# XLore: A Large-scale English-Chinese Bilingual Knowledge Graph

Zhigang Wang†, Juanzi Li†, Zhichun Wang‡, Shuangjie Li†, Mingyang Li†, Dongsheng Zhang†, Yao Shi†, Yongbin Liu†, Peng Zhang†, and Jie Tang†

† DCST, Tsinghua University, P.R. China
{wzhigang, ljz, lsj, lmy, zds, sy, lyb, zp, tangjie}@keg.cs.tsinghua.edu.cn
‡ CIST, Beijing Normal University, P.R. China
zcwang@bnu.edu.cn

**Abstract.** Current Wikipedia-based multilingual knowledge bases still suffer the following problems: (i) the scarcity of non-English knowledge, (ii) the noise in the semantic relations and (iii) the limited coverage of equivalent cross-lingual entities. In this demo, we present a large-scale bilingual knowledge graph named XLore, which has adequately solved the above problems.

## 1 Introduction

Multilingual knowledge bases are important for the globalization of knowledge sharing. Knowledge bases such as DBpedia[1], YAGO[2], and BabelNet[3] are mainly built upon the multilingual Wikipedia. Some problems are to be addressed: (i) The imbalanced sizes of different Wikipedia language versions lead to the highly imbalanced knowledge distribution in different languages. Knowledge encoded in non-English languages is much less than those in English. (ii) The inconsistency of the large category system in Wikipedia causes incorrect semantic relations between concepts that are defined based on categories. For example, "Wikipedia-books-on-people is the `subCategoryOf` People" will lead to the wrong "Wikipedia-books-on-people is `subClassOf` People" in DBpedia's SKOS schema. (iii) Integrated by directly using cross-lingual links in Wikipedia, the amount of integrated multilingual knowledge totally depends on these existing cross-lingual links.

In this demo, we present an English-Chinese bilingual knowledge graph named XLore to adequately solve the above problems. We use much larger heterogenous online wikis to enrich the Chinese knowledge, utilize a classification-based method to correctly semantify the wikis' category systems, and employ a cross-lingual knowledge linking approach to find new cross-lingual links between entities. Besides, we use a cross-lingual structured knowledge extraction method to enrich the semantic relations.

[1] http://dbpedia.org/
[2] http://www.mpi-inf.mpg.de/yago-naga/yago/
[3] http://lcl.uniroma1.it/babelnet/

To the best of our knowledge, XLore is the first large-scale cross-lingual knowledge graph with balanced amount of Chinese-English knowledge. XLore gives a new way for building such a knowledge graph across any two languages.

## 2 Approach

As shown in Figure 1, the building of XLore contains three stages: (1) *Data Preprocessing*: First we collect and clean the data sets from four online wikis, namely English Wikipedia, Chinese Wikipedia, Baidu Baike and Hudong Baike. (2) *Knowledge Graph Building*: Next, we learn the cross-lingual ontology as follows: semantify the online wikis to predict correct semantic relations, conduct cross-lingual knowledge linking to integrate the heterogenous wikis together, and extract the structured knowledge to enrich more relations in the graph. (3) *Knowledge Query*: Finally, we construct an online system for knowledge acquisition.

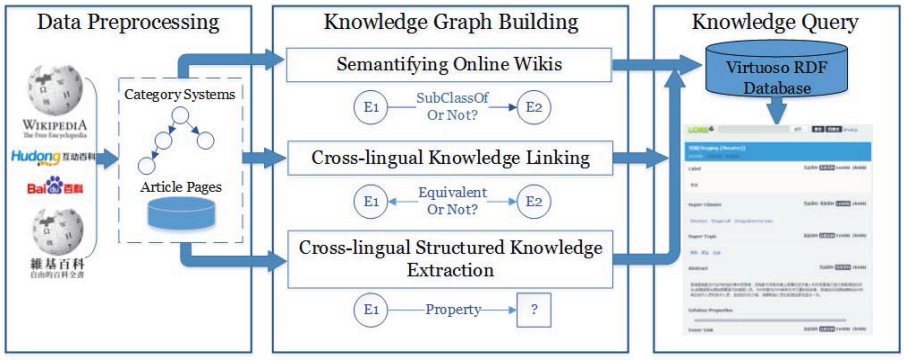



**Fig. 1.** Overview

**Semantifying Online Wikis** To semantify the online wikis, we are to predict the correct `subClassOf` and `instanceOf` relations between two entities. We view both the correct `subClassOf` and the `instanceOf` relations as `is-a` relations. Table 1 shows some examples about the semantic relations.

**Table 1.** Examples of Semantic Relations

| Entity 1 | Relation | Entity 2 | Right or Wrong |
|---|---|---|---|
| European Microstates | `instanceOf` | Microstates | Right |
| European Microstates | `instanceOf` | Europe | Wrong |
| 教育人物(Educational Person) | `subClassOf` | 人物(Person) | Right |
| 教育人物(Educational Person) | `subClassOf` | 教育(Education) | Wrong |

Formally, we learn two series of functions $g_1$ (for English) and $g_2$ (for Chinese) to predict the probabilities to be an `is-a` relation between two entities. We define some literal and structural features and train the Logistic Regression models. The most important features are the head words' singular/plural forms of English entities and the substring relationship between the labels of Chinese entities. By iteratively expanding the training data sets, both the functions achieve over 90.48% F1-score. To keep the semantic relatedness, we treat the incorrect relations as the `subTopicOf` relations and import these relations into the RDF database too.

**Cross-lingual Knowledge Linking via Concept Annotation** To integrate the knowledge in different languages, we proposed learning based approaches for linking equivalent entities in different languages [1][2]. Several features are defined based on the link structures in wikis to assess the similarities between two different entities. Then learning models are trained based on the already known cross-lingual links in Wikipedia, which afterward predict new equivalent entity pairs. In order to find desired number of new cross-lingual links, we use concept annotation to enrich the inner links within wikis, which improves the knowledge linking approach considerably. The knowledge linking process as a whole can execute iteratively, resulting in large number of new cross-lingual links.

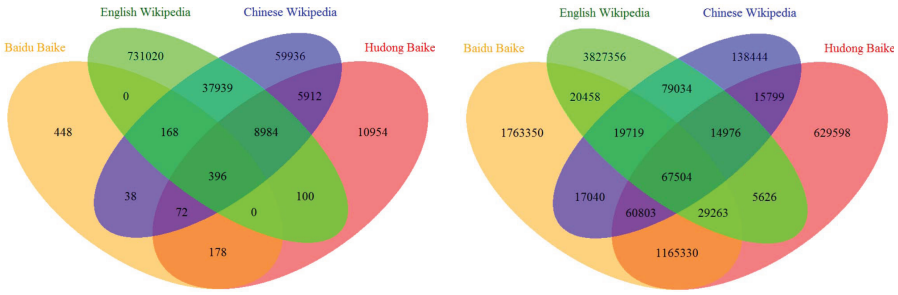
**Cross-lingual Structured Knowledge Extraction** To enrich kinds of relations in XLore, we apply our cross-lingual knowledge extraction framework named WikiCiKE to complete the missing infoboxes [3]. WikiCiKE is based on the hypothesis: one can use the rich auxiliary (e.g. English) information to assist the target (e.g. Chinese) infobox extraction. We treat this task as a transfer learning-based binary classification problem. Given an attribute in the target wiki, WikiCiKE automatically generates the cross-lingual training data and learns the extractor using TrAdaBoost model. Finally, WikiCiKE uses the learned extractor to extract the missing value from the unstructured article texts. Our experiments in [3] demonstrate that WikiCiKE significantly outperforms the monolingual knowledge extraction method and the translation-based method.

## 3 XLore System

We construct a unified knowledge graph in the form of RDF and use the OpenLink Virtuoso server[4] for systematical data management. Using the proposed approach, XLore harvests 856,146 classes , 71,596 properties and 7,854,301 instances across English and Chinese. Figure 2 gives a brief statistics the number of linked entities from different online wikis.

We also deploy an online system to illustrate our XLore. As shown in Figure 3, the system supports the keyword-based or SPARQL queries, gives the statistical information, offers visualization demonstrations, etc. A live demonstration of the system can be found at `http://www.youtube.com/watch?v=QKA-RYFfztA`. We invite the readers to try our XLore prototype at `http://xlore.org`.

[4] `http://virtuoso.openlinksw.com/`




(a) *Number of Linked Concepts* (b) *Number of Linked Instances*

**Fig. 2.** Statistics of the Linked Entities.

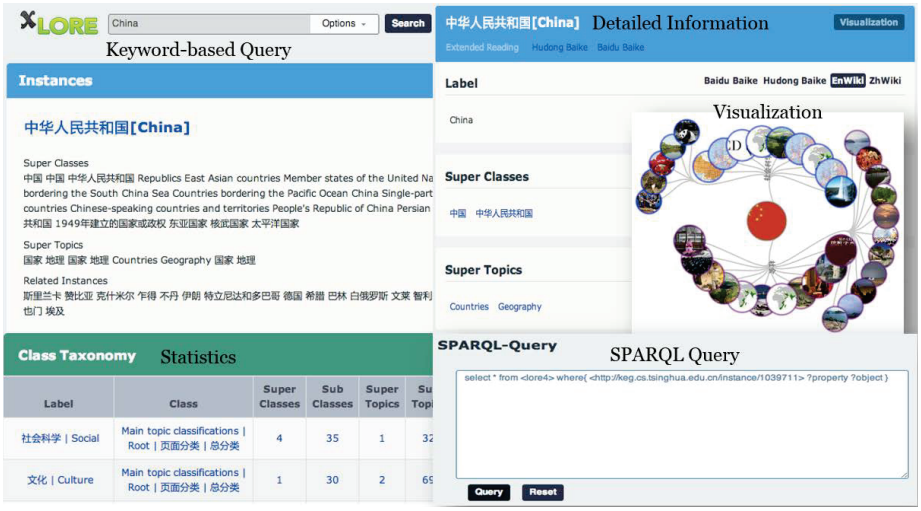



**Fig. 3.** Interface of XLore System

## References

1. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. WWW'12
2. Wang, Z., Li, J., Tang, J.: Boosting cross-lingual knowledge linking via concept annotation. IJCAI'13
3. Wang, Z., Li, Z., Li, J., Tang, J., Z.Pan, J.: Transfer learning based cross-lingual knowledge extraction for wikipedia. ACL'13