Constructing Semantic Knowledge Base based on Wikipedia automation

Wanpeng Niu, Junting Chen, Meilin Chen

Jilin University,Changchun Jilin 130012

Keywords: The learning based on positive example, Feature Engineering, Semantic Relation, entity extraction, Hierarchical Reasoning, Wikipedia

Abstract. We know that Wikipedia is the largest knowledge set in the world, each instance entries can be a semantic entity, and it has richly hyperlinked text. Based on these, we propose a self-training method based on a small number of positive examples to extract the semantic relations and entities from the dynamic construction of semantic knowledge base. At the same time, we use TFIDF in the field of text classification and Feature Engineering in the field of computer linguistics to extract the physical characteristics of each instance and calculate their correlation. These physical features are used to help improve the accuracy and recall rate of the self-training method based on a small number of positive examples. After getting the entity right, it will be stored in the form of XML. Based on the storage structure of the XML document, a new reasoning algorithm is proposed which we called Hierarchical Reasoning. We use Wikipedia XML data in 2007 as the data test set the experimental results show that the filter based on feature selection constraint can obtain high precision and recall rate. In general, the knowledge base is built automatically. This makes it possible to extracting a large amount information from Wikipedia.

Introduction

Knowledge Base is such a tool to store structure knowledge and show the relationship between entities if needed. Today, however, the vast majority of the semantic knowledge base are all hand-built, fully rely on experts in various fields of basic experience and common sense, such as WordNet. For example, a lot of knowledge base construction is subjective, based on religious beliefs, different countries system or democratic will, subjective to change some of the information content and it will not be completely in accordance with the facts to ensure information integrity and objectivity. In the meantime, Wikipedia is a relatively weak knowledge base, and each article of it is a detailed description of an entity. But it is not clear about what is the relationship between entities, and cannot be effective reasoning. Its internal composition is similar to the shape of the tree (there is also sometimes a ring) rather than the network.

The next chapters are the arrangement of this: we introduce the related research work in the second chapter. In the third chapter, the filter based on feature selection constraint is introduced, which is based on the classification of positive examples. In the fourth chapter we introduce the storage structure of the knowledge base and the hierarchical reasoning algorithm. The fifth chapter includes the experiments, the results and the related discussion. At last, the summary and prospect are given.

Related Work

At the present stage, the extraction method of semantic relation can be divided into two categories [3]: feature vector based machine learning method [4] and kernel function based machine

learning method. Wherein the feature vector based machine learning methods need to be constructed in the form of training data, and then use a variety of machine learning algorithms to train the data. Different from the feature vector based machine learning methods, kernel function based method does not need to construct high dimension feature vector space. The feature vector based machine learning methods requires manually data, so it belongs to Supervised Machine Learning. But this kind of method requires manual intervention, so that it's not commonality.

Filtering based on Feature Felection Constraint and classification based on positive examples

1. Entity Recognition

a. Entity Feature Extraction

There exist some structured information in Wikipedia, which defines the semantic types of the corresponding entities[①]. As shown in Figure 1, we define three features that are used to represent the entity's semantic features.

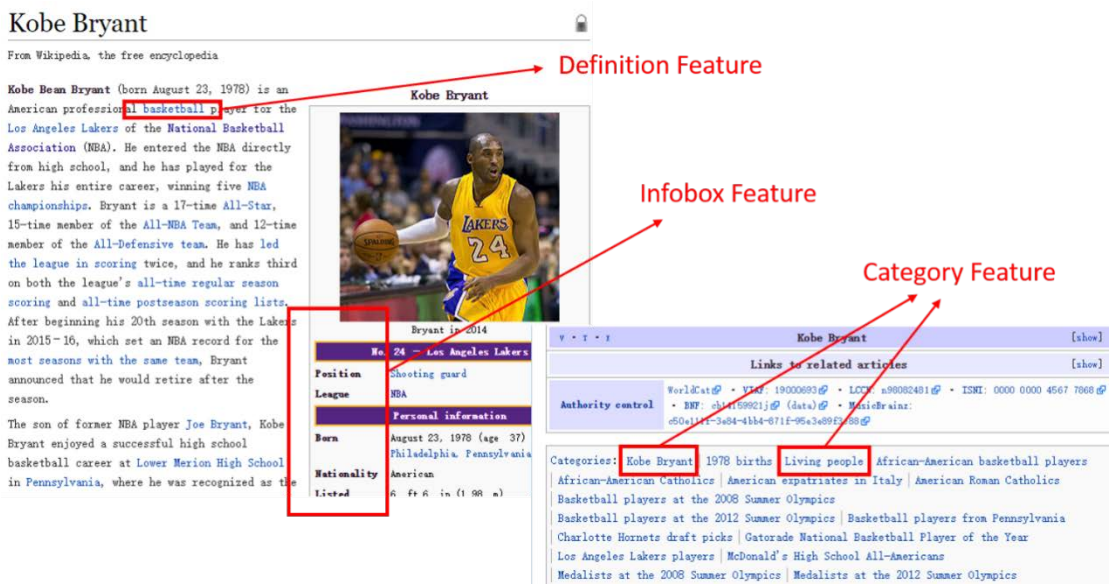

Figure 1 Entity features in Wikipedia entry

◆Definition Feature(DF)

It can be known that the first sentence in the text is often the definition of the entity on Wikipedia. Therefore, the first anchor text that appears in the sentence is extracted as the definition features.

◆Infobox Feature(IF)

Getting the attribute word from the infobox, and making them into an array of strings, such as "Position", "League", "Nationality" to form infobox features.

◆Category Feature(CF)

Because each entity in Wikipedia has a directory structure, and each directory's name is a string, as shown in figure 1. So we can extract several strings directly from the directory collection as category features.

b. Extraction of feature selection constraints

Under the constraint of the same semantic relationship, these entities have all kinds of characteristics, but they also have some common characteristics due to the constraints of the same

① In this article the entity and Wikipedia entries are one meaning

relationship. Any entity that complies with the relationship must have these characteristics. We refer to the common weighting techniques in the field of information retrieval TFIDF[9], some of the relationship sample in relation to R, using the following methods to calculate the correlation degree of physical characteristics:

In the formula:

——Entity feature

——Relational sample set under R constraints

——The complete set of all the sample instances, P is a subset of C

—— An relationship sample contained the entity features f

—— An instance of the relation of the entity feature f

For each semantic relations, calculate related degree of feature entity respectively in subject and object position, and select several entity features related to a higher degree of two sets, as the feature selection constraint of subject and object under the relationship.

2. The classification based on of positive examples

In this paper, the training sample is mainly obtained from the Wikipedia information form (Infobox). So, we only have a small amount of positive examples data to describe the relationship between the data without a corresponding negative examples. Based on this, we combine the transductive learning algorithm based on label propagate-on proposed by Chen, Ji Tan [10] and two - step way of training negative examples proposed by Yu, Zhai, Han [11], putting forward a self-training algorithm (B-POL) to obtain a sufficient number of positive and negative examples. The overall process of the system is shown in Figure 2.

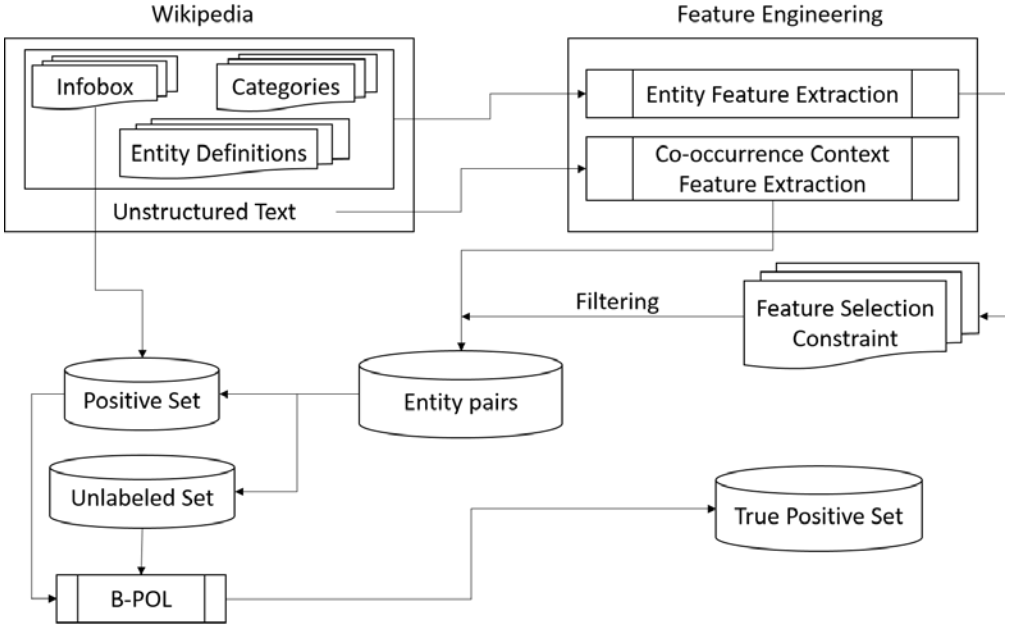

Figure 2 overall system flow chart

Storage Structure of Knowledge Base and Inference engine

1. Storage Structure

Because XML has a good information storage form and is easy to expand. Therefore, XML document is used to store the entity, the entity related semantic relations and another entity under this relationship extracted from Wikipedia. As shown in Figure 2:

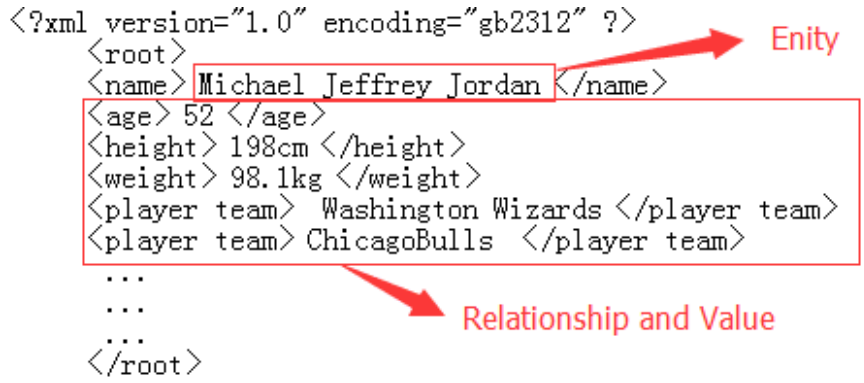

Figure 3 Internal structure of a file

We can see that the document is used to describe the main entity of **Michael Jeffrey Jordan**. In addition, to build two XML documents, one is used to store all entities and their characteristics (Entity.xml), the other one is used to store all relationships and their feature selection constraints (Relation.xml), for inference engine calls.

2. Inference Engine

The Inference engine adopted in this paper starts different reasoning mechanism based on the user input part of speech. The part of speech input by users are divided into two kinds: entity and entity, entity and relationship. Based on the knowledge storage structure we construct, a new reasoning mechanism named **Hierarchical Reasoning** is proposed.

Hierarchical Reasoning is proposed based on the user input that is mostly the entity vocabulary (rather than the relationship type). For example, cutting the user input into entity words, we can get two entities E_1 and E_2, traverse to XML documents storing entities and determine whether these two entities are included. If the match is successful, the entity information is extracted from the corresponding XML file to the two entities, and the hash table is established separately. If E_1 (E_2) matches the value of the hash table of E_2 (E_1), the relationship between the two can be obtained. Otherwise, based on all entities in the two hash tables, the entity information from the corresponding XML file is extracted to construct the second layer hash table again, and the matching action is repeated. When a value of the hash table is successfully matched to another entity, the path between the two entities is the relationship of the two.

Experiments and Results

Chapter III of this paper describes the classification method based on positive examples with the constraint of feature selection and chapter IV describes the Hierarchical Reasoning algorithm. Then the experiments are shown to demonstrate their effectiveness.

1. Filtering based on Feature Selection Constraint and classification based on positive examples

a. Experimental Setting

The data of this experiment comes from the Wikipedia XML data set in 2007[②], which is a collection of XML documents. Each XML document corresponds to a Wikipedia document. Taking into account the limited human resources, we only selected the NBA area of the corpus as the experimental data. We had a total of 6,352 documents. So, in this experiment we have 6,352 entities. By extracting the information from the infobox in these documents, we get 142 relations and 1,594

[②] https://dumps.wikimedia.org/

relational instance. After using OpenNLP[③] to split 6,352 documents, we get 10,719 entity pairs by Co-occurrence context feature engineering. In this process, we only retain the anchor text which is proper noun. In addition, we use the LibSVM[④] toolkit to implement the T-POL and B-POL algorithms. We choose RBF (Radial Basis Function) as the kernel function of SVM model. The parameter V is set to 0.01 in the v-SVM.And the other model parameters use the default parameters in LibSVM.

Considering the human factors, we randomly selected three relationships from the 142 kinds of relationships. In this process, we make as much as possible to ensure that the relationship is sufficiently accurate to demonstrate the effectiveness of the experiment. Before presenting the results, we give the following tag:

C: A collection of co-occurrence context features for all entity pairs

P： A collection of co-occurrence context features for entity pairs that conforms to a particular relationship (Positive set)

U： Unlabeled entity pairs set

a. Experimental Results

We use precision (P) and recall rate (R) to evaluate the experimental results. In many relationships, we selected three kinds of relationships: player-team, team-city, coach-team. We measured the precision and recall rate by changing the size of the positive set (P). The following table shows the accuracy and recall rate of the algorithm T-POL and B-POL under different number of training samples.

The results from the above table can be seen, B-POL did not lose too much precision while improving the recall rate. Especially in the rare positive example, recall rate was significantly improved. For analyzing advantage and disadvantage of the two algorithms, we use line chart to show the differences between B-POL and T-POL. We only select a relationship “player-team” in this paper.

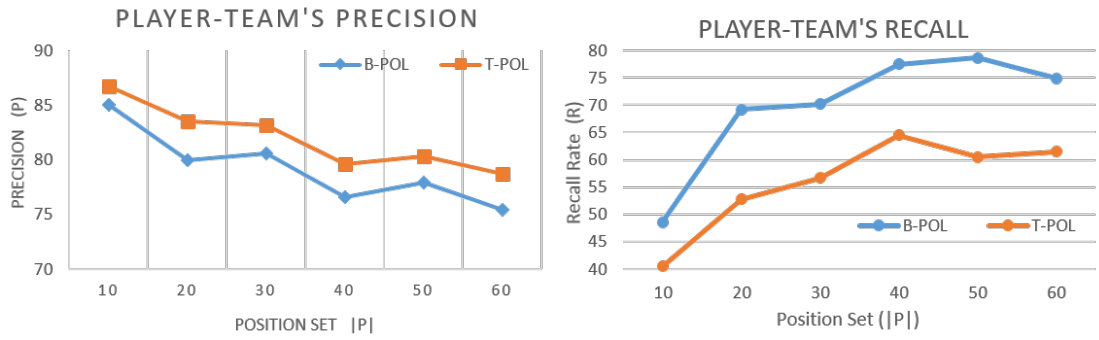

Figure 4 The recall rate and precision of “player-team” under different algorithms

It can be seen that the precision of B-POL is always less than that of T-POL in any case. Because when the training sample is relatively sparse, the positive example boundary of T-POL training is finally close to the original positive example boundary. Although the recall rate is low, but the precision is guaranteed. But when the positive examples are sparse, B-POL significantly improves the recall rate. Part of the precision of the loss is acceptable.

2. Hierarchical Reasoning

Hierarchical Reasoning is used to derive relationships between two or more entities. We

③ https://opennlp.apache.org/
④ https://www.csie.ntu.edu.tw/~cjlin/libsvm/

randomly selected 100 entities from **Entity.xml** to form 50 entity pairs, and push them to the inference engine. Because some entity pairs don't exist any strong relationships, they don't have the result of reasoning. So we abandoned the entity and finally got 34 sets of data.

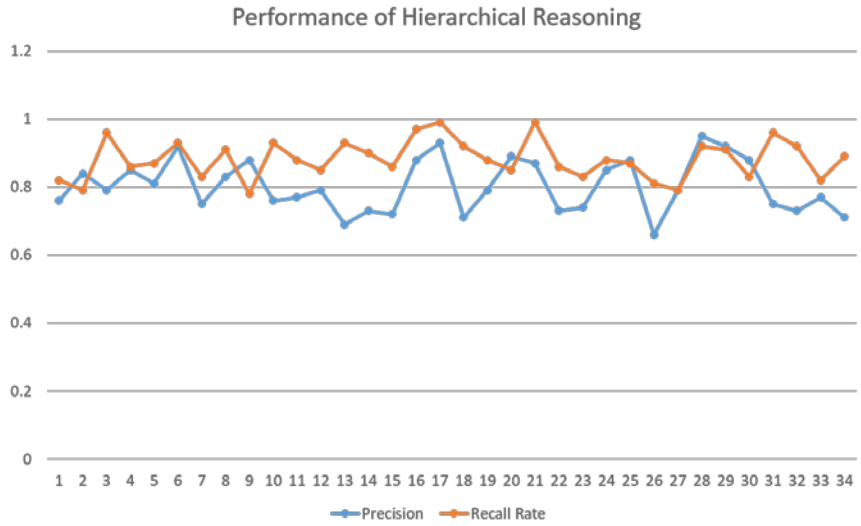

Figure 5 Performance of hierarchical reasoning

From Figure 5 we can see that the overall recall rate of the algorithm is basically stable at around 85%. Because of reasoning to the third layer by default, we ignore the weak relationship between the entities. So there is no complete to be recalled. Reasoning to the third layer is the compromise between the recall rate and the time cost. The precision is stable at 80%, which is acceptable. Because the entire system is automated to extract entities and relationships, it cannot guarantee complete accuracy. Although there are few mistakes, it can ensure maximum objectivity.

Conclusions

In this paper, we discuss an automatic extraction method based on generalized pattern matching, which is based on XML as memory structure, and we propose a new reasoning algorithm, hierarchical reasoning. And we analysis the relationship between the entity and the extraction process, a series of examples occurred in the user's retrieval, and the internal mechanism of these examples in detail. It is relatively accurate for the output result of user's different input part of speech. The knowledge of the reasoning is more complete, with strong robustness. In general, the knowledge base is built automatically, which makes it possible to carry out a large-scale relationship extraction on Wikipedia.

Acknowledgements

First thank teacher Peng Tao, who are in the busy work of teaching, still spend much time for our college students to practice the project guide. Also special thanks to the authors of the paper which we refer to, each of their inspiring research findings as well as brilliant exposition also has a profound impact on this paper.

References

[1]. Wang Gang. Automatic extraction of semantic relations in text. Wikipedia Diss. Shanghai Jiao Tong University, 2008

[2]. Li, Ding, and T. Finin. "Characterizing the Semantic Web on the Web."*Lecture Notes in Computer Science* 4(2006):242-257.

[3]. Culotta, Aron, and J. Sorensen. "Dependency tree kernels for relation extraction." *Proceedings of Annual Meeting on Association for Computational Linguistics* (2004):423--429.

[4]. Banko, Michele, and O. Etzioni. "T.: The tradeoffs between open and traditional relation extraction." *Proceedings of Acl* (2008):28--36.

[5]. Hasegawa, Takaaki, S. Sekine, and R. Grishman. "Discovering Relations among Named Entities from Large Corpora." *Annual Meeting of the Association for Computational Linguistics* (2004).

[6]. Peng, Tao, W. Zuo, and F. He. "SVM based adaptive learning method for text classification from positive and unlabeled documents.."*Knowledge & Information Systems* 16.3(2008):281-301.

[7]. Peng, Tao, L. Liu, and W. Zuo. "PU text classification enhanced by term frequency–inverse document frequency-improved weighting. "Concurrency *& Computation Practice & Experience* 26(2014):728-741.

[8]. Chen, Jinxiu, et al. "Relation Extraction Using Label Propagation Based Semi-Supervised Learning..." *International Conference on Computational Linguistics & Meeting of the Association for Computational Linguistics*2006:129-136.

[9]. Yu, Hwanjo, C. X. Zhai, and J. Han. "Text classification from positive and unlabeled documents..." *In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03*2003:232--239.

[10]. Guodong, Zhou, et al. "Exploring Various Knowledge in Relation Extraction..." *Acl'–30 Ann Arbor Michgan Usa* (2005):419--444.

[11]. Agichtein, Eugene, and L. Gravano. "Extracting Relations from Large Plain-Text Collections." *In Proc. of the 5 th ACM International Conference on Digital Libraries (ACMDL'00* 2000:85-94.

[12]. Pelckmans, Kristiaan, and J. A. K. Suykens. "Transductively Learning from Positive Examples Only..." *In Proc. of the 16th European Symposium on Artificial Neural Networks (ESANN09* 2009.

[13]. Li, Xiaoli, and B. Liu. "Learning to Classify Texts Using Positive and Unlabeled Data..." *International Joint Conference on Artificial Intelligence*2003:587--594.

[14]. Zhu, Xiaojin. "Semi-Supervised Learning Literature Survey." *Computer Science* 37.1(2008):63-77.

[15]. Strube, Michael, and S. P. Ponzetto. "Wikirelate! Computing semantic relatedness using wikipedia." *National Conference on Artificial Intelligence-volume* 2006:1419--1424.

[16]. Leacock, C., and M. Chodorow. "WordNet: An electronic lexical database", Combining local context and WordNet similarity for word sense identification." *Wordnet an Electronic Lexical Database* (1998).