Towards Automatic Construction of Knowledge Bases from Chinese Online Resources

Liwei Chen, Yansong Feng, Yidong Chen, Lei Zou, Dongyan Zhao
Institute of Computer Science and Technology
Peking University
Beijing, China
{clwclw88,fengyansong,chenyidong,zoulei,zhaodongyan}@pku.edu.cn

Abstract

Automatically constructing knowledge bases from online resources has become a crucial task in many research areas. Most existing knowledge bases are built from English resources, while few efforts have been made for other languages. Building knowledge bases for Chinese is of great importance on its own right. However, simply adapting existing tools from English to Chinese yields inferior results.In this paper, we propose to create Chinese knowledge bases from online resources with less human involvement.This project will be formulated in a self-supervised framework which requires little manual work to extract knowledge facts from online encyclopedia resources in a probabilistic view.In addition, this framework will be able to update the constructed knowledge base with knowledge facts extracted from up-to-date newswire.Currently, we have obtained encouraging results in our pilot experiments that extracting knowledge facts from infoboxes can achieve a high accuracy of around 95%, which will be then used as training data for the extraction of plain webpages.

1 Introduction

As the development of world wide web (WWW), the volume of web data is growing exponentially in recent years. Most of the data are unstructured, while a few are manually structured and only a small part of them are machine-readable. How to make these data accessible and useable for end users has become a key topic in many research areas, such as information retrieval, natural language processing, semantic web(Tim et al., 2001) and so on. Among others, constructing knowledge bases (KB) from web data has been considered as a preliminary step. However, it is not trivial to extract knowledge facts from unstructured web data, especially in open domain, and the accuracy is usually not satisfactory. On the other hand, with the development of Web2.0, there are increasing volume of online encyclopedias which are collectively created by active volunteers, e.g., Wikipedia[1]. Surprisingly, experiment evidences show that the confidence of Wikipedia is even comparable with that of British Encyclopedia (Giles, 2005). Therefore, many efforts have been made to distill knowledge facts from Wikipedia or similar resources and further build KBs, for example YAGO(Suchanek et al., 2007), DBpedia(Bizer et al., 2009) and KOG(Wu and Weld, 2008).

In the literature, most KBs constructed recently are in English as it takes up an overwhelming majority on the web, while other major languages receives less attention, for example, Chinese features similar amounts of web pages with English yet is less frequently studied with regarding to building KBs. Although continuous works have been made to process English resources, building Chinese KBs is of great value on its own. To the best of our knowledge, few efforts have been made to construct a KB in Chinese until now. Despite of necessary special preprocessings, e.g., word segmentation, for Chinese, building a Chinese KB from web data is quite different from building English ones, since we have limited resources available in Chinese that are of lower

[1] http://www.wikipedia.com

Proceedings of the 2012 Student Research Workshop, pages 67–72,

quality compared to their English counterparts. This brings more difficulties than that of English. As a result, the approaches used in English may not work well in Chinese.

In this paper, we propose a new framework to build a KB in Chinese from online resources without much human involvement. Since the Chinese portion of Wikipedia is much smaller than its English part, we harvest knowledge facts from a Chinese online encyclopedia, HudongBaike[2]. HudongBaike is the largest Chinese online encyclopedia and features similar managing rules and writing styles with Wikipedia. We first obtain knowledge facts by parsing the infoboxes of HudongBaike. Then we use these triples as seeds, and adopt the idea of distant supervision(Mintz et al., 2009; Riedel et al., 2010; Yao et al., 2010) to extract more facts from other HudongBaike articles and build a KB accordingly. Moreover, to make the knowledge base more up-to-date, we also propose to propagate the KB with news events.

The rest of this paper is organized as follows: we first introduce the related work, and briefly introduce two online encyclopedias. In Section 4 we describe our framework in detail. Our current work are discussed in Section 5. In Section 6 we conclude this paper.

2 Related Work

KB construction is an important task and has attracted many research efforts from artificial intelligence, information retrieval, natural language processing, and so on. Traditional KBs are mostly manually created, including WordNet(Stark and Riesenfeld, 1998), Cyc or OpenCyc(Matuszek et al., 2006), SUMO(Niles and Pease, 2001), and also some domain-specific ontologies such as GeneOntology[3]. These KBs achieve a high accuracy since they are manually built or filtered by domain experts. However, manually creating KB is a time-consuming and labor-intensive work, and continuous annotation is required to keep the KB up-to date. Most of them thus suffers from the coverage issue in practice.

In recent years, many researchers turn to automatically extract knowledge to construct KBs. One kind of methods extract knowledge facts from general text corpus. These approaches, such as TextRunner(Banko et al., 2007) and KnowItAll(Etzioni et al., 2004), use rule based information extraction technologies to extract relations between entity pairs. Recently, TextRunner is expanded by a life long learning strategy, which can acquire new facts. Another type of approaches aims to automatically derive facts from online encyclopedias. Collectively created by many volunteers, online encyclopedias are more reliable than general web pages. They also contain semi-structured knowledge such as handcrafted infoboxes. Therefore, the accuracy of the facts extracted will be higher. Researchers utilize these semi-structured data resources for knowledge extraction, for example, YAGO extract facts from infoboxes and category names of Wikipedia, and use WordNet as its taxonomy(Suchanek et al., 2007). A similar approach is adopted by DBpedia, which also extract knowledge facts from infoboxes(Bizer et al., 2009). Unlike YAGO and DBpedia, Kylin uses the infoboxes and the Wikipedia pages containing these infoboxes to build a training set, and use machine learning methods to extract facts from plain Wikipedia articles(Wu and Weld, 2007). Although Kylin achieves a high precision, it is corpus-specific, which means it can only be used in Wikipedia-like corpora. It is noticed that all the above works focus on building an English KB, and few efforts have been made in building a Chinese one until now.

3 Online Encyclopedia

Wikipedia is known as an accurate online encyclopedia whose accuracy is comparable with Encyclopedia Britannica(Giles, 2005). It's created by thousands of volunteers around the whole world. Until now, the English version of Wikipedia has 3,878,200 content pages, making it the largest English online encyclopedia. The Chinese version contains 402,781 content pages, which is much smaller than the English version.

HudongBaike is the largest Chinese online encyclopedia with over 5 million content pages. Similarly with Wikipedia, HudongBaike is also created by volunteers, and relies on the community to ensure its quality. Many HudongBaike pages also contains

[2]http://www.hudong.com

[3]http://www.geneontology.org

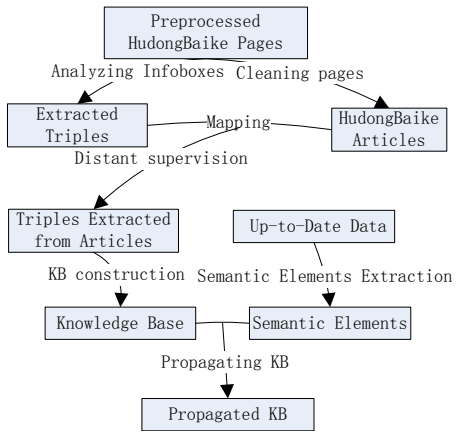

Figure 2: The framework of Our project

a hand-crafted summary box, *infobox*. An infobox summarizes the knowledge of the corresponding entity. The information in the infobox is reliable since these are collaboratively crafted by many volunteers. Figure 1 is an example page with an infobox from HudongBaike, introducing a US general 乔治·马歇尔 (George Marshall).

4 The Framework

In this paper, we formulated the KB construction task in a semi-supervised learning fashion which requires little manual annotation and supports knowledge propagation by up-to-date feeds. Because the Chinese part of Wikipedia is relatively small and may suffer from the coverage problem, we use HudongBaike to build our KB in this project. In future we may merge the Wikipedia part into our KB. After necessary preprocessings including word segmentation and named entity extraction, we are able to apply our framework shown in Figure 2.

In general, our framework contains the following steps: (1)Extracting knowledge from online encyclopedia; (2)Linking triples and building KB; (3)Propagating KB with up-to-date data.

4.1 Entity Relation Extraction

Compared to other resources on the Web, online encyclopedias contain less noises and feature more regular structures, thus are considered easier for us to extract knowledge facts.

Analyzing Infoboxes As mentioned before, many HudongBaike pages contains an infobox, which has high accuracy and can be used directly for relation extraction. We can conveniently parse these infoboxes into $< S, P, O >$ triples. For example, from the first entry of this infobox, we can derive the following triple: $<$ 乔治·马歇尔，出生地，尤尼恩敦 $>$($< GeorgeMarshall, BirthPlace, Uniontown >$). The precision of the extraction is over 95%, and these triples can form a valuable knowledge source.

Extracting relations with Distant Supervision Extracting knowledge from infoboxes is efficient and can achieve a high precision. However, many web pages in HudongBaike do not have infoboxes. There is much richer knowledge in the main articles of HudongBaike, which we should also take into consideration.

Extracting knowledge from unstructured articles is a challenging task. Traditionally, researchers use manually created templates to extract relations. These templates need lots of human efforts and are domain-specific. Recent methods trend to rely on machine learning models, which need a large amount of labeled data. One idea is to utilize the infoboxes to form the training data set, and train an extractor to extract relations from the pages without an infobox(Wu and Weld, 2007). However, the relations extracted from a page are restricted to the infobox template used by the current page category, and their subject must be the entity that this page describes. For example, when we extract relations from the page of 查克·叶格 (Charles Yeager, Ace of US in WWII) which does not contain an infobox, the subject of these relations must be Charles Yeager, and we can only extract the relation types listed in infobox template for a military person. As a result, this method can only be used in online encyclopedias in a Wikipedia style, and the recall will be relatively low.

Distant supervision is widely used in relation extraction in recent years. It hardly need any manual work, and can overcome the above problems. It can be used in any reliable corpus, and doesn't have the strict restrictions as previous methods. We adopt its idea in our framework. The basic assumption of distant supervision is the sentences containing two en-

Figure 1: A HudongBaike page about a US general George Marshall

tities should express the relation between them more or less. It only needs a reliable seed KB (in the form of relation triples) and a corpus. Here, we can use the knowledge facts extracted from infoboxes previously as the seed KB, and the articles of HudongBaike as text corpus. For each triple in the seed KB, we generate positive training data by finding sentences containing both its subject and object in the corpus. For example, we can map the first entry in Figure 1 to the sentence 1880年12月31日，马歇尔出生在尤尼恩敦 (On December 31th, 1880, Marshall was born in Uniontown). The negative training data can be generated by randomly select some sentences which contain neither of the subject and the object. A predictive model such as logistic regression model is trained with the training data. We can use the model to give predictions for the relations in a textual knowledge source. For a HudongBaike page, we should decide the entity pairs we are interested in. A simple strategy is to select all entity pairs. But it will be time-consuming, and may suffer from weak-related entity pairs. So we extract topic entities which have high *tfidf* weights from this page, and generate entity pairs under the restriction that they must contain at least one topic entity. For each entity pair, we find the sentences which contain both the subject and object and use the predictive model to give the possible relations between them and the confidence of the relations.

However, the predictions of distant supervision is less accurate than those of supervised methods. So we should adopt some heuristics to filter the relations extracted. An easy strategy is to set up a threshold for relation confidences to avoid uncertain relations and improve the precision. We adopt this method in our project. Furthermore, we can also use the strategies of Riedel et al. (2010) or Yao et al. (2010).

4.2 Knowledge Base Construction

After the relation extraction, we must link the extracted knowledge triples in order to construct the knowledge base. In our scenario this linking task can be formulated as: given a base KB, a bunch of newly extracted knowledge triples with the sentences describing them and their contexts, the task of entity linking aims to link each of the entity mentions in the plain texts (these sentences mentioned above) to its corresponding entity in the base KB. At the very beginning, we initiate a base KB by using the taxonomy of HudongBaike thus are able to map relations between entities into the KB through entity linking.

In online encyclopedias, the synonyms of an entity are represented by redirect links. Synonyms are important in entity linking because they provide alternative names for entities, and we may miss some mappings without them. For example, we have an entity 美利坚合众国 (United States of America) in the KB, and an mention 美国 (USA) in a piece of text. Redirect links can tell us that we can create a mapping between them. Basically, for each mention, we can find matching candidates for them in a KB through exact matching. However, if we cannot find an exact match for a mention, we will try

fuzzy matching since a mention may not match exactly with its referent entity in KB.

Now we need to solve the entity linking task. Traditional methods did not exploit global interdependence between entity linking decisions. We thus adopt the collective entity linking approach of Han et al. (2011) to solve this problem. This method captures the rich semantic relatedness knowledge between entities, and take the interdependence of linking decisions into consideration. They construct a graph by linking name mentions and candidate entities in pairwise using the semantic relatedness between them. Then they use a random walk algorithm on the graph to solve the problem. However, they did not take the NIL problem into consideration. That is, in entity linking, if the referent entity of an name mention is not in our KB, it should be linked to a pseudo entity NIL. In our case, we should abandon the mapping of the current triple by deciding whether this entity has been listed in the KB(Zheng et al., 2010).

4.3 Knowledge base Propagation

Although we can extract millions of relations and built a KB in previous subsections, it has the same shortage as most existing KBs: the knowledge extracted are mostly statical attributes of entities (such as birthdate or occupation of a person) and can not describe the latest updates of an entity (such as a politician is currently visiting a country).

In order to settle this problem, we use the dynamical knowledge extracted from up-to-date data to expand our KB. One possible solution is extracting semantic event elements from online news. In this project, we will synchronies our KB with a Chinese newspaper, RenMinRiBao (People's Daily).

5 Current Work

Currently, we have extracted triples from the infoboxes of HudongBaike and built the base KB. Manual evaluation shows that the precision of structured content extraction is over 95%. Most errors are caused by the web page's own mistakes or editing errors in infoboxes.

To assess the quality of HudongBaike data, in our preliminary experiments(Yidong et al., 2012), we extract relation facts from plain HudongBaike articles without infoboxes in a way similar to Kylin. We focus on three categories, including 国家 (Nation), 人物 (Person) and 演员 (Actor or Actress). In each category we select several representative attributes from its infobox template. We manually annotated more than 200 testing examples for evaluation: 100 in Person, 33 in Nation and 91 in Actor or Actress. The results shows that the HudongBaike data can be used to extract knowledge facts with a high precision in all three categories: in 人物 the average precision is 79.43%, in 国家 it is 78.9%, and in 演员 it even goes up to 90.8%.

Distant Supervision We further adopt the approach of distant supervision(Mintz et al., 2009) in a Chinese dataset. We generate a dataset from RenMinRiBao with 10000 sentences, and each sentence contains at least a pair of entities which correspond to a knowledge triple in HudongBaike's infobox extraction. We use 60% of the sentences as training set and 40% as the testing set. Our experiments show that when the recall is 10%, we can obtain a high precision of 87%, which indicates the feasibility of our model. However, as the recall raises, the precision drops dramatically. For example, when the recall is 29% the precision is about 65%. This can be remedied by adopting more encyclopedia-specific filtering strategies and assumptions during the distant supervision modeling.

6 Conclusions

In this project, we proposed a framework to build KBs in Chinese. It uses the infoboxes of HudongBaike as a seed knowledge base, the articles of HudongBaike as extra textual resources, adopts the idea of distant supervision to extract knowledge facts from unstructured data and link the triples to build a knowledge base. This framework requires little manual work, and can be used in other reliable knowledge resources. Our preliminary experimental results are encouraging, showing that the HudongBaike provides reasonable resources for building knowledge bases and the distant supervision fashion can be adapted to work well in Chinese.

For the next, we will further adapt our framework into a self-training manner. By using higher threshold for confidence in distant supervision we can make sure the precision of extracted knowledge

is high enough for bootstrapping. Then we put the extracted knowledge facts into the seed KB, and the framework will repeat iteratively. On the other hand, we can extract knowledge facts from other reliable knowledge resource, such as Wikipedia, academic literature, and merge knowledge from different resources into one KB. Moreover, we can also make our KB multilingual by adopting our framework in other languages.

References

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of IJCAI*, IJCAI'07, pages 2670–2676.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7:154–165.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in knowitall. In *Proceedings of the 13th WWW*, WWW '04, pages 100–110.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438:900–901.

Han, X., Sun, L., and Zhao, J. (2011). Collective entity linking in web text: a graph-based method. In *SIGIR*, SIGIR '11, pages 765–774, New York, NY, USA. ACM.

Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In *Proceedings of the 2006 AAAI Spring Symposium*.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of FIOS - Volume 2001*, pages 2–9. ACM Press, New York.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer Berlin / Heidelberg.

Stark, M. M. and Riesenfeld, R. F. (1998). Wordnet: An electronic lexical database. In *Proceedings of 11th Eurographics Workshop on Rendering*. MIT Press.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of WWW*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Tim, B.-L., J., H., and O., L. (2001). The semantic web. *Scientific American*.

Wu, F. and Weld, D. S. (2007). Autonomously semantifying wikipedia. In *CIKM*, CIKM '07, pages 41–50, New York, NY, USA. ACM.

Wu, F. and Weld, D. S. (2008). Automatically refining the wikipedia infobox ontology. In *WWW*, WWW '08, pages 635–644, New York, NY, USA. ACM.

Yao, L., Riedel, S., and McCallum, A. (2010). Collective cross-document relation extraction without labelled data. In *Proceedings of EMNLP*, EMNLP '10, pages 1013–1023, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yidong, C., Liwei, C., and Kun, X. (2012). Learning chinese entity attributes from online encyclopedia. In *Proceedings of IEKB workshop in APWeb 2012*.

Zheng, Z., Li, F., Huang, M., and Zhu, X. (2010). Learning to link entities with knowledge base. In *HLT-NAACL 2010*, pages 483–491, Stroudsburg, PA, USA.