??** shows a typical example of infobox template.

4 Cross-lingual Links Extension

Research of Wang ZhiChun

After the extracting preprocessing, we acquire a series of semi-structured data, including concept information, property pairs, instance information, relations among the three and cross-lingual links. Holding these raw data in hand, we then build an ontology schema for the futher knowledge base.

5 Cross-lingual Knowledge Base Building

To construct an cross-lingual ontology schema with existing semi-structured data, firstly we link the four encyclopedia, which is to say, combining concept, instance and property of the four sources into one if it represents the same thing. Secondly, we prune the taxonomy which generates from concept relationships to retain a more accurate one. At last, we hung instances and properties on to taxonomy to create a complete knowledge base.

5.1 Cross-lingual Linking

combine four ontology into on

5.2 Taxonomy Prune

Research of Wang Zhigang

5.3 Taxonomy Prune

6 Result

6.1 DataSet

The four encyclopedias and their statistics data

6.2 Extracted Knowledge Base

Xlore statistics data.

6.3 Query Interface

Xlore query interface

7 Related Work

8 Conclusion and Future Work

Acknowledgement

Thanks anonymous reviewers for their valuable suggestions that help us improve the quality of the paper. Thanks Prof. Chua Tat-Seng from National University of Singapore for discussion. The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), Tsinghua University Initiative Scientific Research Program (No. 20131089256) and THU-NUS NExT Co-Lab.