DATA PAPER

XLORE2: Large-scale Cross-lingual Knowledge Graph Construction and Application

Hailong Jin, Chengjiang Li, Jing Zhang, Lei Hou†, Juanzi Li & Peng Zhang

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Keywords: Knowledge base completion; Knowledge linking; Property matching; Taxonomy alignment; Type inference; Entity linking

Citation: H. Jin, C. Li, J. Zhang, L. Hou, J. Li, & P. Zhang. XLORE2: Large-scale cross-lingual knowledge graph construction and application. Data Intelligence 1(2019), 77-98. doi: 10.1162/dint_a_00003
Received: April 23, 2018; Revised: September 10, 2018; Accepted: September 18, 2018

ABSTRACT

Knowledge bases (KBs) are often greatly incomplete, necessitating a demand for KB completion. Although XLORE is an English-Chinese bilingual knowledge graph, there are only 423,974 cross-lingual links between English instances and Chinese instances. We present XLORE2, an extension of the XLORE that is built automatically from Wikipedia, Baidu Baike and Hudong Baike. We add more facts by making cross-lingual knowledge linking, cross-lingual property matching and fine-grained type inference. We also design an entity linking system to demonstrate the effectiveness and broad coverage of XLORE2.

1. INTRODUCTION

Wikipedia has become one of the most accessible online encyclopedias. It has extremely high language coverage, containing articles in 298 languages. Among them, the English version of Wikipedia owns more than 5.6 million articles, sitting in the first position. "Everyone can edit" makes its knowledge constantly increase and evolve. However, knowledge in Wikipedia is in the form of free text or attribute-value pairs in infobox. Wikipedia's vast knowledge inspires the emergence of many knowledge base (KB) projects that structure knowledge and link knowledge in different languages.

Several projects construct KBs from Wikipedia, e.g., DBpedia [1], YAGO [2] and BabelNet [3]. Nevertheless, they have different focuses. YAGO pays more attention to the semantic consistency of the

† Corresponding author: Lei Hou (Email: houlei@tsinghua.edu.cn; ORCID: 0000-0002-8907-3526).

same knowledge in different languages. DBpedia does much work on the extraction and alignment of cross-lingual fact triples. BabelNet concentrates on the entity concepts, senses and synsets.

The imbalanced size of different Wikipedia language versions apparently leads to the highly imbalanced knowledge distribution in different languages. This is reflected in the KBs that are based on this imbalance, as knowledge encoded in non-English languages is much less than those in English. To address this issue, XLORE has become the first large-scale cross-lingual KB with a balanced amount of Chinese-English knowledge [4]. It gives a new way for building a knowledge graph across any two languages by utilizing cross-lingual links in Wikipedia. Although XLORE already has a relatively balanced amount of bilingual knowledge, there are still a large number of missing facts that need to be supplemented. After reviewing the quality of XLORE, there are clearly three kinds of facts that require enhancement:

1). The number of cross-lingual links between English instances and Chinese instances is limited. Discovering more cross-lingual links is beneficial to knowledge sharing across different languages;
2). Each language version maintains its own set of infoboxes with their own set of attributes, as well as sometimes providing different values for corresponding attributes. Therefore, attributes in different languages must be matched if we want to get coherent knowledge;
3). The type information of an instance is incomplete. For example, *Yao Ming* should not only be assigned with *Person*, *Athlete* and *Basketball Player*, but also *Businessman*.

Completing these three types of missing facts is a very challenging task. Existing cross-lingual knowledge linking discovery methods heavily depend on the number of existing cross-lingual links. It is a fact that the cross-lingual links in Wikipedia are quite sparse. Existing cross-lingual property matching methods have high precision. But the number of aligned properties is quite small for such a large-scale KB. Existing type inference methods require creation and maintenance of large-scale highly-qualified annotated corpora, which are often difficult to obtain.

In this paper, we present XLORE2, an extension of XLORE, as a holistic approach to the creation of a large-scale English-Chinese bilingual KB, to adequately answer the above problems.

Our approach applies the *cross-lingual knowledge linking* method to find more cross-lingual links between equivalent instances in different languages and the *fine-grained type inference* method to assign specific types for those instances without type information. Further, we perform *subClassOf* and *instanceOf* relations validation in XLORE2 in order to build a high-quality taxonomy. Moreover, in *cross-lingual property matching*, we investigate several effective features and propose entity-attribute factor graphing to find the corresponding attributes between English and Chinese. This strategy uncovers many more facts by completing the attribute knowledge, and addresses to a large extent the obstacle of language imbalance. Last but not least, we design an efficient entity linking system *XLink*, which links the "mentions" in a document to the various entities in XLORE2. As a result, XLORE2 reveals significantly more facts when compared with XLORE.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents the framework of XLORE2. Section 4 introduces our approaches in cross-lingual knowledge building. Section 5 introduces our methods of data quality improvement. Section 6 presents some practical applications of XLORE2. Section 7 gives some statistical analysis of XLORE2. Section 8 then makes a conclusion.

2. RELATED WORKS

Several works have integrated the multilingual data from Wikipedia, but with different focuses.

DBpedia[①]: Using Semantic Web and Linked Data technologies, DBpedia is a crowd-sourced community effort to extract structured, multilingual knowledge from the information created in various Wikimedia projects. The DBpedia knowledge bases are extracted from 125 Wikipedia editions. Altogether the latest DBpedia (2016-10) release consists of 13.1 billion pieces of information (RDF triples) out of which 1.7 billion were extracted from the English edition of Wikipedia, 6.6 billion were extracted from other language editions and 4.8 billion from Wikipedia Commons and Wikidata. The DBpedia project maps Wikipedia infoboxes from 27 different language editions into a single shared ontology consisting of 760 classes, 1,105 object properties, 1,622 datatype properties and 132 specialized datatype properties. In addition to the regular releases, the project maintains a live knowledge base which is updated whenever a page in Wikipedia changes. DBpedia is connected to many resources, e.g., DBpedia offers the YAGO type hierarchy as an alternative to the DBpedia ontology and sameAs links are provided in both directions.

YAGO[②]: YAGO is an extensible semantic KB with high coverage and enhanced quality, derived from Wikipedia, WordNet and GeoNames. YAGO uses the categories in Wikipedia to infer type information about an entity and then links this type information to WordNet so as to pursue the coherence of knowledge. It contains more than 1 million entities and 5 million facts. In YAGO2 some declarative extraction rules were introduced in order to gather and integrate temporal, spatial and semantic information from resources. This kind of space- and time- awareness results in the enrichment of entity-relationship-oriented facts along the dimensions of time and space. YAGO3 maps multilingual infobox attributes to canonical relations, merging equivalent entities into canonical entities by way of help from Wikidata. It can achieve a precision of 95%-100% in attribute mapping and thus results in a gain of roughly 1 million new entities and 7 million new facts over the original English-only YAGO. Generally one of the main differences between YAGO and XLORE2 is that XLORE2 applies the cross-lingual entity alignment method and cross-lingual attribute alignment method to merge equivalent entities and equivalent attributes automatically. YAGO does not align different extractions from different Wikipedias, but rather different Wikipedias with a central clean KB.

Wikidata[③]: Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It intends to provide a common source of data which can be used by Wikimedia projects such as Wikipedia,

① http://wiki.dbpedia.org/
② https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
③ https://www.wikidata.org/wiki/Wikidata:Main_Page

and by anyone else, under a public domain license. This is similar to the way Wikimedia Commons provides storage for media files and access to those files for all Wikimedia projects, all of which are freely available for reuse. Wikidata is powered by the software Wikibase. Wikidata currently contains 45,817,125 items. 655,389,411 edits have been made since the project launch.

BabelNet[④]: BabelNet is a very large wide-coverage multilingual semantic network built from Wikipedia and WordNet, which is quite similar to the YAGO project. The key of the methodology is the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia via an automatic mapping. Inevitably, there are lexical gaps in resource-poor languages. BabelNet then fills the lexical gaps with the aid of machine translation. The result of the mapping process can be defined as an encyclopedic dictionary. This approach provides concepts and instances lexicalized in different languages and is connected with a large number of semantic relations. BabelNet has now reached version 4.0. It covers 284 languages and has 16 million multilingual synsets and 832 million senses. The number of concepts, named entities and lexico-semantic relations exceeds 6.1 million, 9.6 million and 1.3 billion, respectively. The knowledge encoded in BabelNet can be used to perform knowledge-rich, graph-based word sense disambiguation both in a monolingual and multilingual setting.

Summary: All the above KBs are built upon the multilingual Wikipedia. Existing multilingual KBs are lack of Chinese knowledge and suffer from the imbalanced knowledge distribution in different Wikipedia language versions. XLORE is the first large-scale cross-lingual KB with a balanced amount of Chinese-English knowledge. XLORE2 seeks to improve the data quality of XLORE while inferring missing facts based on existing pairs in XLORE.

④ http://babelnet.org/

3. METHODOLOGY

Following the framework shown in Figure 1, the construction of XLORE2 contains four stages:

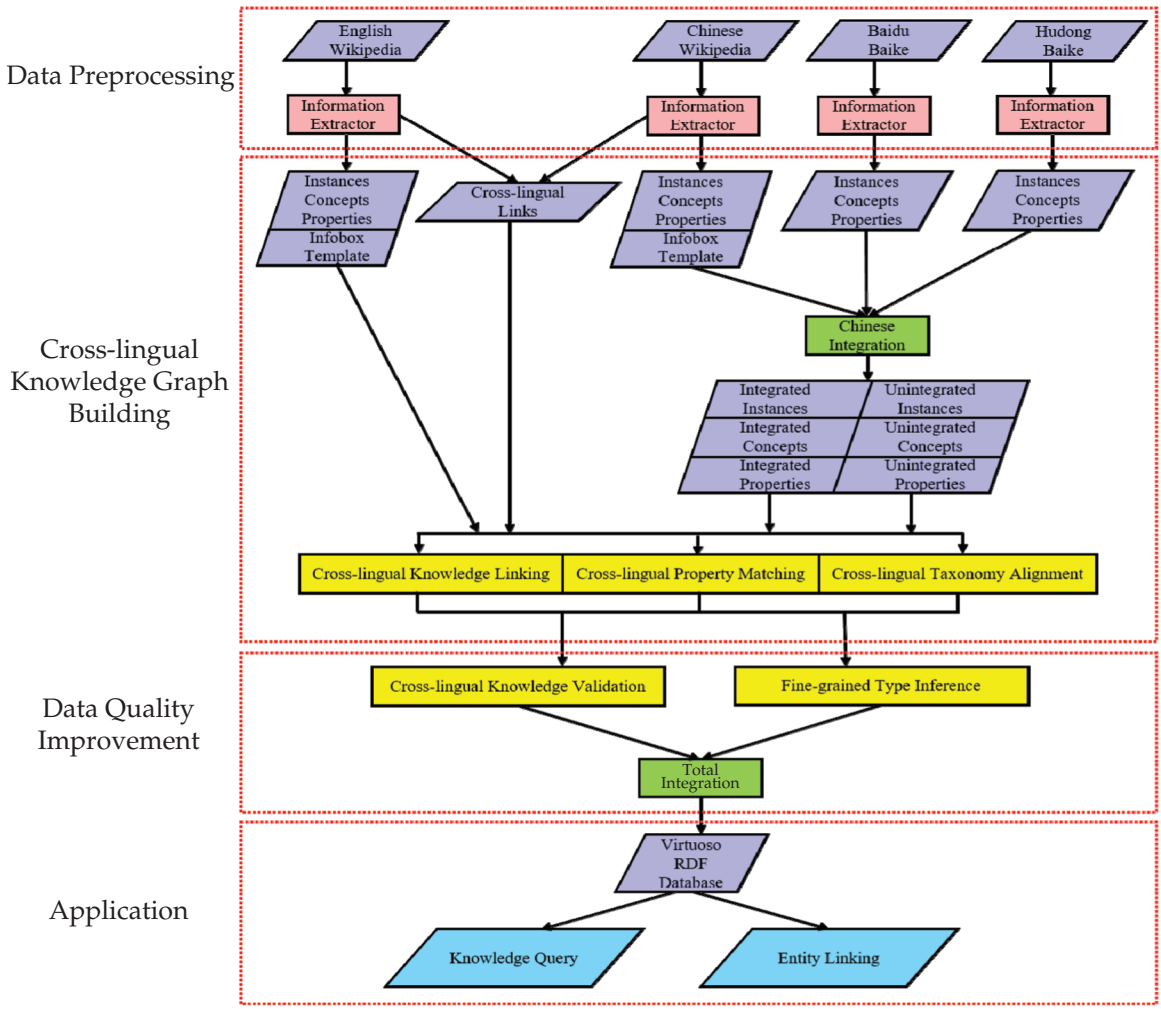

Figure 1. The framework of XLORE2.

1). Data Preprocessing: First, we gather and clean the data from four online wikis, i.e., English Wikipedia, Chinese Wikipedia, Baidu Baike and Hudong Baike. Our extractors parse out instances, concepts, properties and templates. It is worth mentioning that the latest version of the sources above consists of much newer and richer knowledge than those in XLORE.

2). Cross-lingual Knowledge Graph Building: Second, we integrate Chinese knowledge via a Chinese Wikipedia category system. Given existing cross-lingual links, English Wikipedia knowledge and

integrated Chinese knowledge, we perform cross-lingual knowledge linking, cross-lingual property matching and cross-lingual taxonomy alignment to build a cross-lingual knowledge graph. Based on existing cross-lingual links, we aim to find more cross-lingual links between instances, properties and concepts in different languages. This strategy can largely enrich the cross-lingual knowledge and facilitate knowledge sharing across different languages.

3). Data Quality Improvement: Third, we utilize two methods to improve the data quality of XLOR2 by performing cross-lingual knowledge validation. The goal here is to predict whether the subClassOf relation between two concepts is correct and whether the instanceOf relation between an instance and a concept is correct. In addition, to make the best of unintegrated instances, we utilize a fine-grained type inference method to find their missing types.

4). Application: Finally, we construct an online system XLORE2, along with a bilingual entity linking application XLink which uses XLORE2 as the primary datasource.

In the following sections, we will describe each part of this system in detail.

4. CROSS-LINGUAL KNOWLEDGE GRAPH BUILDING

Existing cross-lingual links between English and Chinese is limited. As such we perform cross-lingual knowledge linking, cross-lingual property matching and cross-lingual taxonomy alignment to find more cross-lingual links. This not only globalizes the knowledge sharing of different languages on the Web, but also serves to benefit many online applications by facilitating information retrieval and machine translation.

4.1 Cross-lingual Knowledge Linking

Problem: XLORE2 contains 4.7 million English instances and 10 million Chinese instances. There are currently only 424,000 cross-lingual links between instances of the two languages. One important task required to expand knowledge linking is to discover new links between instances in different languages that are "equivalent" and describe the same thing. If two instances semantically describe the same subject or topic then we can say that they are equivalent. Automatically discovering cross-lingual links between instances in different languages can largely enrich the cross-lingual knowledge and facilitate knowledge sharing across different languages. Figure 2 shows an example for cross-lingual knowledge linking. The instance "*Anaerobic exercise*" is in English and the other instance "*无氧运动*" is in Chinese. There is not a cross-lingual link between them. However, in cross-lingual knowledge linking, our goal is to find an equivalent instance in Chinese for the English instance "*Anaerobic exercise*" in XLORE2. In order to find the equivalent relations between instances, it is helpful to consider different kinds of information. Figure 2 highlights some useful information in the two instances including textual, linkage and semantic information.

Existing methods face the following non-trivial challenges:

1). Feature Expansibility: Existing methods usually rely on many well-defined lexical or structural features (similarities) to predict new cross-lingual links. Though this requires rich background

knowledge and extensive human labor in the feature design and extraction [5, 6]. These types of features largely rely on the Wikipedia internal structure, which in truth is not so expansive. Can we utilize distributed representation learning methods to jointly embed entities in a different language to the same space?

2). Link Sparsity: Existing methods for finding cross-lingual links heavily depend on the number of existing cross-lingual links [7]. Such a method often results in high precision, but low recall. To address the problem of cross-lingual link sparsity, can we incorporate both textual and structural information in representation learning?

Main Idea: Regarding the discovery of semantic equivalence relations between instances from different languages, we propose a method based on representation learning of heterogeneous networks to address the sparseness and poor scalability of traditional similarity features caused by cross-language gaps. Our method represents the cross-lingual instances in a consistent low-dimensional vector space, where we know that the deep representation of textual, linkage and semantic information successfully improves the performance of semantic equivalence relation discovery and in particular the performance in recall.

Figure 2. An example of cross-lingual knowledge linking.

Method: We denote our method as **H**eterogenous **N**etwork **E**mbedding (HNE). The framework of this model is shown in Figure 3. Our method consists of three components:

1). Heterogenous Network Construction: We construct the textual network between instances with designated words for each language, a linkage network joining instances for each language, a semantic network between instances and words for each language as well as an inter-wiki network (existing cross-lingual links connecting the languages pairs).
2). Network Representation Learning: We apply a state-of-the-art embedded network discovery method to learn Chinese and English instance embeddings [8].
3). Cross-lingual Link Discovery: We utilize a logistic regression model to find new cross-linguals between Chinese and English instances.

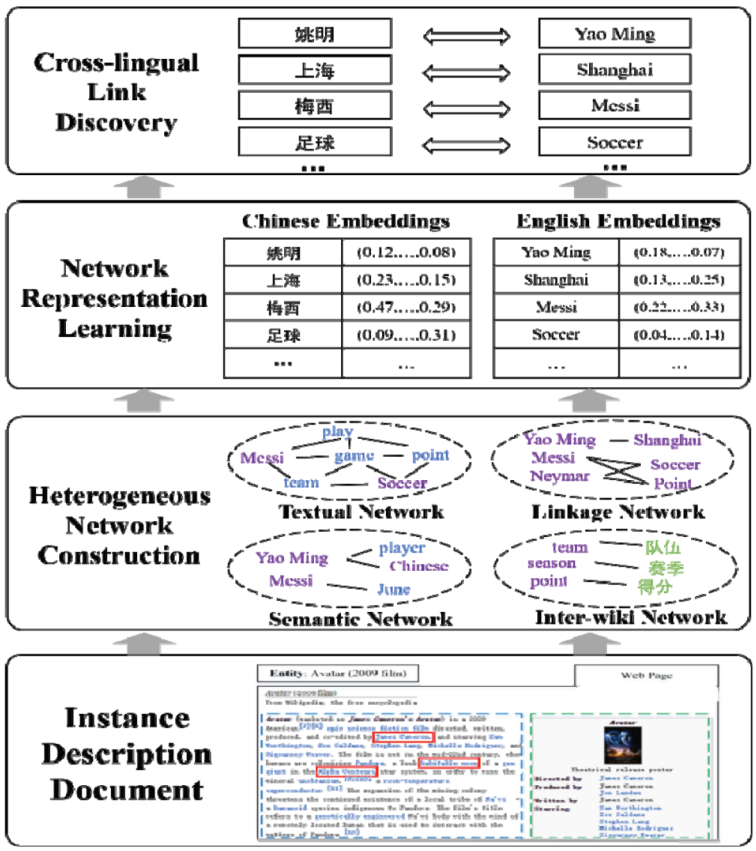

Figure 3. The framework of cross-lingual knowledge linking.

4.2 Cross-lingual Property Matching

Problem: There is a large amount of semantic information contained in Wikipedia infoboxes, which provides semi-structured, factual information in the form of attribute-value pairs. Attributes in infoboxes contain valuable semantic information, which plays a key role in the construction of a coherent large-scale knowledge base [2]. However, each language version maintains its own set of infoboxes with their own set of attributes, as well as sometimes providing different values for corresponding attributes (as shown in Figure 4). Thus, attributes in different Wikipedia must be matched if we want to get coherent knowledge. For instance, inconsistencies among the data provided by different editions for corresponding attributes could be detected automatically. English Wikipedia is obviously larger and of higher quality than low resource languages. This enables us to use attribute alignments to expand and complete infoboxes in other languages, or at least help Wikipedia communities to do so. This is encouraging because the number of existing attribute mappings is limited, e.g., there are more than 100,000 attributes in English Wikipedia but only about 5,000 (less than 5%) existing attribute mappings between English and Chinese.

There are several challenges involved in finding multilingual correspondences across infobox attributes. First, there are *Polysemy-Attributes* (a given attribute can have different semantics, e.g., country can mean nationality of one person or place of a production) and *Synonym-Attributes* (different attributes can have the same meaning, e.g., alias and other names), which leads to worse performance on label similarity or translation based methods. Second, there exist some problems in the values of attributes: 1) *different measurement* (e.g., height of Yao Ming is 2.29m in English edition and 7 feet 6 inch in Chinese) and 2) *timeliness* (e.g., population of Beijing is 21,150,000 (in 2013) in French edition). In this way, labels and values alone are not credible enough for cross-lingual attribute matching.

Figure 4. An example of cross-lingual property matching (attribute matching).

Main Idea: In order to solve the problems above, we must first investigate several effective features considering the characteristics of cross-lingual attribute matching. Then we need to propose an approach based on the factor graph model [9, 10]. The most significant advantage of this model is that it can formalize correlations by joining attributes explicitly. The method is as follows.

Method: Figure 5 contains two parts. Part 1 on the left is a relation graph which represents several relations in two editions of Wikipedia K_1 and K_2. Different language versions are separated by a diagonal line. The attribute layer contains the attributes and template relations among them. Similarly, the article layer contains the articles and category relations. The imaginary lines between the two layers denote the relation of usage between articles and attributes. The red dashed lines denote the existing cross-lingual links. Part 2 on the right is a factor graph. The white nodes are variables and there are two types of variables, x_i and y_i. Each candidate pair is mapped to an observed variable x_i. The hidden variable y_i represents a Boolean label (equivalent or inequivalent) of the observed variable x_i. For example, x_2 in Figure 5 corresponds to a candidate attribute pairs (p_{i3}, p_{j2}), and there exists a cross-lingual link between p_{i3} and p_{j2}. So the hidden variable y_2 equals to 1. The black nodes in the factor graph are factors. There are three types, f, g and h. Each type is associated with a kind of feature function which transforms relations into a computable feature.

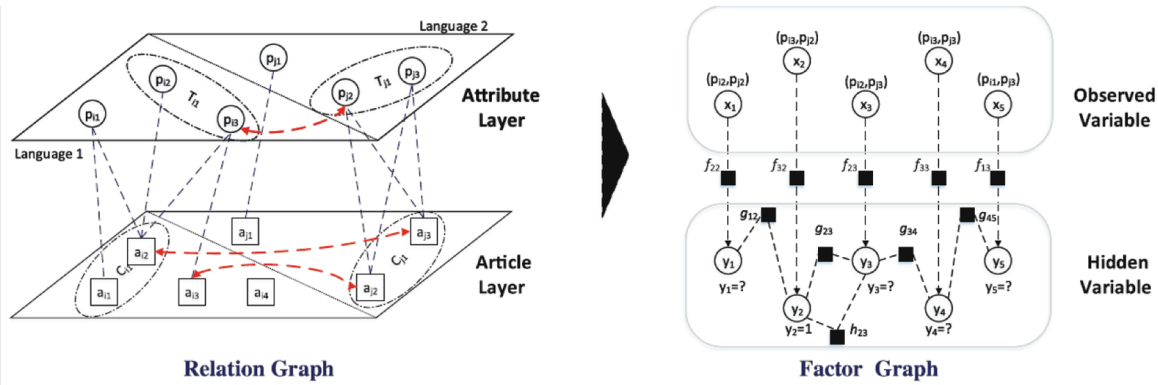

Figure 5. The framework of cross-lingual property matching.

4.3 Cross-lingual Taxonomy Alignment

Problem: Cross-lingual taxonomy alignment aims to map each concept in the source taxonomy of one language onto a ranked list of most relevant concepts in the target taxonomy of another language [11,12,13]. Recently, vector similarities that depend on bilingual topic models have achieved a state-of-the-art performance relative to this task. However, these models only consider the textual context of concepts while outright ignoring explicit concept correlations such as those between the concepts and their co-occurring words in text or those among the concepts of ancestor-descendant relationships in a taxonomy.

Method: Cross-lingual taxonomy alignment in general is non-trivial. Fortunately, our goal is fairly simple, that is to link concepts across different languages. The purpose is merely to construct the cross-lingual knowledge graph, not to find as many cross-lingual concept links as possible. Thus, we are only concerned with the precision and not the recall. In XLORE2 we directly utilize the cross-lingual links between categories provided by Wikipedia as cross-lingual links between concepts. It is a relatively simple method which permits us to investigate a more powerful method to find more cross-lingual links between concepts in different languages as we look toward the future.

5. DATA QUALITY IMPROVEMENT

Because XLORE2 is a large-scale cross-lingual knowledge graph, it naturally contains many mistakes and errors. This is unavoidable, partly because Wikipedia is user-generated (Wikipedia is a source of XLORE2), and also because the world is always changing and evolving. Knowledge bases are therefore always in need of maintenance, updates and corrections. So it is necessary to continually improve the quality of the data in XLORE2. We perform cross-lingual knowledge validation to correct wrong *subClassOf* relations between two concepts and wrong *instanceOf* relations between an instance and a concept. Subsequently we propose to utilize a fine-grained type inference method to find more *instanceOf* relations between instances and concepts.

5.1 Cross-lingual Knowledge Validation

Problem: As mentioned above, the taxonomy in XLORE2 is derived from the Wikipedia category system. The taxonomy directly derived from Wikipedia usually contains many mistakenly imported *subClassOf* and *instanceOf* relations. By treating each category and disambiguated article as one candidate class and instance, respectively, the taxonomies are therefore directly derived from the online wikis by way of transforming the user-generated subsumption relations, namely *subCategoryOf* between two categories and *articleOf* from one article to one category, into the semantic taxonomic relations, which are subClassOf between two classes and instanceOf from one instance to one class. Unfortunately the user-generated subsumption relations in Wikipedia and the semantic taxonomic relations in the knowledge bases are not exactly the same. The well-defined subClassOf and instanceOf essentially represent the *isA* relation, while freely edited subCategory and *articleOf* cover another *topicOf* relation which denotes the topic related relation and generates the noise in the derived taxonomy. As Figure 6 clearly shows, when reasoning is based solely on the derived taxonomy, the system mistakenly concludes that *Barack Obama* (person) *isA Chicago, Illinois* (location) which apparently should be the topic related relation. So in fact it is a really auspicious problem to have, to be presented with the opportunity to correct those wrong subClassOf and instanceOf relations.

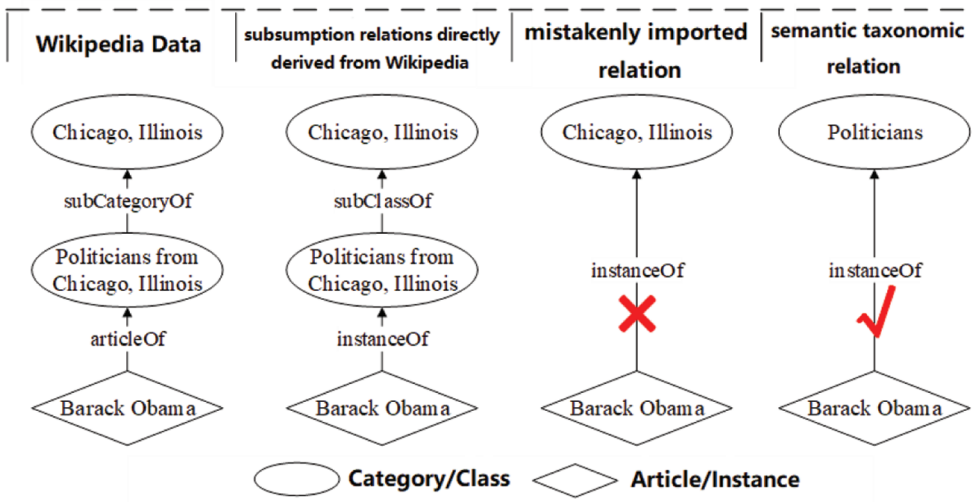

Figure 6. An example of mistakenly derived facts.

Existing approaches typically suffer from the following problems:

1). Language Dependence: The heuristic-based methods that strongly rely on the accuracy of the headword recognition algorithm and the language dependent rules, such as those involving Chinese and Japanese, are much too rigid to handle some languages with no explicit plural/singular forms [14, 15].
2). Limited Corpus: The corpus-based methods depend on large-scale corpora with a high degree of quality, which in fact are often simply unavailable [15]. Thus, the generated taxonomies are often small, mostly domain dependent, and thus have a rather poor performance [16].

Main Idea: We formulate the above problem as the problem of cross-lingual taxonomic relation prediction [17]. We investigate different linguistic heuristics and language independent features, and propose a cross-lingual knowledge validation based dynamic adaptive boosting model to iteratively reinforce the performance of taxonomic relation prediction. Specifically we tackle taxonomic relation prediction as a binary classification problem. We do this by learning the following two functions: the *subClassOf Prediction Function* and the *instanceOf Prediction Function*.

Method: The framework of our model is shown in Figure 7. First, we utilize the binary classifier for the basic learner and use the Decision Tree as our implementation. We analyze some features which are beneficial to taxonomic relation prediction. The defined features include the linguistic heuristic features and the language-independent structural features. Then we propose the Dynamic Adaptive Boosting (DAB) model for cross-lingual taxonomy derivation. To improve the learning performance of taxonomic relation prediction, our model is trained iteratively on an active dynamic training set. The training examples are

weighted samples from the pre-labeled data and the cross-lingual validated predicted data. We utilize a cross-lingual validation method to avoid potential performance degradation.

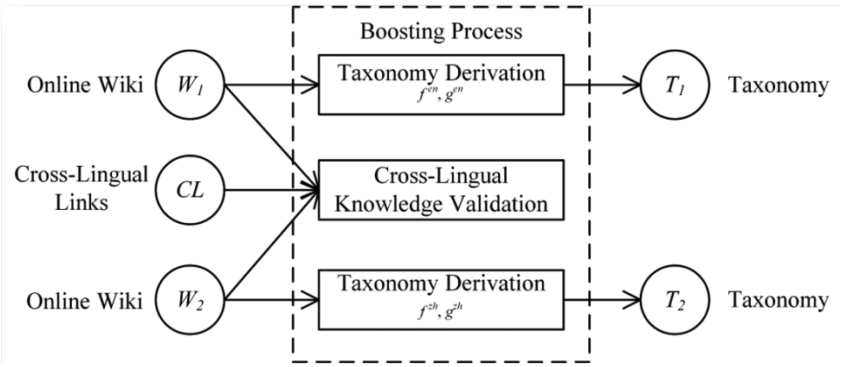

Figure 7. The framework of cross-lingual knowledge validation.

Evaluation: To evaluate our approach, we conduct our experiments using English Wikipedia and Chinese Hudong Baike. We retrieve the English Wikipedia data set containing 561,819 categories and 3,711,928 articles. The Chinese Hudong Baike data set contains 28,933 categories and 980,411 articles. Our approach significantly outperforms the well designed state-of-the-art comparison methods, with 0.3%, 1.3%, 1.7% and 19.4% improvement in F1 pertaining to the English SubClassOf, Chinese SubClassOf, English InstanceOf and Chinese InstanceOf validation tasks, respectively.

5.2 Fine-grained Type Inference

Problem: Type information is very important relative to knowledge bases. But some large knowledge bases lack meaningful type information due to the incompleteness of the knowledge bases themselves. In XLORE more than 18.7% instances are without useful type information, so specifically our target is to identify the semantic type of an instance in XLORE2. This is called *instance type inference*. Traditional type inference methods focus on a small set of types such as *Person*, *Location* and *Organization*. Fine-grained type inference assigns more specific types to an instance, which will normally result in forming new type-paths in the taxonomy [18, 19, 20, 21, 22]. As shown in Figure 8, *Yao Ming* is associated with a type-path */Thing/Agent/Person/Athlete/Basketball Player*. Fine-grained types (e.g., *Athlete* and *Basketball Player*) are more informative than coarse-grained types (e.g., *Person*) because they provide more specific semantic information about an instance. Characterizing an instance with fine-grained types (type-paths) benefits many real world applications such as knowledge base completion, entity linking, relation extraction and question answering.

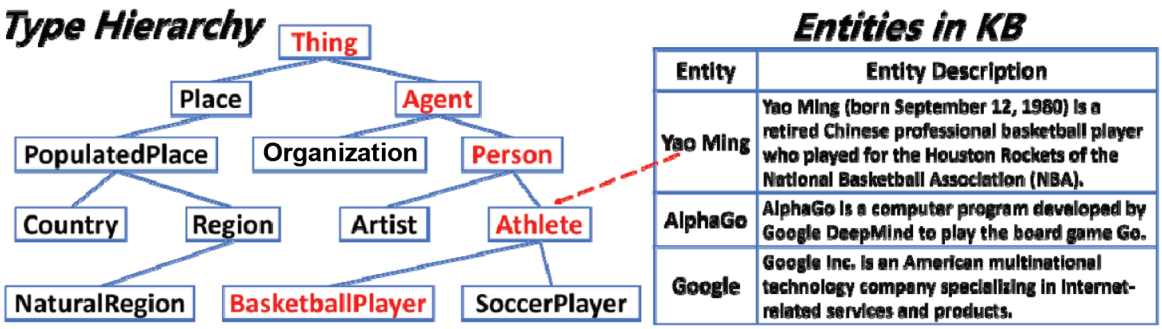

Figure 8. An example of fine-grained type inference. Note: The red items are assigned types.

Existing approaches suffer the following problems:

1). Hand-crafted features: Sentence-level methods focus on classifying an instance mention in the text to a broad set of types so as to exploit the well-defined linguistic features based on the mention itself and the contextual information in text. These require rich background knowledge [18, 19, 20, 21, 22].
2). Annotated corpus: Corpus-level methods utilize annotated corpora to learn low-dimensional representations of the instances then subsequently determine type inference based on the learned embeddings. But such large-scale high-quality annotated corpora are usually quite difficult to obtain [23, 24, 25] whereas access to simple entity text descriptions is often relatively easy to achieve.

Main Idea: To address the above issues we propose an embedding based method [26]. Our model makes type inference based on instance text description. Not all instances are labeled with types. We construct heterogeneous networks which encode different levels of co-occurrence information and labeled information. Then we learn instance, word and type representations jointly via a network embedding method.

Method: The framework of our model is shown in Figure 9. First, we construct four heterogeneous networks to exploit different kinds of information effectively, i.e., word-word, instance-word, instance-type and type-word. Each network encodes a specific kind of semantic information. Then we utilize a heterogeneous network embedding method to learn low-dimensional representations for each instance and type based on the above networks. The benefit from the heterogeneous networks, the learned instance and type embeddings is not only to preserve their semantic closeness, but even more so, to facilitate a higher quality predictive result for the type inference task. To meet the fine-grained demand we use learning to rank algorithms, greatly improving the quality of instance and type embeddings. Finally, we use the learned embeddings to force type inference for each unlabeled instance in XLORE2.

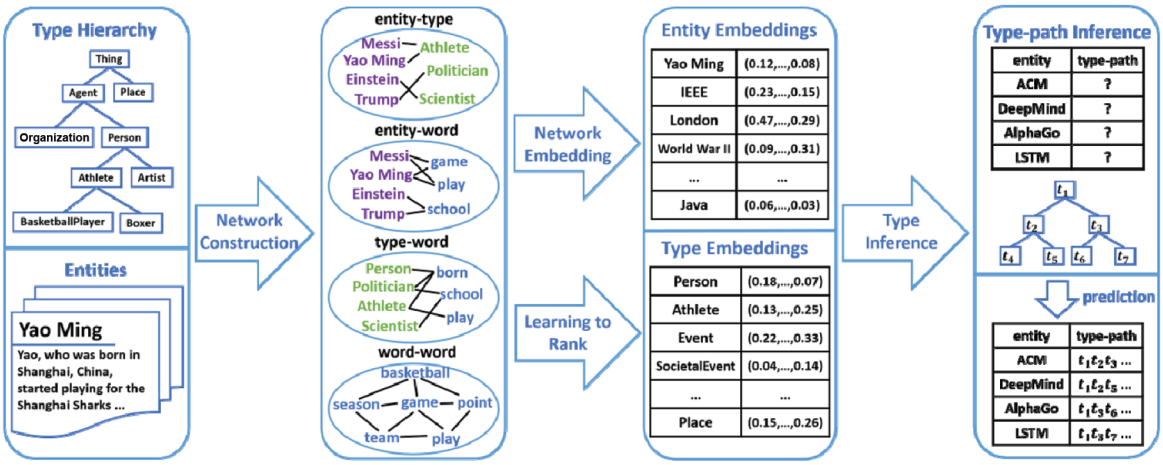

Figure 9. The framework of fine-grained type inference.

Evaluation: We evaluate the proposed method using real-world data sets collected from Wikipedia and DBpedia (entity from Wikipedia and type hierarchy from Dbpedia). Our proposed method outperforms state-of-the-art methods with 2.8% and 4.2%, respectively regarding improvement in Mi-F1 and Ma-F1 specific to entity typing tasks.

6. APPLICATION

Entity Linking (EL) is a fundamental Natural Language Processing and Knowledge Engineering technology. EL builds bridges between plain text and the knowledge base. *XLink* is the entity linking system application of XLORE2. An entity link is the task of linking mentions in the text to the corresponding entities in a knowledge base. Recently entity linking has received considerable attention and several online entity linking systems have been published such as Wikify! [27], AIDA [28], DBpedia Spotlight [29], TagMe [30] and Linkify [31].

Problem: Existing entity linking systems commonly have two components: *mention detection* and *entity linking*. For mention detection, AIDA [28] and Linkify [31] depend on Names Entity Recognition (NER) tools. However, NER tools rely heavily on language and these tools only recognize three types of named entities: *Person*, *Location* and *Organization*. This leaves a significant gap in the types of entities covered in knowledge base [32]. To address the problem of ambiguity and variation in entity linking, the simplest way is to choose the most prominent entity (i.e., the candidate with the largest number of incoming or outgoing links in Wikipedia) for the given mention. However, the different context of mentions leads to different linking results, which in turn becomes too complex to be solved through a mapping of entity priority. An alternative strategy is to calculate the contextual similarity for single mention linking, and to further employ the topical coherence to collectively link all mentions within a document. But unfortunately few of these systems consider the features in a unified and effective manner. Moreover, these systems mainly use

Wikipedia as the knowledge base and rarely handle Chinese documents. Additionally, there is the issue of many large-scale Chinese encyclopedias e.g., Baidu Baike) emerging and evolving. So really it is now time to begin conducting entity linking in both Chinese and English.

Method: To address the above issues, we develop a bilingual online entity linking system named *XLink*. As shown in Figure 10, it conducts a language-independent process for both Chinese and English documents, on-the-fly, via two phases: *mention parsing* and *entity disambiguation*. Mention parsing detects mentions in input documents and generates candidate entities for each mention. The entity disambiguation phase chooses the correct entity in the candidate set. Xlink's purpose is to provide users an online service for linking all important mentions in the text to entities in the knowledge base, both correctly and efficiently. In particular, we first use a parsing algorithm to search a pre-built dictionary to detect mentions instead of using NER tagger. Second, we design a generative probabilistic entity disambiguation method which models contextual feature, coherence feature and prior feature jointly, to guarantee the accuracy of disambiguation. For the system efficiency we use the Aho-Corasick algorithm to parse mentions and introduce word and entity embeddings. This ensures the time efficiency of the disambiguation phase. Finally, the disambiguation method is unsupervised to facilitate easy deployment of the system online.

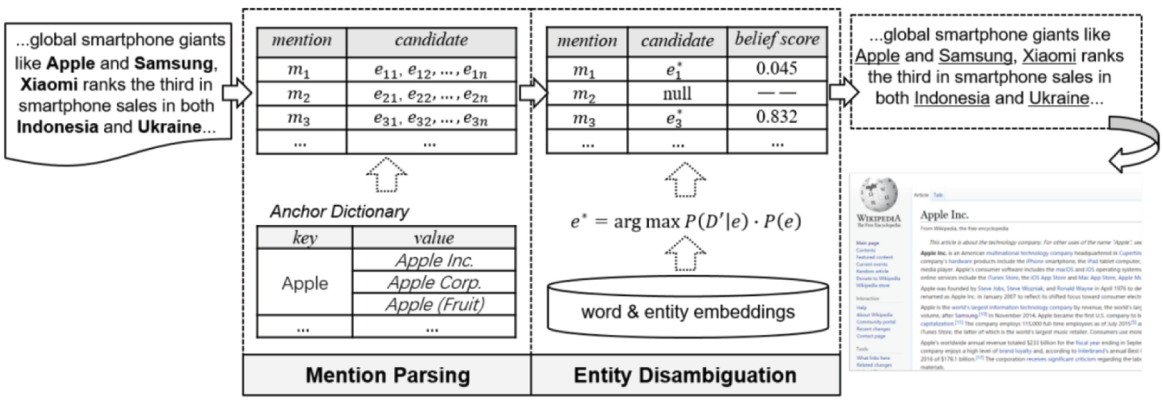

Figure 10. Illustration of XLink.

7. SYSTEM AND DATA STATISTICS

We construct XLORE2 in the form of RDF and use the OpenLink Virtuoso server for systematical data management. Using the proposed approach, XLORE2 harvests 1,371,272 concepts, 512,883 properties and 14,951,135 instances across English and Chinese. Table 1 gives a brief summary analysis of the linked item enhancement and the percentage increase when compared to XLORE.

Table 1. XLORE2 statistics.

	English	Chinese	Total	Linked
Concepts	1,214,820	228,635	1,371,272(+60.17%)	72,183(+51.69%)
Instances	4,769,900	10,605,209	14,951,135(+90.36%)	423,974(+79.21%)
Properties	42,110	460,241	512,883(+616.36%)	10,532(NULL)

We update our online system to illustrate XLORE2. As shown in Figure 11, one new application feature of XLORE2 is that the system supports the keyword-based or SPARQL queries. We also introduce several APIs for readers to access and download instances, concepts and properties in XLORE2. This greatly helps to facilitate relevant research. Another application addition is our new entity linking system XLink, as shown in the right part of Figure 11. XLink is an unsupervised bilingual entity linking system. It conducts mention parsing and entity disambiguation to link the mentions in the input document to entities in XLORE2. We invite the readers to access our XLORE2 system at http://XLORE.org and XLink system at http://xlink.xlore.org/.

8. CONCLUSION

In this paper, we present XLORE2, an extension of XLORE, to adequately solve the problem of limited cross-lingual links and wrong/missing instanceOf/subClassOf relations.

We infer missing facts based on existing ones in XLORE via three methods:

1). We propose to utilize the heterogeneous network embeddings method and regression-based model to predict new cross-lingual links.
2). We investigate several effective features and propose the entity-attribute factor graph to find corresponding attributes between English and Chinese.
3). We propose to utilize the heterogeneous network embedding method to find missing instanceOf relations between instances and concepts automatically.

Finally, XLORE2 realizes significantly more facts when compared with XLORE. So we design an efficient entity linking system XLink, which can link the mentions in a document to entities in XLORE2.

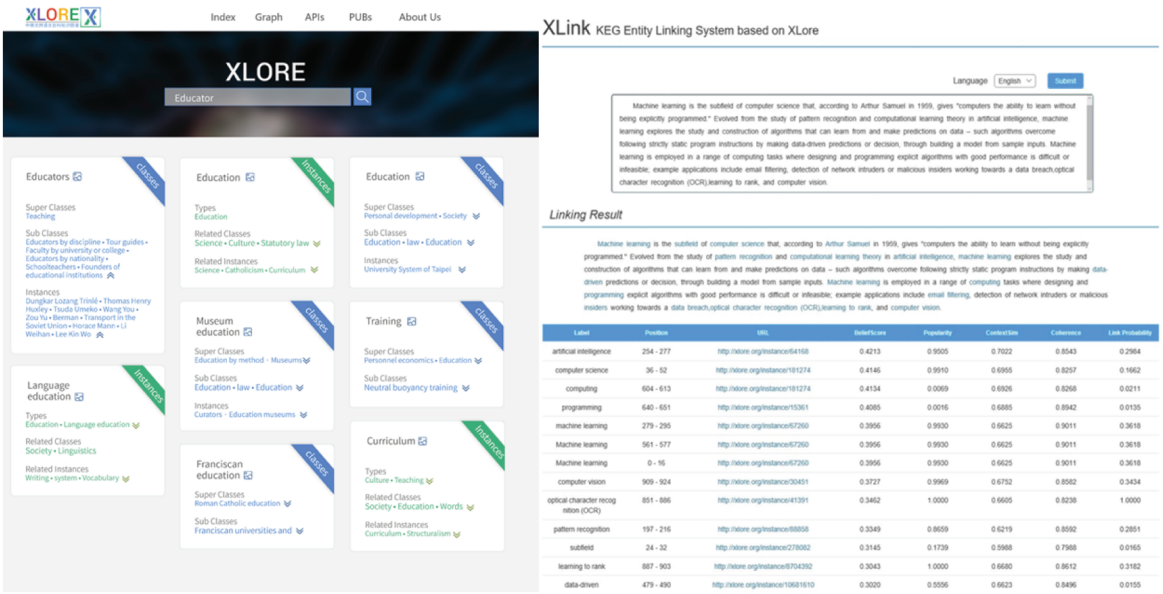

Figure 11. The interface of XLORE2 and XLink.

AUTHOR CONTRIBUTIONS

All of the authors contributed equally to the work. J. Li (lijuanzi@tsinghua.edu.cn) is the leader of the XLORE system, who drew the whole framework of the system. H. Jin (jinhl15@mails.tsinghua.edu.cn) and C. Li (licj17@mails.tsinghua.edu.cn) summarized the methodology part of this paper. J. Zhang (jing-zha15@mails.tsinghua.edu.cn) and L. Hou (houlei@tsinghua.edu.cn, corresponding author) summarized the applications and drafted the paper. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

ACKNOWLEDGEMENTS

The work is supported by National Natural Science Foundation of China (NSFC) key project (No. 61533018, No. U1736204 and No. 61661146007), Ministry of Education and China Mobile Research Fund (No. 20181770250) and THUNUS NExT Co-Lab.

REFERENCES

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, & C. Bizer. DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2)(2015), 167-195. doi: 10.3233/SW-140134.

[2] F. Mahdisoltani, J. Biega, & F.M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In: The 7th Biennial Conference on Innovative Data Systems Research (CIDR 2015), California, USA, 2015. Available at: http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf.

[3] R. Navigli, & S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193(2012), 217-250. doi: 10.1016/j.artint.2012.07.001.

[4] Z. Wang, J. Li, Z. Wang, S. Li, M. Li, D. Zhang, Y. Shi, Y. Liu, & J. Tang. XLore: A large-scale English-Chinese bilingual knowledge graph. In: The 12th International Semantic Web Conference (ISWC2013) on Posters & Demonstrations Track. Available at: https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ab/2013/iswc2013-Nov13/iswc2013.semanticweb.org/content/demos/31.html.

[5] P. Sorg, & P. Cimiano. Enriching the crosslingual link structure of Wikipedia—a classification-based approach. In: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, 2008. Available at: http://www.aifb.kit.edu/images/3/3b/2008_1758_Sorg_Enriching_the_c_1.pdf.

[6] J.H. Oh, D. Kawahara, K. Uchimoto, J. Kazama, & K. Torisawa. Enriching multilingual language resources by discovering missing cross-language links in Wikipedia. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2008, pp. 322-328. doi: 10.1109/WIIAT.2008.317.

[7] Z. Wang, J. Li, Z. Wang, & J. Tang. Cross-lingual knowledge linking across Wiki knowledge bases. In: Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 459-468. doi: 10.1145/2187836.2187899.

[8] J. Tang, M. Qu, & Q. Mei. PTE: Predictive text embedding through large-scale heterogeneous text networks. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1165-1174. doi: 10.1145/2783258.2783307.

[9] F.R. Kschischang, B.J. Frey, & H.A. Loeliger. Factor graphs and the sum-product algorithm. IEEE Transactions on information theory 47(2)(2001), 498-519. doi: 10.1109/18.910572.

[10] Y. Zhang, T. Paradis, L. Hou, J. Li, J. Zhang, & H. Zheng. Cross-lingual infobox alignment in Wikipedia using entity-attribute factor graph. In: C. D'Amato, M. Fernández, V. Tamma et al. (eds.) The Semantic Web—ISWC 2017. Cham: Springer, 2017, pp. 745-760. Available at: https://www.springer.com/cn/book/9783319682877.

[11] T. Wu, L. Zhang, G. Qi, X. Cui, & K. Xu. Encoding category correlations into bilingual topic modeling for cross-lingual taxonomy alignment. In C. D'Amato, M. Fernández, & V. Tamma et al. (eds.) The Semantic Web—ISWC 2017. Cham: Springer, 2017, pp. 728-744. Available at: https://www.springer.com/cn/book/9783319682877.

[12] T. Wu, G. Qi, & H. Wang, K. Xu, & X. Cui. Cross-lingual taxonomy alignment with bilingual biterm topic model. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), ACM, 2016, pp. 287-293. Available at: https://aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12011.

[13] T. Wu, D. Zhang, L. Zhang, & G. Qi. Cross-lingual taxonomy alignment with bilingual knowledge graph embeddings. In: Z. Wang, A.Y. Turhan, & K. Wang, et al. (eds.) Joint International Semantic Technology Conference (JIST) 2017: Semantic Technology. Cham: Springer, pp. 251-258. doi: 10.1007/978-3-319-70682-5_16.

[14] G. de Melo, & G. Weikum. MENTA: Inducing multilingual taxonomies from Wikipedia. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, 2010, pp. 1099-1108. doi: 10.1145/1871437.1871577.

[15] S.P. Ponzetto, & M. Strube. Deriving a large scale taxonomy from Wikipedia. In: Proceedings of the 22nd AAAI Conference on Artificial intelligence (AAAI'07), AAAI, 2007, pp. 1440-1445. Available at: http://www.aaai.org/Papers/AAAI/2007/AAAI07-228.pdf.

[16] C. Brewster. Ontology learning from text: Methods, evaluation and applications. In: P. Buitelaar, P. Cimiano, & B. Magnini (eds.) Computational Linguistics. Cambridge, MA: MIT Press, 2006, pp. 569–572. doi: 10.1162/coli.2006.32.4.569.

[17] Z. Wang, J. Li, S. Li, M. Li, J. Tang, K. Zhang, & K. Zhang. Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online Wikis. In: Proceedings of the 28th AAAI Conference on Artificial

Intelligence (AAAI'14), AAAI, 2014, pp. 180-186. Available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/download/8260/8418.

[18] X. Ren, W. He, M. Qu, L. Huang, H. Ji, & J. Han. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 1369-1378. Available at: http://www.aclweb.org/anthology/D/D16/D16-1144.pdf.

[19] X. Ren, W. He, M. Qu, C.R. Voss, H. Ji, & J. Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1825-1834. doi: 10.1145/2939672.2939822.

[20] X. Ling, & D.S. Weld. Fine-grained entity recognition. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12), AAAI, 2012. Available at: https://aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5152.

[21] D. Yogatama, D. Gillick, & N. Lazic. Embedding methods for fine grained entity type classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, 2015, pp. 291-296. Available at: http://www.anthology.aclweb.org/P/P15/P15-2048.pdf.

[22] M.A. Yosef, S. Bauer, J. Hoffart, M. Spanio, & G. Weikum. Hyena: Hierarchical type classification for entity names. In: Proceedings of COLING 2012, 2012, pp. 1361-1370. Available at: http://www.anthology.aclweb.org/C/C12/C12-2133.pdf.

[23] Y. Yaghoobzadeh, & H. Schütze. Corpus-level fine-grained entity typing using contextual information. arXiv preprint. arXiv: 1606.07901, 2016.

[24] Y. Yaghoobzadeh, & H. Schütze. Multi-level representations for fine-grained typing of knowledge base entities. arXiv preprint. arXiv: 1701.02025, 2017.

[25] Y. Yaghoobzadeh, H. Adel, & H. Schuetze. Corpus-level fine-grained entity typing. Journal of Artificial Intelligence Research 61(2018), 835-862. doi: 10.1613/jair.5601.

[26] H. Jin, L. Hou, & J. Li. Type hierarchy enhanced heterogeneous network embedding for fine-grained entity typing in knowledge bases. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer, 2018.

[27] R. Mihalcea, & A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, ACM, 2007, pp. 233-242. doi: 10.1145/1321440.1321475.

[28] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, & G. Weikum. Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 782-792.

[29] P.N. Mendes, M. Jakob, A. García-Silva, & C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, 1-8. doi: 10.1145/2063518.2063519.

[30] P. Ferragina, & U. Scaiella. Fast and accurate annotation of short texts with Wikipedia pages. IEEE Software 29(1)(2012), 70-75. doi: 10.1109/MS.2011.122.

[31] I. Yamada, T. Ito, S. Usami, S. Takagi, H. Takeda, & Y. Takefuji. Evaluating the helpfulness of linked entities to readers. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, ACM, 2014, pp. 169-178. doi: 10.1145/2631775.2631802.

[32] J.R. Finkel, T. Grenager, & C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 363-370. Available at: https://dl.acm.org/citation.cfm?id=1219885.

AUTHOR BIOGRAPHY

Hailong Jin is currently a PhD student in the Department of Computer Science and Technology, Tsinghua University. He received his Bachelor Degree from Harbin Institute of Technology in 2015. His research interests include semantic Web, knowledge graph and entity typing.

Chengjiang Li is currently a master student in the Department of Computer Science and Technology, Tsinghua University. He received his Bachelor Degree from Tsinghua University in 2017. His research interests include semantic Web, knowledge graph and cross-lingual entity alignment.

Jing Zhang is currently a master student in the Department of Computer Science and Technology, Tsinghua University. She received her Bachelor Degree from Beijing University of Posts and Telecommunications in 2015. Her research interests include knowledge graph, entity linking and named entity recognition.

Lei Hou is currently a postdoctoral researcher in the Department of Computer Science and Technology, Tsinghua University. He received his PhD Degree from Tsinghua University in 2016 under the supervision of Prof. Juanzi Li. He received his Bachelor Degree from Beijing University of Posts and Telecommunications in 2010. His research interests include semantic Web, and news and user-generated content mining.

Juanzi Li received her PhD Degree in Computer Science from Tsinghua University in 2000. She is now working as a professor in the Department of Computer Science and Technology, Tsinghua University. Her research interests include semantic Web, knowledge discovery, social network analysis, news mining and natural language processing. She has published over 100 research papers in major international journals and conferences.

Peng Zhang is currently a PhD student in the Department of Computer Science and Technology, Tsinghua University. He received his Master's Degree from Tsinghua University in 2005, and Bachelor Degree from Tsinghua University in 2002. He is the system designer of XLORE.