The method of Constructing Chinese Knowledge Base based on open source English Knowledge Base

Zhonghe He[a], Yunbao Gong[b], Liang Gan[c] and Xiaohui Chou[d]

Department of Computer, National University of Defense Technology, Changsha 410073

[a]riterhe@outlook.com, [b]738793262@qq.com, [c]gl.nudt@gmail.com, [d]570946291@qq.com

Keywords: Knowledge base, Chinese, YAGO, Translation.

Abstract. Building a large-scale knowledge base is a hot spot in the current research of big data, and it is the basis of Internet search, Question Answering (Q&A), Machine Translation (MT) etc. At present, it has formed serval mature English knowledge base, such as YAGO, Freebase, DBpedia, etc. but we still lack a large-scale Chinese knowledge base. This paper presents an approach to construct the Chinese knowledge base by using the automatic translation technique and the method of building the Chinese knowledge base with the source of English knowledge base. And proposing a quality calibration method based on the search engine. In order to verify the above method, we take the YAGO English knowledge base as an example, the formation of Chinese Knowledge most quality of knowledge is up to 95%, and a few low quality data can be discarded after being calibrated. The method is an important supplement to building a large-scale knowledge base by using text information extraction, crowdsourcing and other methods.

1. Introduction

Building a large-scale knowledge base is a hot spot in the current research of big data, and it is the basis of Internet search, Question Answering (Q&A), Machine Translation (MT) etc. With the influence of “the big search in CyberSpace [1]”, which is aim to construct a framework of the wisdom network of people, material and information search. So, it is imperative to construct a large scale Chinese knowledge base as accurate as possible with a reasonable price.

At present, there are many open source English knowledge base, including DBpeida [2], Freebase, Probase [3] and YAGO [4,5,6] etc. DBpeida is extracted from the structured information in Wikipedia and its associated data in the form of data sharing in Web [7]. Until 2014, the DBpedia knowledge base has more than 3.64 billion items involving person names, place names, music, movies, group, race, and other multiple categories, Freebase is a large-scale open structured data sets, following the knowledge sharing protocol of Creative Commons(CC), using Wikipedia and MusicBrainz as base source, Covering sports, geography, health care, education, government, finance and other aspects. but most of the knowledge is collected, sorted and summarized by artificial. Probase is based on the knowledge integration technology of resolution entity, which is based on probabilistic entity resolution. Existing structured data, such as Freebase, IMDB, Amazon and others were integrated into the Probase [8]. As of 2014, the core concept of the more 2.7 million, the total amount of the concept reached 10 million. Max Planck Institute for Computer Science integrated the data sources such as Wikipedia, WordNet, GeoNames into YAGO. It is a large scale semantic knowledge base with a wide coverage and high data quality.

Except the DBpedia of the above open source knowledge base, others are not support Chinese. Of cause, there are some Chinese knowledge, such as Fudan University's Chinese knowledge mapping platform, Sogou know cubic knowledge base, and HowNet. But these Chinese knowledge base are not open source. The typical methods of constructing Chinese knowledge base include: information relation extraction, artificial construction of experts, and Crowdsourcing [9]. Information extraction(IE) is to extract structured information from unstructured and semi-structured text, but most of the time, IE need to extract knowledge from all kinds of Natural Language documents, using some automatic or semi-automatic algorithm. The construction of large scale knowledge base by experts is difficult to realize, because of its high cost of time and cost.

This paper puts forward a method of constructing Chinese knowledge base by using automatic translation technology, using open source English knowledge base as its source, and puts forward the method of quality calibration based on search engine. On the one hand, using mature open source English knowledge base quickly forms a high degree of accuracy of Chinese knowledge, on the other hand, avoiding the problem of artificial construction of the expert knowledge base. Although there are differences between Chinese and English, but according to the experimental result, the way to understand the world culture of each language are interlinked, similar, even the same. In some respects, such as human knowledge, geographical knowledge, classification knowledge, we convert those English knowledges to Chinese knowledges, and the accuracy reached 95% after the calibration. Of course, In some respects, their accuracy is very low, during the process of calibrating we have discard this part of the data automatically. YAGO2s knowledge base is used as the experimental data to verify the above method.

2. Basic content of YAGO

YAGO2s contains a total of 25 files, respectively record the classification, labels, entities, categories, facts and geographic information, etc. All knowledge in YAGO is in accordance with the WordNet. Table 2-1 lists the contents of several major documents.

Table 2-1 Main documents in YAGO

File Name	Content	File Name	Content
yagoDBpediaClasses	Map of classes and URI	yagoGeonamesGlosses	Description in GeoNames
yagoDBpediaInstances	Map of instances and URI	yagoGeonamesData	Geographic data in GeoNames
yagoTypes	Map of instances and classes	yagoTaxonomy	Classified information of YAGO
yagoLabels	Map of facts and labels	yagoFacts	Relation between Facts
yagoGeonameClasses	Classes in GeoNames		

In these files, file yagoFacts records relationship between entities, file yagoTaxonomy records relationship between classes, and file yagoType records the relationship between classes and entities.

3. Transformation process

The method proposed in this paper uses an automatic translation technology and a calibration technology based on mutual information to construct a Chinese Knowledge Base with the minimum cost. The translation process of the knowledge base uses the online API provided by Baidu translate. Preprocess the data to improving the accuracy of machine translation.

Using Machine Translation to carry out the transformation of knowledge will encounter some problems. Firstly, the content is too small to make corrections based on the context, this makes some of the translations do not conform to the original article logic and meaning. Secondly, Machine Translation relies heavily on corpus, if the various disciplines of the corpus are not perfect enough, or update is not timely, will lead to a limited corpus, thus reducing the accuracy of the results. So it is necessary to calibrate the results of knowledge conversion.

Common calibration methods include manual calibration and automatic calibration. Manual calibration is too costly and is not conducive to the calibration of large scale. Automatic calibration method based on word vector of text understanding, and based on mutual information of text correlation degree method. Based on the search engine's quality calibration method, this paper uses the method of mutual information to calculate the correlation between English and Chinese. Mutual information is a measure of the amount of information contained in a random variable [10].

Definition Mutual information is defined as follows:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)\cdot p(y)} = E_{p(x,y)} log \frac{p(x,y)}{p(x)\cdot p(y)} \quad (1)$$

P (x, y) is joint probability density function of two random variables X and Y, and p(x), p(y) represent marginal probability density function.

Pointwise Mutual Information(PMI) is commonly used to judge the correlation of two strings.

PMI is defined as follows:

$$\mathrm{PMI}(x, y) = log \frac{p(x,y)}{p(x) \cdot p(y)} \quad (2)$$

We use mutual information for the calibration process is based on one assumption, for an original three tuples <S, P, O> and transformed tuples <S', P, O'>, if the latter is correct knowledge, there will be a lot of entries that contain both S' and O' in Search results through search engines. That is to say that the higher the probability of transformation result is right, the smaller the mutual information between S' and O'. We can set a threshold to verify the transformation results.

4. Result analysis

We select yagoDBpediaClasses, yagoDBpediaInstances, yagoType and yagoGeonameClasses as experimental data. The experimental results are shown in table 4-1.

Table 4-1 experimental results

File name	File size	Numbers	Residual after calibration	Accuracy
yagoDBpediaClasses	55MB	450363	11.83%	100%
yagoDBpediaInstances	95.6MB	1144842	75.73%	96.15%
yagoGeonameClasses	3.3MB	13620	71.84%	100%
yagoType	896.2MB	18039592	94.73%	93.68%

The results show that the accuracy after the calibration can meet the expectations, but the conversion accuracy of partial knowledge conversion is low, and for some files the remaining amount is less. As shown in the figure above, *yagoDBpediaClasses* discarded nearly 90% of the content. Compared with *yagoType*, the latter is mainly in terms of words, the phrase is also 2 to 3 words, it is easy to find the corresponding Chinese data through Baidu search. On the contrary, there are more long phrases in *yagoDBpediaClasses*, these phrases are described in general, our word frequency statistics use fully match as a criterion, so it is difficult to complete a search to Chinese data, making final streamlined rate is too high. For *yagoDBpediaInstances* and *yagoGeonameClasses*, these two documents almost abandoned the 3/4, through the analysis of the original English data, it is shown that, the discarded parts of a large number of non-English characters (such as French, Russian, etc.), in the transformation process is filtered, such data although to discard a lot, but the Chinese knowledge had little effect.

Experiments show that by automatically English knowledge conversion technology, to open the English knowledge base for the source to construct Chinese knowledge base method is feasible, although some of the files of the reduction rate is very high, but the final accuracy rate is to meet the demand of the. It is also further pointed out that the improvement of the calibration algorithm to reduce the reduction rate is the most important task for the next.

5. Conclusion

In the lack of open Chinese knowledge library, from English into Chinese has better practicability, through experimental results can be see that although the streamlined process cut three nearly three quarters of the content, but the results accuracy rate after calibration has been at a high level.

According to the experiment, it is found that the problem can be in the following improvement: first of all, to further improved calibration algorithm, which are not fully matched, but the translation correct results retained; second, crowdsourcing platform is established, using the network human resource, the manual calibration again on the discarded knowledge of calibration and improve the knowledge of the capacity and accuracy.

It's a system engineering to construct a large-scale high quality Chinese knowledge base, this paper proposes transformation method provides a new way for knowledge base construction. In the following work, we will combine with the typical knowledge base construction method to form the open source large-scale Chinese knowledge base.

References

[1] Fang Bingxing. The White Paper of large web search technology[D]. China Machine Press. 2015.3

[2] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: A nucleus for a Web of open data. In: Proc. of the 6th Int'l the Semantic Web and the 2nd Asian Conf. on Asian Semantic Web Conf., ISWC 2007. Piscataway: IEEE, 2007. 722-735

[3] Wu W, Li H, Wang H, Zhu KQ. Probase: A probabilistic taxonomy for text understanding. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2012. 481-492

[4] Biega J, Kuzey E, Suchanek FM. Inside YAGO2s: A transparent information extraction architecture. In: Proc. of the 22th Int'l Conf. on World Wide Web. New York: ACM, 2013. 325-328.

[5] Hffart J, Suchanek F, Berberich K, Weikum G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia.Artificial Intelligence Journal, 2013,194(4):28-61 .

[6] Suchanek F, Kasneci G, Weikum G. YAGO—A core of semantic knowledge. In: Proc. of the 16th Int'l Conf. on World Wide Web.New York: ACM, 2007. 697-706 .

[7] Chao Le-men, Zhang Yong, Xing Chun-xiao. DBpedia and Its Typical Applications. 2011, 27(3): 80-87.

[8] Van der Zon R W L. A knowledge base approach for semantic interpretation and decomposition in concept based video retrieval[D]. TU Delft, Delft University of Technology, 2014.

[9] WEI Shuan-cheng. Crowdsourcing Concept and China Enterprise Crowdsourcing Business Model Design 10.3969/j.issn.1004-292X. 2010.01.012

[10] Thomas M. Cover / Joy A. Thomas. Elements of Information Theory. Wiley-Blackwell. 2006-7