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.

我觉得就是language啊

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 affects the quality of a bilingual knowledge base. Especially when there are scarce 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

structure有构建的意思啊

改过，不知道行不行

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 endpoint 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. 这句话留下来了
- 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.
- Both website and SPARQL endpoint is provided for public query operations over our knowledge base.

The rest of the paper is organized as follows. Section 2 introduces the four involved encyclopedias. Besides, the procedure are formalized in this section. Section 3 presents the extraction approach in concept, instance and property level. Section 4 describes the procedure of building a knowledge base using extracted result. Section 5 shows the data situation of established knowledge base. Section 6 induces related work about this paper. Section 7 gives the conclusion.

2 Problem Definition

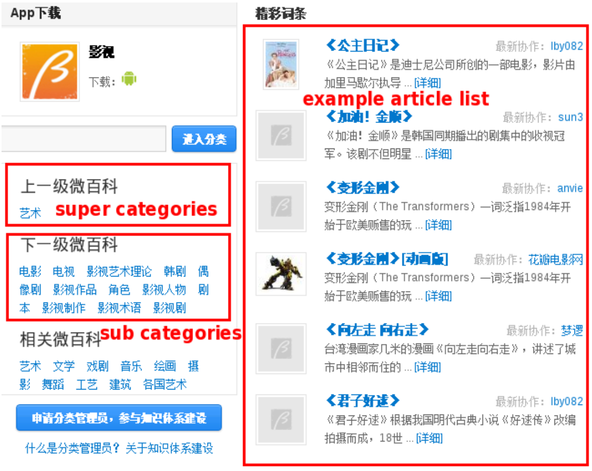
In this section, we firstly introduce the four encyclopedias used to build and enrich our knowledge base. Then we give some related definitions to formalize our ontology model.

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 languages by 2015. Out of these, English articles contribute most. The imbalance of different language articles makes ontologies based on Wikipedia-only behave badly in cross-lingual aspect. Thus, more Chinese encyclopedias are necessary to enrich Chinese source.

Among the large-scale monolingual Chinese encyclopedias currently, 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. Meanwhile, Baidu Baike maintains more than 11 million articles.

Articles from the four sources are similar in structure. Usually they provide two important elements with potential semantic information, category taxonomy and articles. A taxonomy presents the relations between categories. Usually a category has 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:


App下载
影视
下载：
进入分类
上一级微百科
艺术
super categories
下一级微百科
电影 电视 影视艺术理论 韩剧 偶像剧 影视作品 角色 影视人物 剧本 影视制作 影视术语 影视剧
相关微百科
sub categories
艺术 文学 戏剧 音乐 绘画 摄影 舞蹈 工艺 建筑 各国艺术
申请分类管理员，参与知识体系建设
什么是分类管理员？ 关于知识体系建设
精彩词条
example article list

《公主日记》 最新协作：lby082
《公主日记》是迪士尼公司所创的一部电影，影片由加里马歇尔执导 ... [详细]

《加油！金顺》 最新协作：sun3
《加油！金顺》是韩国同期播出的剧集中的收视冠军。该剧不但明星 ... [详细]

《变形金刚》 最新协作：anvie
变形金刚（The Transformers）一词泛指1984年开始于欧美贩售的玩 ... [详细]

《变形金刚》[动画版] 最新协作：花瓣电影网
变形金刚（The Transformers）一词泛指1984年开始于欧美贩售的玩 ... [详细]

《向左走 向右走》 最新协作：梦逻
台湾漫画家几米的漫画《向左走向右走》，讲述了城市中相邻而住的 ... [详细]

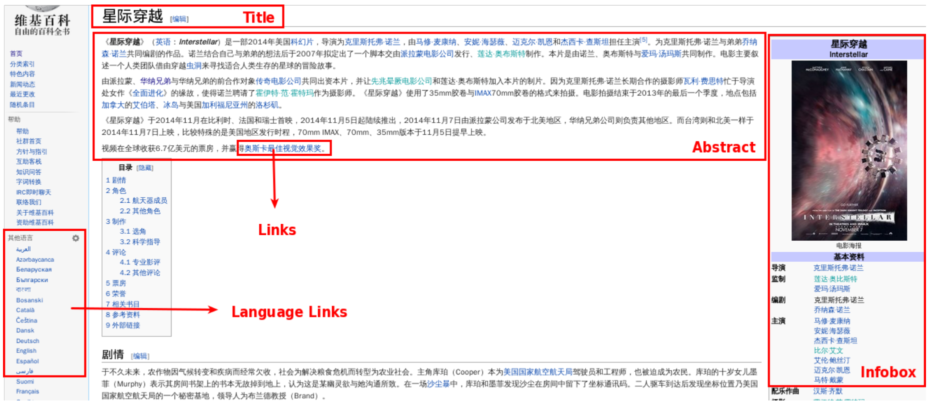
《君子好逑》 最新协作：lby082
《君子好逑》根据我国明代古典小说《好逑传》改编拍摄而成，18世 ... [详细]

Fig. 1. Taxonomy in Hudong

- Title: A Title is the label of an entity, which is unique so that it can be used to distinguish entities.
- Abstract: An abstract is a brief summary of the entity. It's always the first paragraph of an article.
- Infobox: Most of articles contain infobox. An infobox maintains structured data which are 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 belongs to one or more categories.

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

Notably, articles in Wikipedia follow templates specified by Wikipedia when being editted. A template defines items that a group of article should fill. Besides,



星际穿越 [编辑]

《星际穿越》（英语：*Interstellar*）是一部2014年美国科幻片，导演为克里斯托弗·诺兰，由马修·麦康纳、安妮·海瑟薇、迈克尔·凯恩和杰西卡·查斯坦担任主演[5]。为克里斯托弗·诺兰与弟弟乔纳森·诺兰共同编剧的作品。诺兰结合自己与弟弟的想法后于2007年拟定出了一个脚本交由派拉蒙电影公司发行、莲达·奥布斯特制作。本片是由诺兰、奥布斯特与爱玛·汤玛斯共同制作。电影主要叙述一个人类团队借由穿越虫洞来寻找适合人类生存的星球的冒险故事。

由派拉蒙、华纳兄弟与华纳兄弟的前合作对象传奇电影公司共同出资本片，并让先兆晕厥电影公司和莲达·奥布斯特加入本片的制片。因为克里斯托弗·诺兰长期合作的摄影师瓦利·费思特忙于导演处女作《全面进化》的缘故，使得诺兰聘请了霍伊特·范·霍特玛作为摄影师。《星际穿越》使用了35mm胶卷与IMAX70mm胶卷的格式来拍摄。电影拍摄结束于2013年的最后一个季度，地点包括加拿大的艾伯塔、冰岛与美国加利福尼亚州的洛杉矶。

《星际穿越》于2014年11月在比利时、法国和瑞士首映，2014年11月5日起陆续推出，2014年11月7日由派拉蒙公司发布于北美地区，华纳兄弟公司则负责其他地区。而台湾则和北美一样于2014年11月7日上映，比较特殊的是美国地区发行时程，70mm IMAX、70mm、35mm版本于11月5日提早上映。

视频在全球收获6.7亿美元的票房，并赢得奥斯卡最佳视觉效果奖。

目录 [隐藏]

1 剧情
2 角色
2.1 航天器成员
2.2 其他角色
3 制作
3.1 选角
3.2 科学指导
4 评论
4.1 专业影评
4.2 其他评论
5 票房
6 荣誉
7 相关书目
8 参考资料
9 外部链接

剧情 [编辑]

于不久未来，农作物因气候转变和疾病而经常欠收，社会为解决粮食危机而转型为农业社会。主角库珀（Cooper）本为美国国家航空航天局驾驶员和工程师，也被迫成为农民。库珀的十岁女儿墨菲（Murphy）表示其房间书架上的书本无故掉到地上，认为这是某幽灵欲与她沟通所致。在一场沙尘暴中，库珀和墨菲发现沙尘在房间中留下了坐标通讯码。二人驱车到达后发现坐标位置乃美国国家航空航天局的一个秘密基地，领导人为布兰德教授（Brand）。

星际穿越
Interstellar

电影海报

基本资料

导演	克里斯托弗·诺兰
监制	莲达·奥比斯特 爱玛·汤玛斯
编剧	克里斯托弗·诺兰 乔纳森·诺兰
主演	马修·麦康纳 安妮·海瑟薇 杰西卡·查斯坦 比尔·艾文 艾伦·鲍丝汀 迈克尔·凯恩 马特·戴蒙
配乐作曲	汉斯·齐默

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

infoboxes are generated based on templates too. For example, The infobox in film 星际穿越(Interstellar) is edited according to the Template *Infobox film*, which maintains a property set of films.

Moreover, in Wikipedia, some article pages have language links which help readers switch to other language-version 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.

2.2 Definitions

Here we list some definitions involved in the later sections:

Definition 1: An 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.

According to Section 2.1, the elements of each article a can be defined as follow:

$$a =< Ti(a), Ab(a), Li(a), In(a), C(a), U(a) > \tag{1}$$

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

The topic described by a is defined as Entitye. Ariticles from different sources may talk about the same entity. In such a situation, we say the articles are linked.

Definition 2: As to an article a containing multi-language content in Wikipedia, L_e and L_z denotes its article links, usually titles, in English and Chinese separately. Thus, to a cl in CL, $cl(a) =< L_e(a), L_z(a) >$

Definition 3: An Infobox $In(a)$ 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 webpage. 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.

Definition 4: An ontology is a formal specification of a group of entitie. In our work, an ontology is described as a 6-tuple:

$$O =< C, I, P, H^C, H^I, H^P > \tag{2}$$

where C, I, P are the sets of concepts, instances, and properties, respectively. H^C, H^I, H^P represents the hierarchical relationships of concept-concept, concept-instance, instance-property. A Taxonomy$T$ is comprised by H^C and H^I, which seperately repesent $SubClassOf$ and $InstanceOf$ relations.

2.3 Cross-lingual Knowledge Base Building

A cross-lingual knowledge base is a database conform to a cross-lingual ontology. Taking advantage of language links in CL, several mono-ontologies generated from various sources can be merge into one cross-lingual ontology. Thus a knowledge base can be defined as:

$$KB =< O_i, CL > \tag{3}$$

where O_i denotes the ith mono-ontology, CL represents the cross-lingual link set. We build the cross-lingual ontology according to the procedure in Fig. 2.3 We extract 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 get four datasets with concept lists, instance lists, property lists and their relations of each source. Meanwhile, an initial Chinese-English

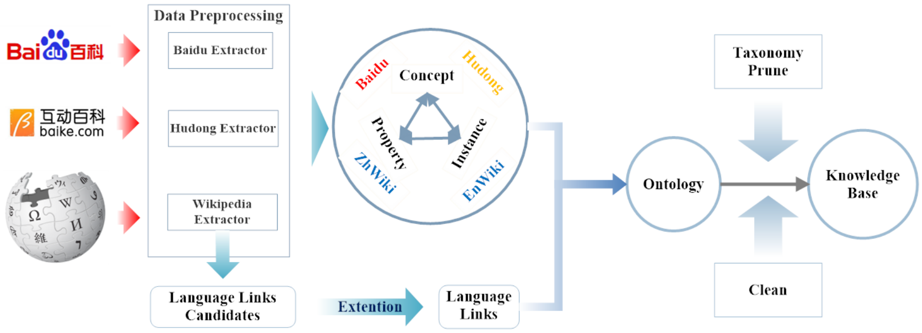


Fig. 3. Procedure of building our cross-lingual knowledge base

language-link set is generated, which will be increased by employing 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.

3 Semantic Data Extraction

We prepare data for the knowledge base by a series of extracting 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.

3.1 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.

这部分留下来了

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.

需要再斟酌

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 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 more precise concept taxonomy, we do some refining works as follow:

- Delete list categories or template categories in Wikipedia.
- Delete categories that relate to less than two articles.

The cleaning work is carried out consistently in all encyclopedias when extracting. Remaining categories comprise an original concept hierarchical relationshipH^C of T. However, among the relations there are still incorrect samples. For example, *Tsinghua University* is not an entity of *Haidian District*, but relates to. Thus we will prune the taxonomy later.

这个example是一个真实的例子啊

3.2 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.

留下来了

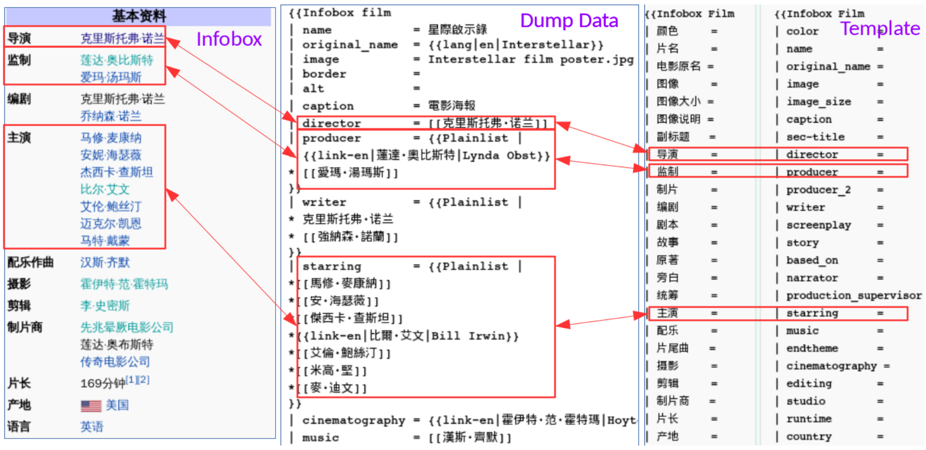
General-properties Characteristics of an entity are regarded as general-properties, including label, abstract, and url. General-properties describe specific information of an entity. Given an article$a$, title$Ti(a)$ is the value of label property, which identifies a unique entity. $Ab(a)$, usually the first paragraph of article, corresponding the abstract property, which provides a brief description of an entity. The url property saves the resource url of an entity, whose value is $U(a)$. All of them are datatype properties.

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. Each attribute in the infobox is related to a 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 上映时间(release date) can be defined as a datatype property as its value is a datetime string. Meanwhile, 导演(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:

- In Wikipedia, the attribute label displayed in the webpage infobox is inconsistent with it in the published dump file. Fig.3.2 gives a mapping result of display label and dump label in *Interstellar*' infobox. The left is infobox, the middle is a snap from dump file in Wikipedia. As we've seen, attribute label in infobox is different from it in dump file. As a result we have to explore the display labels in Wikipedia rather than dump labels extracted from data file.
- There are special characters in labels. Wikipedia usually use hyphen ”-” or dot ”•” to mark sublabels. For example, *population* property has 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. For example, movie infobox follows template *Infobox film*, which is shown on the right of Fig.3.2, where both display labels and dump labels come from. When a dump label in triple-bracket occurs in dump file, we replace it by its mapping display



基本资料	
导演	克里斯托弗·诺兰
监制	莲达·奥比斯特 爱玛·汤玛斯
编剧	克里斯托弗·诺兰 乔纳森·诺兰
主演	马修·麦康纳 安妮·海瑟薇 杰西卡·查斯坦 比尔·艾文 艾伦·鲍丝汀 迈克尔·凯恩 马特·戴蒙
配乐作曲	汉斯·齐默
摄影	霍伊特·范·霍特玛
剪辑	李·史密斯
制片商	先兆晕厥电影公司 莲达·奥布斯特 传奇电影公司
片长	169分钟[1][2]
产地	美国
语言	英语

```
{{Infobox film
| name             = 星際啟示錄
| original_name    = {{lang|en|Interstellar}}
| image            = Interstellar film poster.jpg
| border           =
| alt              =
| caption          = 電影海報
| director         = [[克里斯托弗·诺兰]]
| producer         = {{Plainlist |
 {{link-en|蓮達·奧比斯特|Lynda Obst}}
* [[愛瑪·湯瑪斯]]
}}
| writer           = {{Plainlist |
* 克里斯托弗·诺兰
* [[強納森·諾蘭]]
}}
| starring         = {{Plainlist |
*[[馬修·麥康納]]
*[[安·海瑟薇]]
*[[傑西卡·查斯坦]]
*{{link-en|比爾·艾文|Bill Irwin}}
*[[艾倫·鮑絲汀]]
*[[米高·堅]]
*[[麥·迪文]]
}}
| cinematography = {{link-en|霍伊特·范·霍特瑪|Hoyt
| music            = [[漢斯·齊默]]
```

```
{{Infobox Film          {{Infobox Film
| 颜色      =           | color            =
| 片名      =           | name             =
| 电影原名 =             | original_name    =
| 图像      =           | image            =
| 图像大小 =             | image_size       =
| 图像说明 =             | caption          =
| 副标题    =           | sec-title        =
| 导演      =           | director         =
| 监制      =           | producer         =
| 制片      =           | producer_2       =
| 编剧      =           | writer           =
| 剧本      =           | screenplay       =
| 故事      =           | story            =
| 原著      =           | based_on         =
| 旁白      =           | narrator         =
| 统筹      =           | production_supervisor
| 主演      =           | starring         =
| 配乐      =           | music            =
| 片尾曲    =           | endtheme         =
| 摄影      =           | cinematography =
| 剪辑      =           | editing          =
| 制片商    =           | studio           =
| 片长      =           | runtime          =
| 产地      =           | country          =
```

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

label. After convert all the dump labels, we make a filter to redress the label text.

3.3 Instance Extraction

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

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

After the preprocessing above, we harvest two types of information. One is the characteristics of instance, including $Ti(a)$, $Ab(a)$ of article$a$. The other includes relationships, containing H^I and H^P associated with article a.

4 Cross-lingual Integration

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

是merge呀

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.

4.1 Cross-lingual Linking

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% between English Wikipedia and Baidu Baike.

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 means T_e and T_z are cross-lingual pairs, find the display labels mapping the same dump label. That is to say, to two pairs, p_e in T_e and p_z in T_z, if tl_e is equal to tl_z, $< dl_e, dl_z >$ are property cross-lingual; (2) Given the English and Chinese infoboxes of a matched instance, compare their templates, which are English template$T_e(a_i)$ and a Chinese template$T_z(a_j)$ in which a_i and a_j direct to the same entity, find the matched display labels 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.

定义了

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 instance, we merge instances by the following method:

Step 1 Given an instance extracted from Chinese Wikipedia, find the entity in Hudong and Baidu Baike with the same title.
Step 2 Given an instance extracted from Hudong, find the entity which is not in Chinese Wikipedia but in Baidu Baike with the same title.
Step 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.
Step 4 To a $L_z(a)$, 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.
Step 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.

4.2 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 prefix/suffix 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.

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

5 Result

5.1 Extracted Knowledge Base

We collect the resources from 4 wiki sites, English and Chinese Wikipedia dump files in 2015 April, Hudong html pages until 2014 May, and Baidu html pages until 2014 September. 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. Table1 the result we get after elementary extraction in 4 different wiki sources.

Table 1. Statistics of elementary extraction result

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

To build a knowledge base, we create URIs to identify each element and provide corresponding information if users look up elements over HTTP protocol to achieve the knowledge base. Table 2 lists the defined URIs.

Table 2. 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

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.

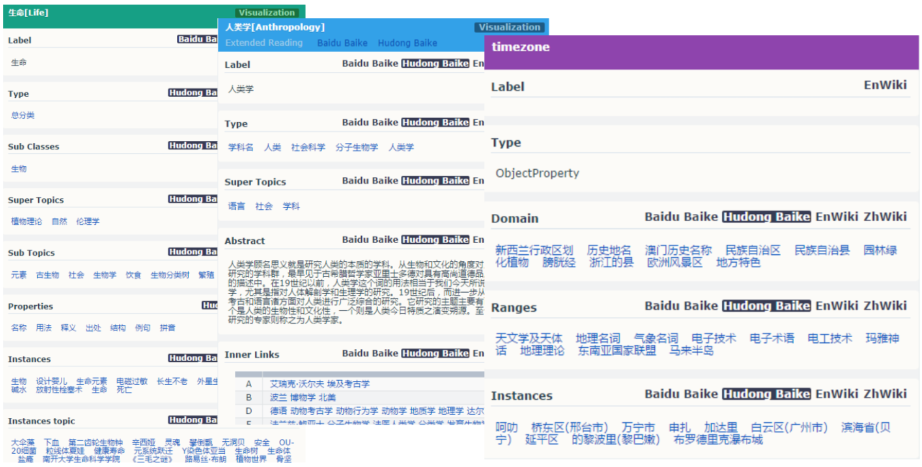
Table 3. Statistics of our knowledge base

	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.

5.2 Web Access to Knowledge Base

We provide a website to present an intuitive graph of our knowledge in the forms of class, instance and property. Fig.5.2 shows sample pages of the integrated data. 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 provide two ways to access our knowledge base: via the search engine shown



生命[Life] Visualization

Label Baidu Ba

生命

Type Hudong Ba

总分类

Sub Classes Hudong Ba

生物

Super Topics Hudong Ba

植物理论 自然 伦理学

Sub Topics Hudong Ba

元素 古生物 社会 生物学 饮食 生物分类树 繁殖

Properties Huc

名称 用法 释义 出处 结构 例句 拼音

Instances Hudong Ba

生物 设计婴儿 生命元素 电磁过敏 长生不老 外星生
碱水 放射性栓塞术 生命 死亡

Instances topic Hudong Ba

大伞藻 下血 第二齿轮生物钟 辛西娅 灵魂 攀倒甑 无洞贝 安全 OU-
20细菌 粒线体夏娃 健康寿命 元系统跃迁 Y染色体亚当 生命树 生命体
盐瘾 南开大学生命科学学院 《三毛之谜》 路易丝·布朗 植物世界 骨坚

人类学[Anthropology] Visualization

Extended Reading Baidu Baike Hudong Baike

Label Baidu Baike Hudong Baike En

人类学

Type Baidu Baike Hudong Baike En

学科名 人类 社会科学 分子生物学 人类学

Super Topics Baidu Baike Hudong Baike En

语言 社会 学科

Abstract Baidu Baike Hudong Baike En

人类学顾名思义就是研究人类的本质的学科。从生物和文化的角度对
研究的学科群，最早见于古希腊哲学家亚里士多德对具有高尚道德品
的描述中。在19世纪以前，人类学这个词的用法相当于我们今天所说
学，尤其是指对人体解剖学和生理学的研究。19世纪后，而进一步从
考古和语言诸方面对人类进行广泛综合的研究。它研究的主题主要有
个是人类的生物性和文化性，一个则是人类今日特质之演变朔源。至
研究的专家则称之为人类学家。

Inner Links Baidu Baike Hudong Baike En

A	艾瑞克·沃尔夫 埃及考古学
B	波兰 博物学 北美
D	德语 动物考古学 动物行为学 动物学 地质学 地理学 达尔

timezone

Label EnWiki

Type

ObjectProperty

Domain Baidu Baike Hudong Baike EnWiki ZhWiki

新西兰行政区划 历史地名 澳门历史名称 民族自治区 民族自治县 园林绿
化植物 膀胱经 浙江的县 欧洲风景区 地方特色

Ranges Baidu Baike Hudong Baike EnWiki ZhWiki

天文学及天体 地理名词 气象名词 电子技术 电子术语 电工技术 玛雅神
话 地理理论 东南亚国家联盟 马来半岛

Instances Baidu Baike Hudong Baike EnWiki ZhWiki

呵叻 桥东区(邢台市) 万宁市 申扎 加达里 白云区(广州市) 滨海省(贝
宁) 延平区 的黎波里(黎巴嫩) 布罗德里克瀑布城

Fig. 5. Sample Pages of Class, Instance and Property

in Fig.**??**, or via SPARQL endpoint shown in Fig.**??**. For users who know little about semantic web, they can query by inputing related text into searchbox and search to get entities with similar label. 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 and cross-lingual knowledge linking methods are referred to.

6.1 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 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 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.

6.2 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]

7 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.: 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)