ADAPTIVE NAMED ENTITY RECOGNITION BASED ON CONDITIONAL RANDOM FIELDS WITH AUTOMATIC UPDATED DYNAMIC GAZETTEERS

Xixin Wu[1,2], *Zhiyong Wu*[1,2], *Jia Jia*[2], *Lianhong Cai*[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
xixinwood@gmail.com, zywu@sz.tsinghua.edu.cn, {jjia, clh-dcs}@tsinghua.edu.cn

ABSTRACT

This paper presents a hybrid model which combines conditional random fields (CRFs) with dynamic gazetteers (DGs) for the task of Chinese named entity recognition (NER). In the previous work of NER, gazetteers were widely used. But their gazetteers were all static ones which cannot adapt themselves to the new domains and new out-of-vocabulary named entities (OOVNEs). In this work, we build and maintain DGs to solve the problems and propose a method to automatically update DGs along with the recognition process of the named entities (NEs). With this method, the DGs can be updated to contain more and more new NEs and features of NEs that are not found in the training data. These newly added items make the DGs become more aware of the knowledge about new domains and hence be more adaptive to new domains for the recognition of OOVNEs. Experiments on the People's Daily corpus demonstrate that our method is effective, and can improve the average F-score by 1%~2%.

Index Terms**—** Named entity recognition (NER), Dynamic gazetteers (DGs), Conditional random fields (CRFs)

1. INTRODUCTION

Named entity recognition (NER) is one of the essential problems in natural language processing (NLP) related researches, such as information extraction (IE), information retrieval (IR), machine translation (MT), as well as general domain text-to-speech (TTS) synthesis. It aims to classify every word in a document into some predefined categories or "not-a-named-entity". The categories consist of person name, organization name, location, currency, date, time, monetary and percentage

In TTS, two major problems affecting the performance of Chinese text processing are the word segmentation ambiguities and the OOV words. Studies [1][2] have shown that the influence of OOV words is 5 times more serious than that of word segmentation ambiguities and most of OOV words are NEs. Furthermore, if a word is a NE (i.e. name of person, organization or location) is an important clue for grapheme-to-phoneme conversion. For example, "单"(single) is pronounced as "dan1" when it serves as an adjective in sentences, while when it serves as a surname, it should be pronounced as "shan4". It's essential to perform NER for text processing in TTS synthesis. In Chinese, currency, date, time, monetary and percentage can be easily recognized by grammatical rules. According to the requirement of TTS system and the characteristic of Chinese, we focus on NER for person name, organization name and location.

The corpora from different domains contain quite different sets of named entities and the features of NEs. The features of NEs include both internal and external features. The internal features describe the characteristics of NEs, e.g. the prefix or suffix of NE. The external features are related to the context information of NEs, e.g. the character directly before or after NE. Table 1 illustrates the difference between documents in sport and politics domains, where percentage (%) denotes the appearance frequency of a suffix in the domain and is obtained from the training data of our corpus. As can be seen, the frequency of the suffix "队(team)" is as high as 54.06% in sport domain, while it is only 0.43% in politics domain. Since NEs in different domains have different features, it is necessary to consider features separately in different domains.

Table.1. Difference between documents in different domains

Domain	Named Entity	Suffix (%)

Sport	上海东方队(Shanghai Dongfang Team), 巴西足协 (Brazilian football association)	队(Team) (54.06)
		政府(Government) (0.72)

Politics	美国国会(United States Congress), 国务院(State Council)	政府(Government) (6.59)
		队(Team) (0.43)

Early methods for NER are roughly rule based by virtue of grammar-based techniques. Development of appropriate rules is very time consuming. Recent studies of NER have been focusing on machine learning methods, such as hidden Markov model (HMM) [3], support vector machine (SVMs) [4], CRFs [5], etc. In these methods, gazetteers have been widely used to improve the performance. There are mainly two kinds of gazetteers, the NE gazetteer and the feature gazetteer, which are used to store common NEs and features of NEs respectively. Many studies have been devoted on the automatic development of high quality gazetteers. Kozareva [6] generated gazetteers by choosing items with capital letter as candidates and building a directed graph to connect family and given names appearing simultaneously in the corpus. The connections that appear less than a threshold were removed, and the remaining connected nodes were taken as gazetteer items. This method is limited to person names, and may generate lots of not-a-person-name items because many proper nouns also begin with capital. Toral and Munoz [7] proposed a solution to build gazetteers by analyzing the nouns in the first sentence of entries form Wikipedia [8] with the aid of WordNet [9]. The gazetteers were updated to match the latest Wikipedia pages by simply executing the construction procedure over and over again, which will be very time consuming and may extract the same items repeatedly. Chen, Zhang and Isahara constructed gazetteers of NE features by extracting N-grams from training data and created a temporary list of the recognized NEs for the current under-processing document [5]. This temporary list was then used to help correct the recognition result in the post-processing procedure, but the list was not saved for later reuse.

Nevertheless, the gazetteers in the above method are all constructed before or within the training process of NER model and therefore they are tied up with the training corpora or the knowledge resources. Since a large number of new OOVNEs spring up every day, no matter what are the sizes of the gazetteers, they cannot cover the new OOVNEs. What's more, the sets of NEs and the features of NEs are quite different in various domains. A NER system with gazetteers containing only NEs and NE features in one domain (e.g. sport) cannot perform well in recognizing NEs for the document from another domain (e.g. politics) with quite different NEs and NE features.

To solve the problem, this paper proposes a method to improve the adaptability of the NER system by building and maintaining dynamic gazetteers (DGs). These DGs consist of NE gazetteers and feature gazetteers. Unlike the method in Chen, Zhang and Isahara [5], our method saves the recognized NEs into the NE gazetteers and updates the gazetteers along with the recognition process of NEs. Furthermore, both internal and external feature gazetteers are used. Internal features consist of prefix (the first character or word) and suffix (the last character or word) of NE. External features include precursor (the characters right before NE) and successor (the characters right after NE) of NE.

2. ADAPTIVE NER SYSTEM

Our NER system takes the framework of conditional random field (CRF). A CRF may be viewed as undirected graphic model, which relaxes the independence assumption and avoids the bias label problem in comparison with HMM [10]. The dynamic gazetteers (DGs) that can be automatically updated with the recognition process are further proposed and incorporated into CRF framework so that our NER system can be adapted to different domains.

2.1. CRF Model

The task of NER is to assign each Chinese character in the input document with a tag from a predefined tag set for identifying named entities. The BIO-style tags are adopted, where 'B' stands for 'Begin' meaning that the character is at the beginning of an entity; 'I' for 'Intermediate' meaning that the character is in the entity but not at the beginning; and 'O' for 'Other' meaning that the character is not part of an entity. The BIO-style tags are combined with the three categories of NEs, including person name (PERN), organization name (ORGN) and location (LOCN). The final NE tag set consists of seven tags: "PERNB", "PERNI", "ORGNB", "ORGNI", "LOCNB", "LOCNI" and "OTHER".

CRF provides an efficient framework for sequence labeling. The probability of a particular tag sequence y given character sequence x takes the form:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \left[\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i) \right] \right) \quad (1)$$

where x is the given character sequence; y is the tag sequence; y_i is the tag at position i; $t_j(y_{i-1},y_i,x,i)$ represents j-th transition feature function of the character sequence x and the assigned tags y_i and y_{i-1} at positions i and i-1; $s_k(y_i,x,i)$ is k-th state feature function of the tag y_i at position i and the character sequence x; λ_j and μ_k are the weights of transition feature functions and state feature functions respectively and can be obtained by training. In brief, the feature function is an indicator to show that whether there is a feature like this at position i. Take $s_k(y_i,x,i)$ as example, it means the given character sequence is x and the assigned tag is y_i at position i. $Z(x)$ is a normalization factor to make the entire probabilities sum up to 1, which is calculated as:

$$Z(x) = \sum_y \exp\left(\sum_i \left[\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i) \right] \right) \quad (2)$$

Feature templates are defined so that features, including state features and transition features, can be automatically generated. According to the internal and external evidences of the named entities, a set of feature templates have been defined in our work. Some of the templates are shown in Table 2. For example, with template X_1, it will generate a feature, the next character, for each character in the training data.

Table.2. Definitions and examples of feature templates

Type	Templates	Definition
Unigram	X_{-1}, X_0, X_1	The previous, current and next character
Bigram	$X_{-1}X_0$, X_0X_1	The previous (next) character and current character
Gazetteer	G_{-1}, G_0, G_1	The previous, current and next character is in Gazetteers
	$G_{-1}G_0$, G_0G_1	The previous (next) character and current character are both in Gazetteers

2.2. Dynamic Gazetteers

The DGs used in this paper consist of NE gazetteers and feature gazetteers. NE gazetteers contain the items for person name, organization name and location, and feature gazetteers contain the internal or external context features of NEs.

Building of DGs is motivated by the following observations:

- One NE can be easily recognized in certain contexts with clear evidence, while may be extremely hard to recognize in some other contexts without obvious external features.
- Some NEs lacking internal features are always missed by NER system if only internal features are considered. But they may be easily recognized using the external features.
- The most common features are regular for different documents belonging to the same domain. While they are quite different for different domains.

For the 1st observation, the NEs recognized easily and confidently in some cases should be added to the NE gazetteers for the succeeding NE recognition. For the 2nd and 3rd observations, the features appearing more frequently should be added to the feature gazetteers to enhance the predicting power of gazetteers.

Table.3. Common features in three NE categories

Category	Feature	Most common feature (with frequency %)
LOCN	Suffix	国(country)(22.3), 州(state)(4.98)
ORGN	Suffix	社(agency)(3.99), 公司(company)(3.22)
PERN	Suffix	泽民(Zemin)(3.98), 鹏(Peng)(1.65)

Table 3 lists the most common features and their frequencies, where the frequencies are calculated among different features in three categories respectively. Although there are various NE features, some features are much more common than others. For example, for the suffix of location, "国(country)" appears more frequently than "州(state)". The feature gazetteers containing the most common features can cover most informative information for the recognition of NEs of that category. Furthermore, the items of the most common features are quite different for three different categories. Because of this, four gazetteers (prefix, suffix, precursor and successor) are built for three NE categories.

2.3. Updating Method of Dynamic Gazetteers

Initial gazetteers are created at training stage. The NEs and their features in the training data are taken down into our initial gazetteers. In our experiment, we found that, in the test set, only 40% of external features are included in the initial gazetteers before dynamical update.

The procedure for gazetteer update is as follows:

- Firstly, the recognized NEs and their features are added to temporary gazetteers as candidates;
- Secondly, the NEs or features in the temporary gazetteers that match the following conditions are added into DGs:

a) The length of the NE should be greater than 2.

b) The frequency of the NE should be greater than a threshold calculated previously at training stage.

c) The frequency of the feature should be greater than a threshold calculated at training stage.

Condition a) is to exclude NEs that are too short. According to the statistics of our experiments, only 0.24% of NEs have the length of 1, and 35.35% of NEs have the length of 2. 47.06% of such NEs are locations such as "北京" (Beijing/Peking). As organization names often begin with such a short location name, e.g. "北京大学" (Peking University), the existing of these short names in the gazetteers will seriously degrade the recall rate of NER result. Furthermore, the recognition accuracies of these short location names are fairly high (95.07%) even they are not included in the gazetteers. Because of this, condition a) is introduced to limit the length of NEs included in the gazetteers and exclude those short confusing items.

Condition b) rejects rare NEs and those recognized by chance. The NEs with low frequencies are usually rare or error NEs, and they offer few help for recognizing new NEs. Including these NEs in the gazetteers will increase the sizes of gazetteers dramatically, which decreases the performance of the gazetteers. The threshold is the average frequency extracted from the training data. With this threshold, which is around 4, most of NEs recognized by mistake can be successfully excluded.

As for condition c), the features, which are rare and exceptional with low frequencies in temporary gazetteers, should not be collected either. The threshold is also extracted from the training data. Experiments in section 3 demonstrate the necessity of these frequency conditions.

3. EXPERIMENTS

We used the CRF++ toolkit [12] to build the NER system and conducted experiments using the People's Daily (P.D.) corpus which contains the Chinese news articles of the year 2000.

3.1. The Corpus

The topic of each news article in the P.D. corpus is related to one of the 12 different categories, including art, economy, politics, sport, education, entertainment, health, history, science and so on. The news articles in the same topic category are considered belonging to the same domain. Hence, we have documents from 12 different domains for the experiments.

The text of each document in the P.D. corpus is tokenized into Chinese words (with space) and annotated with Part-of-Speech (POS) tags (in the form of "/POS"). An example of the annotated text is shown in Figure 1(a). Among all the POS tags, 5 special tags ("nt", "ns", "nr", "nrf", "nrg") have been used to annotate the POS of the named entities, where "nt" is the POS tag for organization name, "ns" is for location, "nrf"/"nrg" are for the family/given name of a person name respectively, and "nr" is used for the person name without family or given name.

We generated the data files for NER experiments automatically from the P.D. corpus according to the word tokenization and POS tagging results. Each line of the data file contains the information of one Chinese character or punctuation, in three columns separated by white space, including the Chinese character, the POS tag of the word which contains the character, and the NE tag of the character. As for the NE tag, it is generated by combining the BIO tag with the named entity type, where the BIO tag is derived from the word tokenization information of the P.D. corpus and the named entity type (i.e. "ORGN", "LOCN", "PERN", "OTHER") is derived from the POS tag. An empty line in the data file indicates the boundary between sentences. An example piece of the data file is shown in Figure 1(b), which is generated from the text in Figure 1(a).

"成都/ns 举办/v 画展/n 迎接/v 新/a 千年/t 。/w"
(Chengdu holds art exhibition to welcome the new millennium.)

(a)

```
成 ns LOCNB
都 ns LOCNI
举 v OTHER
办 v OTHER
画 n OTHER
展 n OTHER
……
```

(b)

Fig.1. An example text from the P.D. corpus with word tokenization and POS tagging results (a), and the sample piece of the corresponding data files for named entity recognition (b).

3.2. Experimental Setup and Results

Four experiments were conducted to evaluate if our proposed method with dynamic gazetteers (DGs) can be automatically adapted to new domains and can achieve better performance than the conventional methods with static gazetteers (SGs).

We evaluate the results of our experiments with F-score (*F*) and Increase rate (*I*) of F-score, defined as follows:

$$F = \frac{2C}{A+B} \qquad (3)$$

$$I = \frac{F_{DGs} - F_{SGs}}{F_{SGs}} \qquad (4)$$

where *A* is the number of NEs in the test data file, *B* is the number of the recognized NEs (by DGs or SGs system), *C* is the number of NEs that are correctly recognized. F_{DGs} and F_{SGs} are the F-scores for the DGs and SGs systems respectively.

3.2.1. Threshold Experiment

The first experiment is designed to demonstrate the necessity of threshold. We compare two systems which update gazetteers with and without thresholds respectively. In the no-threshold system, any NE and corresponding features recognized are collected in gazetteers, without considering the frequencies. In the with-threshold system, only the NEs and features that match the threshold conditions as explained in section 2.3 are added to gazetteers. Each system is evaluated with 5 data files one by one, which are from 6 domains including art, international, health, history, law and society, and the NER result is shown in Figure 2, where the vertical axis represents the F-score of the NER, and the horizontal axis shows the No. of data file. As can be seen, with the update of gazetteers, the with-threshold system outperforms the no-threshold one, and the performance of the no-threshold system decreases seriously because of the negative influence of the unnecessary NEs and features added to the gazetteers. This indicates that the thresholds for the system are necessary.

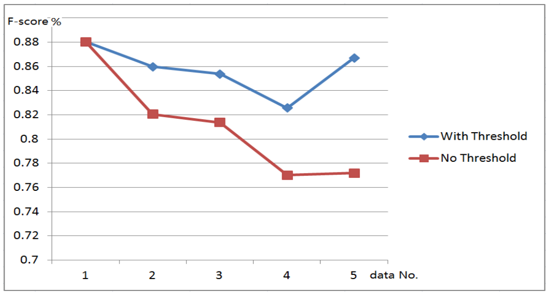

Fig.2. NER result of the two systems with and without thresholds.

3.2.2. Out-Of-Domain Experiment

The second experiment is designed to evaluate whether our proposed method with DGs can improve the performance of NE recognition for unknown domains when the CRF model and the initial gazetteers are trained with sufficient amount of data from a wide range of different domains.

In this experiment, the two systems were trained with the data from 6 domains including art, international, health, history, law and society. The training data contains about 480,000 characters.

The two systems (DGs and SGs) were then evaluated with the test data from another 3 domains of politics, economy and sport. These 3 domains are not among the domains of the training data. For each domain, 3 data files were used, with each data file contains about 6,000 characters. The systems were tested by recognizing NEs from 3 data files one by one for each domain. The DGs of our method were restored to the initial gazetteers before recognizing NEs of the 1st data file of a new domain, and were continuously updated while recognizing NEs of all 3 files.

The result of this experiment is tabulated in Table 4. As can be seen, our method with DGs outperforms the conventional one with SGs in identifying the NEs. This indicates that the DGs in our method can successfully update themselves to new unknown domains, and this update is accomplished in the fully automatic manner during the recognition process of NEs.

Table.4. NER result for the out-of-domain experiment

Data File	System	Politics		Economy		Sport	
		F	*I*	*F*	*I*	*F*	*I*
1st	SGs	92.25	0.99	87.07	0.56	85.83	3.09
	DGs	93.16		87.56		88.48	
2nd	SGs	91.47	1.16	91.16	0.27	83.51	4.48
	DGs	92.53		91.41		87.25	
3rd	SGs	92.28	1.25	89.27	1.37	83.70	3.93
	DGs	**93.43**		**90.49**		**86.99**	

3.2.3. In-Domain Experiment

The third experiment is designed to evaluate if our method with DGs can still improve the NER performance even when the test data belongs to the same domain of the training data.

In this experiment, the training data from all 12 common domains were used to train the two systems. The training data contains around 960,000 characters. The test data and the test procedures were the same as that in the second experiment. And the test data were not covered by the training data.

Table 5 shows the result of this experiment, which indicates that our proposed method with DGs can further improve the NER performance of the models which are trained with the data covering a wide range of domains.

Table.5. NER result for the in-domain experiment

Data File	System	Politics		Economy		Sport	
		F	*I*	*F*	*I*	*F*	*I*
1st	SGs	93.44	0.22	88.84	0.08	87.62	0.63
	DGs	93.65		88.91		88.17	
2nd	SGs	92.80	0.36	92.05	1.10	85.15	3.17
	DGs	93.13		93.06		87.85	
3rd	SGs	93.54	0.45	90.98	1.17	85.51	1.19
	DGs	**93.96**		**92.04**		**86.53**	

Table.6. NER result for the domain sensitivity experiment

Data File	System	Politics		Economy		Sport	
		F	*I*	*F*	*I*	*F*	*I*
1st	SGs	89.42	0.07	85.55	-0.25	85.24	0.47
	DGs	89.48		85.34		85.64	
2nd	SGs	89.21	0.27	87.71	-1.14	82.09	0.77
	DGs	89.45		86.71		82.72	
3rd	SGs	90.39	0.31	86.30	0.86	82.36	2.10
	DGs	**90.67**		**87.04**		**84.09**	

3.2.4. Domain Sensitivity Experiment

The fourth experiment is conducted to validate if our method is sensitive to the number of domains of the training data.

In this experiment, the two systems were trained with the training data from only one domain (law domain). The training data contains about 40,000 characters. The test data and the test procedures were also the same as that in the above experiments.

Details of the result are shown in Table 6. As can be seen, from the 1st data file to the 3rd data file, the Increase rate (*I*) of F-score between the DGs system and the SGs system tends to be bigger. This is consistent with our expectation that the DGs system becomes more adaptive to the new domains. Nevertheless, in the tests in economy domain with the first 2 test data files, the performance of DGs system is worse than the SGs one. This is because the DGs system is trained with data files in law domain; the features obtained in law domain are quite different from those in economy domain which cause a few errors. However, the DGs system catches up with the SGs one in the test with the last test file, which demonstrates that our method is effective.

4. CONCLUSIONS

This paper presents a hybrid model which combines CRFs with DGs for the task of Chinese NER. We build and maintain DGs to improve the adaptability of NER system and propose a method to automatically update DGs along with the recognition process of NEs. With this method, the DGs can be updated to contain more and more new NEs and features of NEs that are not found in the training data. These newly added items make the DGs become more aware of the knowledge about new domains and hence be more adaptive to new domains for the recognition of OOVNEs. Experiments indicate the effectiveness of our method.

5. ACKNOWLEDGEMENTS

This work is partially supported by the National Natural Science Foundation of China (60928005, 60805008 and 61003094).

6. REFERENCES

[1] S. Richard, and T. Emerson, “The First International Chinese Word Segmentation Bakeoff”, In: *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing*, pp. 133-143, 2003.

[2] C. Huang, and H. Zhao, “Chinese Word Segmentation: A Decade Review”, *Journal of Chinese Information Processing*, The Commercial Press, China, 21(3): pp. 8-19, 2007.

[3] G. Zhou, and J. Su, “Named Entity Recognition using an HMM-based Chunk Tagger”, In: *Proc. of 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, pp. 473-480, 2002.

[4] H. Isozaki, and H. Kazawa, “Efficient Support Vector Classifiers for Named Entity Recognition”, In: *Proc. of the 19th International Conference on Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, 1: pp. 1-7, 2002.

[5] W. Chen, Y. Zhang, and H. Isahara, “Chinese Named Entity Recognition with Conditional Random Fields”, In: *Proc. of the 5th SIGHAN Workshop on Chinese Language Processing*, Association for Computational Linguistics, Stroudsburg, pp. 118-121, 2006.

[6] Z. Kozareva, “Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists”, In: *Proc. of the 11th Conference of the European Chapter of the ACL*, Association for Computational Linguistics, Stroudsburg, pp. 15-21, 2006.

[7] A. Toral, and R. Munoz, “A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by using Wikipedia”, In: *Proc. of the EACL-2006 Workshop on NEW TEXT-Wikis and blogs and other dynamic text sources*, Association for Computational Linguistics, Stroudsburg, pp. 56-61, 2006.

[8] http:// www.wikipedia.org.

[9] G. Miller, “Wordnet: A Lexical Database for English”, *Communications of ACM*, Association for Computing Machinery, New York, 38(11):39-41, 1995.

[10] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, In: *Proc. of the 18th International Conf. on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, pp. 282-289, 2001.

[11] J. Morris, and E. Fosler-Lussier, “Combining Phonetic Attributes Using Conditional Random Fields”, In: *Proc. of INTERSPEECH*, pp. 597-600, 2006.

[12] http://chasen.org/~taku/software/CRF++/.

[13] C. Parada, M. Dredze, and F. Jelinek, “OOV Sensitive Named-Entity Recognition in Speech”, In: *Proc. of INTERSPEECH*, pp. 2085-2088, 2011.