XLORE2: Large-Scale Cross-Lingual Knowledge Graph Construction and Application Postprint

Authors: Jin, Hailong, Li, Chengjiang, Zhang, Jing, Hou, Lei, Li, Juanzi, Zhang, Peng, Hou, Lei

Date: 2022-11-25T00:00:00+00:00

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.

Full Text

Preamble

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 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, with extremely high language coverage containing articles in 298 languages. Among them, the English version 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 infoboxes. 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 same knowledge in different languages. DBpedia does much work on the extraction and alignment of cross-lingual fact triples. BabelNet concentrates on entity concepts, senses and synsets.

The imbalanced size of different Wikipedia language versions apparently leads to highly imbalanced knowledge distribution across languages. This is reflected in KBs based on this imbalance, as knowledge encoded in non-English languages is much less than that 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 provides 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 WORK

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

DBpedia . Using Semantic Web and Linked Data technologies, DBpedia is a crowd-sourced community effort to extract structured, multilingual knowledge from 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 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 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 . YAGO is an extensible semantic KB with high coverage and enhanced quality, derived from Wikipedia, WordNet and GeoNames. YAGO uses categories in Wikipedia to infer type information about an entity and then links this type information to WordNet to pursue knowledge coherence. It contains more than 1 million entities and 5 million facts. In YAGO2, some declarative extraction rules were introduced to gather and integrate temporal, spatial and semantic information from resources. This space- and time-awareness results in 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 with help from Wikidata. It can achieve a precision of 95%-100% in attribute mapping and thus gains roughly 1 million new entities and 7 million new facts over the original English-only YAGO. Generally, one main difference between YAGO and XLORE2 is that XLORE2 applies cross-lingual entity alignment and cross-lingual attribute alignment methods to merge equivalent entities and attributes automatically. YAGO does not align different extractions from different Wikipedias, but rather aligns different Wikipedias with a central clean KB.

Wikidata . 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, 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 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 project launch.

BabelNet . BabelNet is a very large wide-coverage multilingual semantic network built from Wikipedia and WordNet, quite similar to the YAGO project. The key methodology is integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia via automatic mapping. Inevitably, there are lexical gaps in resource-poor languages. BabelNet fills these lexical gaps with aid from machine translation. The mapping process result can be defined as an encyclopedic dictionary. This approach provides concepts and instances lexicalized in different languages and connected with many semantic relations. BabelNet has now reached version 4.0, covering 284 languages with 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. Knowledge encoded in BabelNet can be used to perform knowledge-rich, graph-based word sense disambiguation in both monolingual and multilingual settings.

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

3. METHODOLOGY

Following the framework shown in Figure 1 [Figure 1: see original paper], XLORE2 construction contains four stages:

Data Preprocessing. First, we gather and clean data from four online wikis: 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 these sources consists of much newer and richer knowledge than those in XLORE.

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 cross-lingual knowledge and facilitate knowledge sharing across languages.

Data Quality Improvement. Third, we utilize two methods to improve XLORE2 data quality by performing cross-lingual knowledge validation. The goal 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 best use of unintegrated instances, we utilize a fine-grained type inference method to find their missing types.

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 describe each part of this system in detail.

4. CROSS-LINGUAL KNOWLEDGE GRAPH BUILDING

Existing cross-lingual links between English and Chinese are 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 knowledge sharing of different languages on the Web, but also benefits 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 discovering 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, we can say they are equivalent. Automatically discovering cross-lingual links between instances in different languages can largely enrich cross-lingual knowledge and facilitate knowledge sharing across languages. Figure 2 [Figure 2: see original paper] 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. To find 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, requiring rich background knowledge and extensive human labor in feature design and extraction [5, 6]. These features largely rely on Wikipedia' s internal structure, which is not so expansive. Can we utilize distributed representation learning methods to jointly embed entities in different languages into 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 methods often result in high precision but low recall. To address cross-lingual link sparsity, can we incorporate both textual and structural information in representation learning?

Main Idea. Regarding discovery of semantic equivalence relations between instances from different languages, we propose a method based on representation learning of heterogeneous networks to address sparseness and poor scalability of

traditional similarity features caused by cross-language gaps. Our method represents cross-lingual instances in a consistent low-dimensional vector space, where deep representation of textual, linkage and semantic information successfully improves performance of semantic equivalence relation discovery, particularly recall performance.

Method. We denote our method as Heterogenous Network Embedding (HNE). The framework is shown in Figure 3 [Figure 3: see original paper]. 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, and an inter-wiki network (existing cross-lingual links connecting language 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-lingual links between Chinese and English instances.

4.2 Cross-Lingual Property Matching

Problem. Wikipedia infoboxes contain large amounts of semantic information, providing semi-structured factual information in the form of attribute-value pairs. Attributes in infoboxes contain valuable semantic information, playing a key role in construction of a coherent large-scale knowledge base [2]. However, each language version maintains its own set of infoboxes with their own attributes, sometimes providing different values for corresponding attributes (as shown in Figure 4 [Figure 4: see original paper]). Thus, attributes in different Wikipedia versions must be matched to obtain coherent knowledge.

For instance, inconsistencies among 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 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.

Several challenges are 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 a person or place of production) and Synonym-Attributes (different attributes can have the same meaning, e.g., “alias” and “other names”), which lead to worse performance on label similarity or translation-based methods. Second, problems exist in

attribute values: 1) different measurements (e.g., Yao Ming' s height is 2.29m in English edition and 7 feet 6 inches in Chinese) and 2) timeliness (e.g., Beijing' s population is 21,150,000 (in 2013) in French edition). Thus, labels and values alone are not credible enough for cross-lingual attribute matching.

Main Idea. To solve these problems, we must first investigate several effective features considering characteristics of cross-lingual attribute matching. Then we 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.

Method. Figure 5 [Figure 5: see original paper] contains two parts. Part 1 on the left is a relation graph representing several relations in two Wikipedia editions K1 and K2. Different language versions are separated by a diagonal line. The attribute layer contains attributes and template relations among them. Similarly, the article layer contains articles and category relations. Imaginary lines between the two layers denote usage relations between articles and attributes. Red dashed lines denote existing cross-lingual links. Part 2 on the right is a factor graph. White nodes are variables with two types: xi and yi. Each candidate pair maps to an observed variable xi. The hidden variable yi represents a Boolean label (equivalent or inequivalent) of the observed variable xi. For example, x2 in Figure 5 corresponds to candidate attribute pair (pi3, pj2), and there exists a cross-lingual link between pi3 and pj2, so the hidden variable y2 equals 1. Black nodes in the factor graph are factors. There are three types: f, g and h. Each type is associated with a feature function which transforms relations into computable features.

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 depending on bilingual topic models have achieved state-of-the-art performance for this task. However, these models only consider textual context of concepts while outright ignoring explicit concept correlations such as those between concepts and their co-occurring words in text or those among concepts of ancestor-descendant relationships in a taxonomy.

Method. Cross-lingual taxonomy alignment is generally non-trivial. Fortunately, our goal is fairly simple: 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 precision, not recall. In XLORE2, we directly utilize cross-lingual links between categories provided by Wikipedia as cross-lingual links between concepts. This is a relatively simple method which permits investigation of more powerful methods to find more cross-lingual links between concepts in different languages in future work.

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 (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 XLORE2 data quality. 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 utilizing 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 candidate class and instance, respectively, taxonomies are directly derived from online wikis by transforming user-generated subsumption relations—namely subCategoryOf between two categories and articleOf from one article to one category—into semantic taxonomic relations: subClassOf between two classes and instanceOf from one instance to one class. Unfortunately, user-generated subsumption relations in Wikipedia and semantic taxonomic relations in knowledge bases are not exactly the same. Well-defined subClassOf and instanceOf essentially represent the isA relation, while freely edited subCategory and articleOf cover another topicOf relation which denotes topic-related relations and generates noise in the derived taxonomy. As Figure 6 [Figure 6: see original paper] clearly shows, when reasoning is based solely on derived taxonomy, the system mistakenly concludes that Barack Obama (person) isA Chicago, Illinois (location), which apparently should be a topic-related relation. Thus, it is an auspicious problem to have the opportunity to correct those wrong subClassOf and instanceOf relations.

Existing approaches typically suffer from the following problems:

1. **Language Dependence.** Heuristic-based methods strongly rely on accuracy of headword recognition algorithms and language-dependent rules, such as those involving Chinese and Japanese, making them too rigid to handle languages with no explicit plural/singular forms [14, 15].
2. **Limited Corpus.** Corpus-based methods depend on large-scale corpora with high quality, which are often simply unavailable [15]. Thus, generated taxonomies are often small, mostly domain-dependent, and have rather poor performance [16].

Main Idea. We formulate the above problem as 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 taxonomic relation prediction performance. Specifically, we tackle taxonomic relation prediction as a binary classification problem by learning two functions: the subClassOf Prediction Function and the instanceOf Prediction Function.

Method. Our model framework is shown in Figure 7 [Figure 7: see original paper]. First, we utilize a binary classifier as the basic learner and use Decision Tree as our implementation. We analyze features beneficial to taxonomic relation prediction, including linguistic heuristic features and language-independent structural features. Then we propose the Dynamic Adaptive Boosting (DAB) model for cross-lingual taxonomy derivation. To improve learning performance of taxonomic relation prediction, our model is trained iteratively on an active dynamic training set. Training examples are weighted samples from pre-labeled data and cross-lingual validated predicted data. We utilize cross-lingual validation to avoid potential performance degradation.

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

5.2 Fine-Grained Type Inference

Problem. Type information is very important for knowledge bases, but some large knowledge bases lack meaningful type information due to incompleteness. In XLORE, more than 18.7% of instances lack useful type information, so our target is to identify semantic types of instances 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, normally resulting in forming new type-paths in the taxonomy [18, 19, 20, 21, 22]. As shown in Figure 8 [Figure 8: see original paper], Yao Ming is associated with 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.

Existing approaches suffer from the following problems:

1. **Hand-crafted features.** Sentence-level methods focus on classifying instance mentions in text to a broad set of types to exploit well-defined linguistic features based on the mention itself and contextual information

in text. These require rich background knowledge [18, 19, 20, 21].

2. **Annotated corpus.** Corpus-level methods utilize annotated corpora to learn low-dimensional representations of instances then subsequently determine type inference based on learned embeddings. However, such large-scale high-quality annotated corpora are usually difficult to obtain [23, 24, 25], whereas access to simple entity text descriptions is often relatively easy.

Main Idea. To address these 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 encoding different levels of co-occurrence information and labeled information, then learn instance, word and type representations jointly via a network embedding method.

Method. Our model framework is shown in Figure 9 [Figure 9: see original paper]. First, we construct four heterogeneous networks to exploit different information effectively: word-word, instance-word, instance-type and type-word. Each network encodes specific semantic information. Then we utilize a heterogeneous network embedding method to learn low-dimensional representations for each instance and type based on these networks. Benefiting from heterogeneous networks, learned instance and type embeddings not only preserve semantic closeness but even more so facilitate higher quality predictive results for type inference. To meet fine-grained demand, we use learning-to-rank algorithms, greatly improving instance and type embedding quality. Finally, we use learned embeddings to force type inference for each unlabeled instance in XLORE2.

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

6. APPLICATION

Entity Linking (EL) is fundamental Natural Language Processing and Knowledge Engineering technology that builds bridges between plain text and knowledge bases. XLink is the entity linking system application of XLORE2. An entity link is the task of linking mentions in text to 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 Named Entity Recognition (NER) tools. However, NER tools rely heavily on language and only recognize three types of named entities:

Person, Location and Organization, leaving a significant gap in entity types covered in knowledge bases [32]. To address ambiguity and variation in entity linking, the simplest way is choosing the most prominent entity (i.e., the candidate with the largest number of incoming or outgoing links in Wikipedia) for a given mention. However, different contexts of mentions lead to different linking results, which becomes too complex to solve through entity priority mapping. An alternative strategy calculates contextual similarity for single mention linking and further employs topical coherence to collectively link all mentions within a document. Unfortunately, few systems consider features in a unified and effective manner. Moreover, these systems mainly use Wikipedia as the knowledge base and rarely handle Chinese documents. Additionally, many large-scale Chinese encyclopedias (e.g., Baidu Baike) are emerging and evolving, making it time to begin conducting entity linking in both Chinese and English.

Method. To address these issues, we develop a bilingual online entity linking system named XLink. As shown in Figure 10 [Figure 10: see original paper], 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 from the candidate set. XLink' s purpose is to provide users an online service for linking all important mentions in 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 taggers. Second, we design a generative probabilistic entity disambiguation method which models contextual feature, coherence feature and prior feature jointly to guarantee disambiguation accuracy. For system efficiency, we use the Aho-Corasick algorithm to parse mentions and introduce word and entity embeddings, ensuring time efficiency of the disambiguation phase. Finally, the disambiguation method is unsupervised to facilitate easy online deployment.

7. SYSTEM AND DATA STATISTICS

We construct XLORE2 in RDF form and use the OpenLink Virtuoso server for systematic 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 brief summary analysis of linked item enhancement and percentage increase 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(+90.32%)	422,974(+79.21%)
Properties	42,110	460,241	512,883(+616.36%)	10,532(NULL)

	English	Chinese	Total	Linked

We update our online system to illustrate XLORE2. As shown in Figure 11 [Figure 11: see original paper], one new application feature is that the system supports keyword-based or SPARQL queries. We also introduce several APIs for readers to access and download instances, concepts and properties in XLORE2, greatly facilitating relevant research. Another addition is our new entity linking system XLink, shown in the right part of Figure 11. XLink is an unsupervised bilingual entity linking system conducting mention parsing and entity disambiguation to link mentions in input documents to entities in XLORE2. We invite 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 problems 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 utilizing heterogeneous network embeddings and a 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 utilizing heterogeneous network embeddings to find missing instanceOf relations between instances and concepts automatically.

Finally, XLORE2 realizes significantly more facts compared with XLORE. We design an efficient entity linking system XLink, which can link mentions in documents to entities in XLORE2.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGEMENTS

The work is supported by the National Natural Science Foundation of China (NSFC) key project (No. 61533018, No. U1736204 and No. 61661146007), the 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, ···& 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)*, 2015, pp. 1–11. 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, ··· & J. Tang. XLore: A large-scale English-Chinese bilingual knowledge graph. In: *The 12th International Semantic Web Conference (ISWC2013) on Posters & Demonstrations Track*, 2013, No. D31. 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, pp. 1–6. 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*, 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*, 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*, 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 et al. (eds.) *The Semantic Web—ISWC 2017*. Cham, Switzerland: 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 et al. (eds.) *The Semantic Web—ISWC 2017*. Cham, Switzerland: 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 30th AAAI Conference on Artificial Intelligence (AAAI' 16)*, 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 et al. (eds.) *Joint International Semantic Technology Conference (JIST) 2017: Semantic Technology*. Cham, Switzerland: 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*, 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)*, 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)*, 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 embed-

ding. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 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*, 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)*, 2012, pp. 94–100. 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)*, 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: M. Sun et al. (eds.) *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Cham, Switzerland: Springer, 2018, pp. 170–182. doi: 10.1007/978-3-030-01716-3_{15}.

[27] R. Mihalcea, & A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In: *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 2007, pp. 233–242. doi: 10.1145/.

[28] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, ···& G. Weikum. Robust disambiguation of named entities in text. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 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, 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*, 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*, 2005, pp. 363-370. Available at: https://dl.acm.org/citation.cfm?id=1219885.

AUTHOR BIOGRAPHY

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

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

Jing Zhang is currently a graduate student at the Department of Computer Science and Technology, Tsinghua University. She received her Bachelor's 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 at the Department of Computer Science and Technology, Tsinghua University. He received his PhD degree from Tsinghua University in 2016 under supervision of Prof. Juanzi Li, and his Bachelor' s 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 is a Professor at the Department of Computer Science and Technology, Tsinghua University. She received her PhD degree in Computer Science from Tsinghua University in 2000. 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 at the Department of Computer Science and Technology, Tsinghua University. He received his Master's degree from

Tsinghua University in 2005 and Bachelor' s degree from Tsinghua University in 2002. He is the designer of the XLORE system.

Note: Figure translations are in progress. See original paper for figures.