Challenges in Chinese Knowledge Graph Construction

Chengyu Wang, Ming Gao, Xiaofeng He, Rong Zhang*

Shanghai Key Laboratory of Trustworthy Computing
Data Science and Engineering Institute, East China Normal University
3663 North Zhongshan Road, Shanghai, China
chengyuwang@ecnu.cn, {mgao,xfhe,rzhang}@sei.ecnu.edu.cn

***Abstract*—The automatic construction of large-scale knowledge graphs has received much attention from both academia and industry in the past few years. Notable knowledge graph systems include Google Knowledge Graph, DBPedia, YAGO, NELL, Probase and many others. Knowledge graph organizes the information in a structured way by explicitly describing the relations among entities. Since entity identification and relation extraction are highly depending on language itself, data sources largely determine the way the data are processed, relations are extracted, and ultimately how knowledge graphs are formed, which deeply involves the analysis of lexicon, syntax and semantics of the content. Currently, much progress has been made for knowledge graphs in English language. In this paper, we discuss the challenges facing Chinese knowledge graph construction because Chinese is significantly different from English in various linguistic perspectives. Specifically, we analyze the challenges from three aspects: data sources, taxonomy derivation and knowledge extraction. We also present our insights in addressing these challenges.**

I. Introduction

Automatic construction of knowledge graph based on crowd-sourced data has attracted significant interest from both academia and industry due to its wide application in areas like semantic search, machine reading and question answering. Projects such as DBPedia[1], YAGO[2], Kylin/KOG([3], [4]) and BabelNet[5] aim at building knowledge graphs by harvesting entities and relations from Wikipedia, one of the largest crowd-sourced, multilingual wikis on Web. Other projects, such as NELL[6], TextRunner[7] and Probase[8], take more aggressive approach by extracting knowledge from unstructured text in Web pages.

In the wide-spread mood of enthusiasm on knowledge graph, we notice that its construction is quite language-dependent. Data sources as well as the NLP or other methods with which to process the data are unique among languages, especially for those belonging to different language families. Currently, most projects are concerning knowledge graph systems in English language. Because Chinese belongs to a different language family, directly translating English knowledge graphs into Chinese is not always feasible, hence Chinese knowledge graph construction is of great significance.

In this paper, we focus on issues raised during Chinese knowledge graph construction, discussing major challenges in this area. The significances of our paper are:

* Corresponding author

1) The data sources in Chinese have quite different characteristics from English one. A number of online Chinese wikis are publicly available, such as Chinese Wikipedia, Baidu Baike (baike.baidu.com), Hudong Baike (www.baike.com), etc. However, they differ in data size and format; user generated tags and the data quality vary, too. Currently Chinese Wikipedia contains only 0.8M articles, while Baidu Baike and Hudong Baike have over 10M, respectively. Knowledge extraction and integration from these heterogeneous data sources poses greater challenge.
2) Resources for building Chinese knowledge graphs are limited. Some Chinese equivalents used as important parts in English knowledge graph construction are not readily available. For instance, there are no public knowledge repositories (e.g., *Freebase.com*) and semantic networks (e.g., WordNet) in Chinese for fact generating and taxonomy building.
3) Most information extraction algorithms ([9], [10]) are language-dependent. Chinese is different from English in vocabulary, semantics and grammar. For example, in Chinese nouns there are no explicit singular/plural forms which are used to detect conceptual entities in building English knowledge graphs. New approaches are needed since it is doomed to fail if directly applying existing techniques to Chinese scenarios.

Motivated by these observations, we describe the research challenges in Chinese knowledge graph construction in following three aspects: (i) *quality of data sources*, (ii) *taxonomy derivation* and (iii) *knowledge harvesting*.

In this paper, we discuss these challenges in details and present our insights into them based on our practice.

II. Research Challenges

A. Quality of Data Sources

Although Wikipedia is usually treated as high-quality data source for many knowledge graphs such as YAGO and DBPedia, there are several quality issues in Chinese Wikipedia that should be paid special attention to.

1) **Data Sparsity:** Wikipedia is a multi-lingual online encyclopedia. However, there exists a clear imbalance between different language versions. There are over 4 million articles in

English Wikipedia, while only about 0.8 million in Chinese. Furthermore, English Wikipedia contains 13 times more infoboxes than Chinese Wikipedia[11]. The sparsity due to the lack of infoboxes results in difficulty in knowledge harvest, which can be denoted by following two questions:

- A lot of semantic relations between entities and entity properties are missing. How can we construct a "dense" graph out of such data sources?
- Chinese Wikipedia contains much fewer tail entities because of the small number of wiki pages. How can we design a mechanism that can identify the missing tail entities from other Web sources such that the knowledge graph has high coverage?

2) **Information Accuracy:** In Web 2.0 era, most information on the Web is generated by Internet users. Inaccuracy and errors are inevitable, and Wikipedia is not immune. For articles introducing professional knowledge, errors occur due to the lack of expertise of the contributor. On the other hand, articles related to sensitive issues or events may be written in favor of the contributor's attitude. For instance, the page about PX (short for P-Xylene, a chemical material with toxicity slightly higher than ethanol) on Baidu Baike was modified back and forth tens of times with extremely different toxicity description, due to the polarized attitudes of the contributors towards the construction of a PX factory in city of Xiamen, China. Information extracted from pages like this will certainly affect the correctness of the knowledge graph. Attention should be paid to pages with high number of modifications while some attribute descriptions are quite different from version to version. Sophisticated NLP methods will be required to accomplish this task.

3) **Linking Quality:** Wiki pages contain hyperlinks to other wiki pages about entities appearing in anchor text. The link structure of these pages can be leveraged to construct semantic networks and perform tasks like entity linking[12]. In these scenarios, the link structure in Wikipedia serves as the "gold standard" to provide support for other tasks. However, in Chinese Wikipedia, we have identified many errors in links, in that the entity in anchor text is different from the target entity. Here is an example: the entity *Wu Mei* (a professor in Peking University) appeared in the page *May Fourth Movement* (a social and political movement in China) is linked to *Wu Mei* (a dubbing actress in Hong Kong). This error link was caused by simply matching entities according to surface names without considering contextual information.

We perform detailed analysis on the error linking problem in Chinese Wikipedia and find that most of the errors are related to Chinese person names, with a few related to organization names. It may be due to the phenomenon that some popular names are frequently used in China. This problem causes increased difficulty in tasks such as entity matching and record linkage, thus lower the quality of knowledge graph.

In summary, these quality issues should not be overlooked when building a machine-readable Chinese knowledge graph. We argue that it is difficult to find a ground truth and put it into the knowledge graph directly. Instead, quality control mechanisms should be developed to solve these problems. Furthermore, to address the sparsity issue, multiple data sources can be considered together, each of which has confidence scores indicating the level of correctness. Errors can be detected through various techniques such as entity integration [13] and reasoning-based methods.

B. Taxonomy Derivation

The taxonomy is the core of large-scale knowledge graphs in that it provides a hierarchical type system for entities. Ideally, a taxonomy provides two types of relations: the *subClassOf* relation between two classes and the *instanceOf* relation between a class and an entity. In a knowledge graph, every entity e should belong to a class c indicating the type of the entity, represented as $(e, instanceOf, c)$. A class can be a subclass of another class, or the *root class* itself. The goal of taxonomy derivation is to identify such a hierarchical structure of classes from data sources, thus plays an important role in building knowledge graphs.

In wikis, the categories and their hierarchical structures can help derive the taxonomy, but these user-generated categories are often *topical* or *thematic*, rather than *semantically taxonomic*. The resulting problem is that the Wikipedia category system does not provide a taxonomic structure between classes. We also found that many categories in other Chinese wikis such as Baidu Baike only have the semantic associativity with the entity (i.e., *relatedTo* or *topicOf*), instead of the strict *instanceOf* relation. For example, for Chinese president *Xi Jinping*, *political leader* is a valid class rather than *politics* or *China*.

Intensive research has been conducted on English content. Because WordNet contains abundant semantic classes and their relations, Suchanek et al.[2] combined the hypernymy/hyponymy relations in WordNet and facts derived from Wikipedia to generate taxonomy with high accuracy and coverage. In WikiTaxonomy[9], semantic relations between categories are classified into *isa* and *notisa* relations through connectivity network and lexico-syntactic patterns.

However, there are two challenges when trying to extract such information in Chinese language. First, there is no Chinese version of WordNet publicly available to construct the *subClassOf* relations between semantic classes. Second, mapping entities to their corresponding classes (i.e., establishing *instanceOf* relations) is a non-trivial task because heuristic-based methods mentioned above strongly rely on the head words of category names, and these language specific rules and patterns do not work in Chinese where no explicit singular/plural forms exist.

The basic research issues we suggest in this field include:

1) study of lexico-syntactic patterns of category names in Chinese;
2) classification of semantic relations between classes and entities based on machine learning techniques;
3) machine translation based methods which map established semantic relations in other languages to Chinese;

4) construction of a complete taxonomy from individual *subClassOf* and *instanceOf* relations.

C. Knowledge Harvesting

A knowledge graph is basically a collection of classes, entities and relations between them (i.e., facts). Here we discuss some challenges generating relational facts in Chinese.

1) **Hearst Patterns:** Hearst patterns[14] are high quality lexico-syntactic patterns that indicate hyponymy relations from text. For example, in pattern "NP_0 $such\ as$ $\{NP_1, NP_2, \ldots, (and|or)\}$ NP_n", where NP_i are noun phrases, we can infer NP_0 is the hyponym of NP_i ($i > 0$). Hearst patterns can be introduced to extract concepts and *isa* relations from text. They were used to build the largest taxonomy from Web pages[8].

The *isa* relations are valuable information for taxonomy derivation in knowledge graph construction process. However, to best of our knowledge, there is no comprehensive study on Hearst patterns for Chinese language yet. The challenges lie in following aspects:

- Basic NLP analyses such as word segmentation and part-of-speech (POS) tagging do not perform well in Chinese, especially for Web texts where informal (Web style) words are frequently used.
- For large-scale knowledge extraction, precision is more important than recall due to the high noise nature of the mass data. But high quality Hearst patterns in Chinese are difficult to find. The grammar of Chinese language is relatively flexible compared to English, thus patterns with high extraction power, i.e., having high precision in extraction results, are rare.
- In Chinese, *isa* relations are sometimes not explicitly expressed in text. Considering textual patterns only will harm the recall.

Despite the aforementioned difficulties, we argue that in-depth analysis on Hearst patterns in Chinese can have following benefits: (i) enriching current knowledge graphs; (ii) promoting better Chinese text understanding; (iii) providing lexical database (like WordNet) for future research.

2) **General Relation Extraction:** It is prohibitive to manually compile patterns to extract instances of arbitrary target relations. Most relation extraction (RE) systems including Snowball[15], KnowItAll[16] addressed this issue based on the observation of *duality of facts and patterns*.

Even with above efforts, building a general Chinese RE system still remains a challenge. The root cause lies in the flexibility in Chinese language expression. We now face a dilemma when generating patterns: if too many constraints are imposed, then the coverage of extraction will drop rapidly; in contrast, general patterns can lead to the problem of *semantic drift*.

We suggest that advanced pattern generation strategies be tailored to Chinese language, such as adding statistical or NLP-based features. Also, because the results extracted from semi-structured and structured input have higher precision, facts with high confidence can help supervise the RE process from natural language input. How mechanisms such as *distant supervision* can be employed in Chinese RE is worthy of being considered in developing large-scale Chinese RE systems.

III. Conclusion

In this vision paper, we describe several challenges facing the construction of Chinese knowledge graphs, and propose the research directions addressing these challenges. Specifically, we identify and analyze the research challenges from three aspects encountered during our research. We discuss some typical quality issues in data sources that have not been well addressed in existing work. The taxonomy derivation focuses on automatic taxonomy construction where existing lexical databases and language rules are not easily adopted into Chinese scenarios. Besides, we present the challenges of knowledge harvesting in Chinese from linguistic perspective. We hope our insights can be of help for the work of Chinese knowledge graph, and serve as useful hints for applications in other languages too.

Acknowledgement

This work is partially supported by NSFC under Grant No. 61402177, and the Key Program of National Natural Science Foundation of China under Grant No.61232002.

References

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2014.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, 2007, pp. 697–706.

[3] D. S. Weld, R. Hoffmann, and F. Wu, "Using wikipedia to bootstrap open information extraction," *SIGMOD Record*, vol. 37, no. 4, pp. 62–68, 2008.

[4] F. Wu and D. S. Weld, "Automatically refining the wikipedia infobox ontology," in *WWW*, 2008, pp. 635–644.

[5] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *ACL*, 2010, pp. 216–225.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *AAAI*, 2010.

[7] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *IJCAI*, 2007, pp. 2670–2676.

[8] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in *SIGMOD Conference*, 2012, pp. 481–492.

[9] S. P. Ponzetto and M. Strube, "Deriving a large-scale taxonomy from wikipedia," in *AAAI*, 2007, pp. 1440–1445.

[10] G. de Melo and G. Weikum, "MENTA: inducing multilingual taxonomies from wikipedia," in *CIKM*, 2010, pp. 1099–1108.

[11] Z. Wang, Z. Li, J. Li, J. Tang, and J. Z. Pan, "Transfer learning based cross-lingual knowledge extraction for wikipedia," in *ACL*, 2013, pp. 641–650.

[12] W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: linking named entities with knowledge base via semantic knowledge," in *WWW*, 2012, pp. 449–458.

[13] Z. Nie, J. Wen, and W. Ma, "Statistical entity extraction from the web," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2675–2687, 2012.

[14] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *COLING*, 1992, pp. 539–545.

[15] E. Agichtein and L. Gravano, "*Snowball*: extracting relations from large plain-text collections," in *ACM DL*, 2000, pp. 85–94.

[16] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-scale information extraction in knowitall: (preliminary results)," in *WWW*, 2004, pp. 100–110.