A Novel Online Encyclopedia-Oriented Approach for Large-Scale Knowledge Base Construction

Ting Wang
College of Computer Science and Technology, Beijing University of Technology, Beijing, China
Email: basten0926@163.com

Ruihua Di, Jicheng Song
College of Computer Science and Technology, Beijing University of Technology, Beijing, China
Email: drh@bjut.edu.cn, sjc@emails.bjut.edu.cn

***Abstract*—In the process of constructing large-scale knowledge base, manual-based construction approach lacks efficiency as well as flexibility. Therefore, automatically extracting of massive knowledge from online encyclopedia has attracted attention from an increasing number of scholars. Current research is mainly focused on the extracting of data from English online encyclopedia, whereas research about knowledge extraction from Chinese or other language data sources is rare. For such reason, the present paper proposes an automatic construction scheme for large-scale knowledge base based on Chinese online Encyclopedia. (i)In the first phase of the scheme, self-expanded learning is performed on the semantic relations between subjects and objects among the knowledge triples. (ii)In the second phase, semantic relations between the marked attributes and their entities is predicted using Conditional Random Fields (CRFs) and Support vector machine (SVM) classifier. A large-scale knowledge base is automatically constructed based on the scheme, and the experiment results indicate that the scheme possesses feasibility and effectiveness.**

***Index Terms*—knowledge base, online encyclopedia, knowledge discovery, conditional random fields, support vector machine**

I. Introduction

The vision of the Semantic Web is to create a "web of data", so that a machine is able to understand the semantic information on the internet [1]. In order to achieve the goal of Semantic Web, many semantic data sets in different areas are being published on the internet, and the interlinking between them has been established.

Wikipedia is a Web-based open encyclopedia, which has become very popular as new information resources [2]. Since Wikipedia has rich vocabulary, good updatability, and semi-structuredness, there is less difference between Wikipedia and ontologies when compared with free text. Thus, ontology learning from Wikipedia becomes popular.

Corresponding author: Ting Wang, E-mail:basten0926@163.com

doi:10.4304/jsw.9.2.482-489

However, due to cross-language barrier, the number of Chinese vocabulary entries on Wikipedia is only ten percent of English entries. Accordingly, ontology model of DBpedia is English-based, thereby making it hard for Chinese-based knowledge base to contribute in the process of constructing linking open data.

In China, there are also several Chinese online encyclopedias, and the knowledge base construction also attracts more and more scholars' attention. However, at present there is very little contribution concerning building a web scale semantic knowledge base like YAGO in Chinese language, especially constructed in a near automatic fashion with less human involvement.

Therefore, this paper is focused on the exploiting of a construction method to build a large-scale knowledge base based on Chinese online encyclopedia. The present paper is organized into eight sections. First section gives an introduction. Section 2 introduces representative work regarding knowledge base construction based on web encyclopedia. Section 3 presents the overall architecture for the knowledge extraction system which acquires data and knowledge from Chinese online encyclopedia. From section 4 to section 6, detailed analysis and illustration about the implementation of key modules for the system are covered. In section 7, an evaluation index system is proposed and the experiment results of the knowledge extraction system are analyzed and discussed. The last section gives conclusions and points out the possible areas for improvement in the system for future researches.

II. Related Work

With the development of Web 2.0, the collaborative creation in specific or open domain has appeared in the online encyclopedia such as Wikipedia, Baidu Baike[3], Soso Baike[4] and Hudong Baike[5]. YAGO[6] and DBpedia[7] are based on Wikipedia to build the large-scale semantic knowledge base. They extract knowledge from structured information of Wikipedia pages. With the "is-a" relationship between the concept of hierarchical relationships in the taxonomy system, the Infobox

information on the entry's page contains some knowledge <*S,P,O*> triples. Wu and S.Weld. developed Kylin[8][9] system which not only is used for structured content, but also attempts to extract knowledge triples from unstructured text in Wikipedia's articles.

DBpedia is proven to be a successful structured knowledge base. Serving as a Hub that links data, it has connected semantic data from different fields and forms a large-scale web of data. However, most of the semantic data is expressed in English while research and publishing linked data in Chinese is still in its infancy.

Many scholars and institutions have made efforts to build domain ontology with clear semantic structure, and rich relationship between entities[10][11].

Zhichun Wang and other scholars put forward an idea of retrieving hierarchical relationships between concepts based on the classification system of Chinese encyclopedia, acquiring the conception attribute and Wikipedia entry instances on entry web pages embedded with Infobox, and then establishing the Chinese encyclopedia knowledge base to form a co-reference with DBpedia with the simple keyword matching strategy[12]. However, their idea only concentrates on the semi-structured information in the Chinese online encyclopedia system and doesn't cover the work of extracting the knowledge triples from massive unstructured plain-text information.

Yidong Chen and other scholars proposed to use the attribute values from Infobox on Chinese encyclopedia to automatically extract the training sample and then acquire massive knowledge triples from the unstructured text with statistical learning model[13]. However, the amount of words used that can reveal the semantic relations between subjects and objects is deficient and the strategy to automatically generate the training sample is also flawed, causing low precision and recall ratios.

In allusion to the problems mentioned above, this paper proposes a semi-supervised learning method, which uses the self-expansion algorithm for semantic relations and collaborative classifiers based on SVM and CRF to automatically extract the knowledge triples from Chinese online encyclopedia.

III. Preliminary

This section gives brief introductions of the Conditional Random Fields and Support Vector Machine.

A. Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) are commonly used to perform lexical analysis involving Chinese word segmentation and part-of-speech tagging. Without the strict independence assumption under which Hidden Markov Models are built, CRFs possess several advantages. They are capable of expressing the characteristic of long-distance dependence and overlapping, solving the problem of tagging bias and normalizing all the characteristics to obtain the global optimal solution. On the other hand, the entity recognition problem is actually about sequence tagging, or in other words, judging whether the observed words belong to the predefined feature set, which is exactly the advantage of CRFs. Undoubtedly, CRFs model is well suited to named entity recognition [14].

For the named entity recognition, a CRFs model is defined as equation (1) for a given word sequence $x=(x_1,x_2,\ldots,x_n)$ and tagging sequence $y=(y_1,y_2,..,y_n)$:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, x)\right) \quad (1)$$

where $Z(x)$ is the normalizing factor, n is the length of the give word sequence, $f_k(y_{i-1},y_i,x)$ is the characteristic function which can express not only the transferring characteristics $e(y_{i-1},y_i,x)$ of the undirected graph , but also the status characteristics $v(y_i,x,i)$ of the nodes and λ_k is the weight coefficient for the Kth characteristic function.

B. Support Vector Machine (SVM)

The implementation of SVM is performed by mapping the input vectors to a high dimensional feature space through a pre-selected nonlinear mapping (kernel function) and then constructing the optimal separating hyper-plane in that space[15].

Classification margin from a sample point to the hyper-plane is defined as $\delta_i=y_i(wx_i+b)$. After normalization, w and b can be replaced with $w/\|w\|$ and $b/\|w\|$, and the margin is converted into $\delta_i=(1/\|w\|)*|g(x_i)|$, which is called geometry margin which reflects the distance from points to the hyper-plane. Since geometry margin is inversely proportional to $\|w\|$, identifying the minimum $\|w\|$ is the solution to the optimal separating hyper-plane.

IV. Architecture

Extraction system for knowledge triples from Chinese online encyclopedia is shown as the figure 1.

Through the system, knowledge triples can be automatically extracted from entry web pages that don't include Infobox. This system can automatically retrieve training sample from pages with Infobox based on the structured information provided by Infobox, and then retrieve structured knowledge from the plain-text information on pages without Infobox.

This system is mainly composed of the following modules:

(1) **Traning Data Creation:** This module is used to automatically acquire training samples from entry pages with Infobox for CRF and SVM models and is also used to extract attribute values from unstructured text.

(2) **Semantic Relations Extraction:** This module is used to perform self-expansion learning on semantic relations between subjects and objects. Through the proposed self-expansion algorithm, the not all-around semantic relations in "Chinese Thesaurus" can be enriched, which will then improve the performance of the extraction classifier for knowledge triples.

(3) **NER Process for Attribute-Value Pair:** By

adjusting the extraction classifier for knowledge triples, this module is able to extract attributes and attribute values from plain-text information on entry web pages that don't have Infobox. The identification of named entities is implemented based on CRF.

(4) **Relation Prediction between Attribute and Value:** Function of this module is to predict and determine the potential semantic relations between the acquired attributes and attribute values by extraction classifier for knowledge triples. As for the prediction of relations between entities is implemented based on SVM model.

Detailed illustration will be given to each of the four mentioned modules.

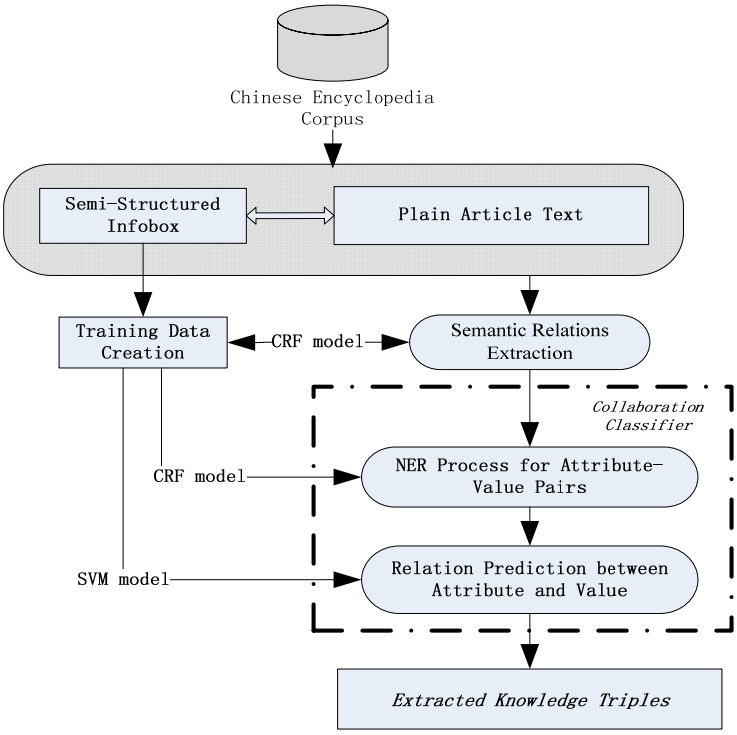


Fig. 1. Architecture of Knowledge Triple Extraction System

V. Training Data Creation

A. Structured Knowledge Extraction from Infobox

Infobox appearing on entry's web page can provide various semantic knowledge; that is, Infobox implies a lot of triples by taking the term name as the subject, attribute name as the predicate, and attribute value as the object. Therefore, well-structured knowledge can be extracted from the Infobox automatically. For example, the fact that "Beijing is the capital of China" is represented in the Infobox of the term:"China" as "Capital: Beijing". Hence, a knowledge triple as "<China, Capital, Beijing>" can be extracted directly. After crawling all the wiki pages with Infoboxes in Hudong encyclopedia, over 5.8 million knowledge triples represented in each Infobox can be resolved. All of them are described in Chinese literals.

B. Candidate Sentence Extraction

Infobox knowledge triples extracted from each wiki entry pages are used to automatically acquire the well-structured training sample. Based on the common characteristics of each word, information described by Infobox knowledge triples can be positioned to a certain sentence in the original text. For example, for triple "Mao Zedong, Birthplace, Hunan-China", if the subject, predicate and object of the knowledge triple all appear in the same sentence such as "Mao Zedong was born in a farmer's family in Shaoshanchong, Hunan, China", then this sentence is considered as the best candidate sentence for training samples.

Since elements of the triples don't always appear in the same sentence, Precedence Chart[14] is introduced to give a quantitative calculation on the importance weights of subject, predicate and object. In the end, all the sentences on the entry pages with Infobox can be graded.

Importance of the three elements of the knowledge triple is defined in the following way: objects(o) have higher importance than subjects(s), which then have higher importance than predicates (p). According to formula (2) of the Precedence Chart algorithm, the calculated weight for each element is $o=0.6$, $s=0.3$, and $p=0.1$. All three elements are compared in pairs. If element i is more important than element k, then $\alpha_{ik}=1$; if they are equally important, then $\alpha_{ik}=0.5$; otherwise, $\alpha_{ik}=0$.

$$\lambda_i = \frac{\sum_{k=1}^{n} \alpha_{ik}}{n(n-1)/2 + 0.5n} \tag{2}$$

The scoring formula for the corresponding sentence is shown as equation (3):

$$Score = s \times Sub + p \times Pre + o \times Obj \quad (3)$$

Assumption: If the subject in a particular entry triple in an Infobox appears in a sentence from plain texts on the web page for that particular entry, then the *Sub* equals to 1; otherwise, it is 0. For the same reason, if the predicate (or its semantic relation word) appears in this sentence, then the *Pre* is 1, otherwise, it is 0; if the subject appears, the *Obj* equals to 1; otherwise, it is 0.

Specifically, for a certain predicate p_i, the sentence set with its weight being 0.9 will be used as training samples for Semantic Relations Extraction module. Whereas the sentence set with weight being 1 will be used as training samples for NER Process for Attribute-Value Pair module.

C. Semantic Relations Extraction

Because of the openness and informality of online encyclopedia, a large amount of words which can reflect the semantic relations between subjects and objects cannot be obtained from the HIT's "Tongyici Cilin"(extended version) [17], whereas entry pages include Infobox only account for 10% of the total entries[10]. As a result, based on the idea of semi-supervised learning, semantic words will be extracted from all web pages with Infobox, and the newly extracted predicates will be added to the semantic word set.

To solve this problem, the paper proposes the self-expansion semantic relation extraction algorithm to automatically retrieve semantic relation and improve the robustness of the training sample.

Self-expansion algorithm is a progressive learning method which only requires small amount of information or seed words as base to continuously study so as to achieve effective expansion of the original semantic relation vocabulary. With the algorithm, semantic words that reflect the relations between subjects and objects in sentences can be automatically acquired.

Specifically, words of the same kind always have the same context [18], that is, if subjects and objects are often surrounded by certain words, then these words can be used to interpret the semantic relations between subjects and objects. This paper uses the semi-structured information from the Infobox on the online encyclopedia entry pages to automatically acquire the training samples to be used by the extended classifier for the semantic relation seeds of a certain predicate.

The proposed self-expansion algorithm is able to acquire the semantic words which can reflect the relations between entities. First of all, all synonyms of the predicate P_i are extracted from the HIT's "Tongyici Cilin"(extended version) [17] to construct the initial seed set. After that, from the text on the wiki pages with Infobox, training samples can be automatically acquired to be used for the expansion of semantic relations. Iterative training is then performed using CRF statistical learning model and the newly acquired semantic words are filtered and then added to the seed set R_i of predicate P_i based using TF-IDF algorithm, thereby improving the robustness of extraction classifier for attribute values. The specific algorithm is illustrated as follows:

Input: seed set R_i of semantic words for the initial predicate P_i and the training sample sentence set T_i which is obtained from R_i.

Output: Expanded semantic word set R_i' and expanded training sample sentence set T_i'

Step 1: Set $R_i' = R_i$ and $T_i' = T_i$.
Step 2: Sentence set that is extracted based on seed set R_i' with weight value as *1* is used as training sample T_i'. Sentences with current weight value as *0.9* are used as test sample. Semantic words are tagged based on CRF model.
Step 3: All the new semantic words that are obtained from each tagging are weighted using TF-EDF and sorted.
Step 4: TOP 2 of the sorted words are added to the semantic word seed set R_i'.
Step 5: Determine whether R_i' is equal to R_i. If yes, then the algorithm is terminated. If not, step (6) is needed.
Step 6: The new seed set is substituted in equation (2) to obtain the new sentence set with weight value as 1, which will be added to the training sample set T_i'.
Step 7: Return to step 2.

VI. Online Encyclopedia Knowledge Extraction based on CRF-SVM Fusion

Currently, research in named entity recognition by most scholars is focused on sentence level, which inevitably results in errors in acquiring sample sentences. In other words, manual extraction of sentences requires the involvement of filed experts and is time-consuming, which makes it hard for large-scale extraction. As for automatic extraction of test samples, the generated errors will cause disturbance for identity of named entity. Previous researches only cover the extraction of attribute values and fail to consider whether there is semantic relation between the attribute values tagged by classifier and predicate from Infobox, which can then cause the subject deviation problem. In addition, their strategy of automatically extracting test sample sentences is flawed, resulting low precision and recall rate.

With a little more calculation, combination of multiple classifiers can generate a higher ratio of correct classification than a single classifier because different classifiers can be complementary to each other. For a particular feature, different classifiers have different performances. As a result, choosing the suitable classifier for different features and combining them can facilitate the improving of the accuracy of final decision. Experiments in various fields indicate that, combination

of multiple classifiers can indeed yield higher recognition rate and improve performance and stability. Suen and Kittler put forward the concept and theory of classifier fusion respectively.

To avoid the problems mentioned above, the present paper proposes an extraction method that is implemented in paragraph level instead of in sentence level. As for subject deviation, SVM classifier is used to eliminate the falsely identified entities. Also in the present paper, an approach based on CRF and SVM is put forward together with the classifier fusion so as to improve the precision of the system.

Predicates and objects in the plain-text message are considered as entities in the present paper so that the attributes and attribute values in the knowledge triples can be extracted and marked based on the idea of classification and named entity recognition. Consequently, corresponding relations between entities and attributes can then be predicated.

First, attributes and attribute values in the knowledge triples are deemed as entities, and CRF is then used to perform entity recognition on training corpus modeling. However, the two kinds of entities after recognition bear no relations at all. Therefore, the next step is to predict the relations between entities.

Vector is marked as 1 if relationship between attributes and attribute values exists; otherwise, vector is marked as 0. SVM is then used to construct a model to predict relations in pairs. With a complete paragraph from the Chinese plain-text entries in Hudong Baike as a unit size, experiment is conducted to predict the relationship between named entities that are paragraph level. Results of experiment indicate that the method possesses feasibility.

A. Entity Recognition of Attributes and Attribute Values based on CRF

1) ***Corpus Mark Conversion:*** According to the recognition task for attributes and attribute values of the knowledge triples, several mark sets are defined as *L=(B-ATTR, I-ATTR, B-OBJECT, I-OBJECT, O)*, which refer to beginning of semantic relation words, interior of semantic relation words, beginning of attribute values, interior of attribute values and others respectively.

2) ***Feature Selection and Design for Feature Template:*** For entity recognition using machine learning method under CRF, the most important step is to select the entity features and design the feature template, which is defined to filter features. In the process of formulating the feature template, single, dual and multi features are combined. In the experiments, selected features for the attributes and attribute values of knowledge triples include: the word itself, part-of-speech, combination feature and context information.

According to the plain-text corpus from Chinese encyclopedia and the results of multiple experiments, when the word, part-of-speech and multi-combination of word and part-of-speech are selected as features, optimal experiments results can be obtained. An example of feature template for CRF is shown as table 1.

TABLE 1.

FEATURE TEMPLATE FOR CRF CLASSIFICATION

Feature Set Number	Template item	Indication
1	U01:%x[-1,0]	One row above current row, column 0 (word)
2	U02:%x[0,0]	Current row, column 0 (word)
3	U03:%x[-1,1]	One row above current row, column 1 (part-of-speech)
4	U04:%x[0,1]	Current row, column 1 (part-of-speech)
5	U05:%x[0,0]/%x[1,0]	Current row, and one row below, column 0 (word, word)
6	U06:%x[0,1]/%x[1,1]	Current row column 1,next row column 1 (part-of-speech, part-of-speech)
7	U07:%x[-1,0]/ %x[0,0]/%x[1,0]	One row above current row, current row, next row column 0 (word, word, word)

3) ***Model Training and Testing:*** In the process of CRFs model training, one commonly used training method is iterative gradient method, such as GIS and IIS method, which are easy to be implemented but have a shortcoming of slow convergence. Wallach [19] used the combination of slope change method and second-moment method for model training and achieved satisfactory results. Therefore, the present paper adopts the method proposed by Wallach, and CRF++ toolkit is used in the system.

For the certain Infobox predicate P_i, the sentence set in which sentence weight is *1* will be used as the training sample for entity recognition of P_i's corresponding attribute words and attribute values. Paragraphs from plain-text descriptions on Infobox-free entry pages that are under the sub-category of a certain encyclopedia category will used for entity recognition of attribute words and attribute values in knowledge triples.

B. Prediction of Entity Relations between Attributes and Attribute Values based on SVM

1) **Feature Selection for SVM**: Based on the analysis of sentence structure of training corpus, part-of-speech of the number of k words that exist in the left, center and right of attributes and attribute value words play a good role in distinguishing the relations between them. In addition, the order and the distance in which attributes and attribute values appear are also the key factors in determining their relations. Due to the diversity

of Chinese description, literal values of words can hardly be unified to small dimensions. Therefore, when determining the training feature vector for SVM, words per se will not be considered as attribute features. An example of SVM feature model is shown as table 2.

TABLE 2.
FEATURE SELECTION FOR SVM CLASSIFICATION

Feature Set Number	Selected Feature
1	Part-of speech of the entity pair and the two words before and after them
2	Order in which the entities in the entity pair appears and the distance between them

2) ***Model Training and Testing:*** During the process of SVM model training, a satisfactory relation prediction result is obtained via effective feature selection and adjusting of training parameters by using a mixed function with strong learning and. In this system, libsvm toolkit is used for SVM model.

For the certain Infobox predicate P_i, the sentence set in which sentence weight is 1 will be used as the training sample to predicate the relations between entities, i.e. the obtained training sample set R_i' in 5.2 will also be used in this process. Since the predicates and objects are already marked, the only step left is to vectorize the sentences in the sentence set according to the given feature set from SVM classification, i.e. format conversion.

In the meanwhile, for the paragraphs from plain-text description on encyclopedia entries where Infobox is not involved, the context information of attributes words and attribute values words that are tagged by classifier of predicate P_i also needs to be converted based on SVM feature set.

VII. EXPERIMENTS

In the experiment, the system firstly adopts the word segmentation software ICTCLAS50 [20] developed by Institute of Computing Technology, Chinese Academy of Sciences, so as to complete word segmentation and part-of-speech tagging on plain-text information from encyclopedia entry pages and then uses CRF++[21] and libsvm[22] toolkit to perform recognition of named identities and construction of classifier for entity relation prediction.

In order to compare with the experiment results in reference [13], category “Country”, “People” and sub-category “Actors” are selected from Hudong Baike as the source for entry page instances in this experiment. Entry pages under these categories will be fetched and those don’t have Infobox are used as final test sample for attribute values extraction.

Currently, researches regarding knowledge extraction from Chinese online encyclopedia are relatively less. As a result, no authoritative corpus is available for conducting this research and that’s the reason for what the present paper constructs the Chinese open encyclopedia corpus.

In the experiment, plain-text messages from 300 entry pages under each of the three categories mentioned above are randomly selected as test corpus, which include 900 entry pages in total. This step fully utilizes the redundancy of Web information to effectively overcome the data sparsity problem. After that, attributes and attribute values of knowledge triples are tagged for the calculation of precision rate and recall rate.

A. Evaluation Criteria

The present paper uses precision and recall rate on attribute values recognition, and F-measure as the final evaluation criteria:

$$Precision(P) = \frac{A}{B} \times 100\%$$

$$Recall(R) = \frac{A}{C} \times 100\%$$

$$F - measure(F1) = \frac{2 \times P \times R}{P + R} \times 100\%$$

A: represent the number of entities with correctly marked attribute values
B: represent the number of entities in the recognition results
C: represent total number of entities in the standard result

B. Results and Analysis

Experiment results in table 3 are compared and analyzed. Extracted results indicate that the CRF-model-based attribute values extraction experiment processed with semantic relation self-expansion algorithm is able to cause about the highest 4% increase in Precision and 6% increase in Recall. The main reason is that context feature of the extracted predicate synonymy is strengthened, which means the number of recognized entities and correctly recognized entities have increased, thereby improving the value of *P* and *R*. Increase in precision, on the other hand, is lees than the increase in recall rate.

After the processing of collaborative classifier, number of recognized entities is decreased, leading to a higher precision rate. This is because false entities are eliminated from the attribute value results set, which causes value of *P* to rise while the value of *R* stays basically the same or experiences an decrease of less than 1%) since some of the correctly recognized entities may have been eliminated as well by mistake. As far as the task—automatically acquiring knowledge triples—is concerned, higher precision rate is the expected target. Also shown in the experiment results, knowledge extraction method based on collaborative classifier can increase value of *F1*, that is, the overall system can benefit from the proposed approach.

TABLE 3.

KNOWLEDGE TRIPLE EXTRACTION RESULTS FROM PLAIN ARTICLES

Category	Infobox Property	Reference[11]			This Paper (Semantic Relation Self-Expansion)					
					CRFBased			Collaborative Classifier		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Person	Birthday	86.4%	73.1%	79.2%	89.7%	79.4%	84.2%	90.3%	79.1%	84.3%
	Birthplace	75.5%	58.7%	66.0%	77.8%	65.1%	70.9%	83.9%	64.4%	72.9%
	Graduated University	76.4%	75.0%	75.7%	79.4%	76.4%	77.9%	85.7%	75.6%	80.3%
Country	Official Language	86.9%	81.1%	83.9%	80.5%	75.2%	77.8%	84.4%	75.0%	79.4%
	Area	70.8%	60.7%	65.4%	73.0%	62.5%	67.3%	75.8%	62.2%	68.3%
	Capital	79.0%	53.6%	63.9%	80.3%	59.9%	68.6%	82.8%	59.9%	69.5%
Actor	Masterpiece	90.8%	65.0%	75.8%	90.1%	66.2%	76.3%	91.0%	65.3%	76.0%
Avg F1				72.8%			74.7%			75.8%

VIII. CONCLUSIONS AND FUTURE WORK

The present paper proposes an approach to automatically tag training corpus to decrease manual work. The proposed approach can select feature set based on the relations between entities in plain-text descriptions from Chinese online encyclopedia and then use the self-expansion algorithm to automatically acquire semantic words that can reflect those relations and add them to the feature set. Ultimately, knowledge triples extraction from unstructured plain-text messages on Chinese online encyclopedia can be achieved with the constructed collaborative classifier.

Focus of future work is to build links between this knowledge base and other data sets such as the E-government domain so as to publish Chinese linked data. Furthermore, the interlinking of Chinese knowledge with DBpedia or other knowledge bases which is described in different languages can enhance the heterogeneous knowledge sharing of different languages.

Furthermore, because of the openness on encyclopedia taxonomy; therefore, many instances and Infobox's triples have been classified improperly; in the future we will find the appropriate clustering algorithm to classify these triples accurately, so as to optimize the knowledge extraction schema.

ACKNOWLEDGMENTS.

This work was supported in part by a grant from National Natural Science Foundation of China (No. 61202075) and IBM Shared University Research (SUR) Project (E-government Data Semantic Integration and Knowledge Discovery on the Cloud).

REFERENCES

[1] Berners-Lee, T., Hendler, J., et al: The Semantic Web. Scientific American (2001)
[2] http://www.wikipedia.org/
[3] http://baike.baidu.com/
[4] http://baike.soso.com/
[5] http://www.hudong.com/
[6] Suchanek, F.M., Kasneci G., et al. YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Sem. 6(3), 203–217 (2008)
[7] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., et al: DBpedia - A Crystallization Point for the Web of Data. Journal of Web Semantics 7(3), 154–165 (2009)
[8] Wu, F., Weld, D.S. Automatically Refining the Wikipedia Infobox Ontology. In: Proceeding of 17th International Conference on World Wide Web, pp. 635-644. ACM (2008)
[9] Wu, F., Weld, D.S. Autonomously Semantifying Wikipedia. In: Proceedings of the 16th ACM International Conference on Information and Knowledge Management, pp. 41–50. ACM Press, New York (2007)
[10] Jie Liu, Yun Ma. Reuse of Chinese domain ontology for the restricted domain question answering system[J]. Journal of Computers(Finland), v7, 2012:2684-2690
[11] Jike Ge, Zushu Li, Taifu Li. A novel chinese domain ontology construction method for petroleum exploration information[J]. Journal of Computers(Finland), v7, 2012:1445-1452
[12] Z. Wang, Z. Wang, J. Li et al. Knowledge extraction from Chinese Wiki Encyclopedias[J]. Journal of Zhejiang University - Science C, vol 13, no. 4, pp. 268–280, 2012.
[13] Chen Yidong, Chen Liwei, Xu Kun. Learning Chinese Entity Attributes from Online Encyclopedia[C]. APWeb 2012:179-186
[14] Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. 2001.
[15] Cortes C, Vapnik V. Support Vector Machine[J]. Machine learning, 1995, 20(3): 273-297.
[16] Liu, Wenyuan., Wu, Lixia., Wang, Baowen. Multidimensional Sparse Fuzzy Reasoning Method Based on Weight of Precedence Chart[J]. Journal of Computer Engineering, 2009:210-215
[17] Mei, J.J., et al: Tongyici Cilin. Shanghai Lexicographical Publishing House (1983)
[18] Bai, Lei. Computer-Assisted Discovery on Language Knowledge (In Chinese)[M]. Science Press, 1995.
[19] Wallach H. Efficient Training of Conditional Random

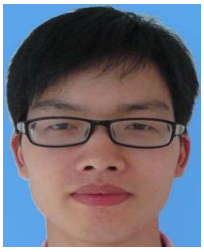
Fields[D]. Master's thesis, University of Edinburgh, 2002.
[20] http://ictclas.org/
[21] http://crfpp.sourceforge.net/
[22] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Ting Wang was born in Beijing, China, in 1985. He received the B.Eng degree in computer science from Beijing University of Technology, China, in 2008. His research interests include the semantic web, distributed computing and cloud computing.

Now he is the Ph.D Student at College of Computer Science and Technology, Beijing University of Technology, Beijing, China.

Ruihua Di was born in Hebei Province, China, in 1947. She received the B.Eng degree in electrical engineering from Beijing University of Technology in 1970. Her research interests include the distributed computing, grid computing and cloud computing.

Now she is a professor at College of Computer Science and Technology, Beijing University of Technology, Beijing, China.



Jicheng Song was born in Henan Province, China, in 1987. He received the B.Eng degree from Jilin University in 2010. His research interests include the semantic web and high performance computing.

Now he is the M.Eng Student at College of Computer Science and Technology, Beijing University of Technology, Beijing, China.