Journal of Computational Information Systems5:6(2009) 1585-1592
Available at http://www.JofCI.org

Research on Chinese Proper Nouns Recognition Based on Pattern Matching

Hu ZHANG†, Jiaheng ZHENG, Xingyi WANG

School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China

Abstract

Chinese proper nouns recognition (CPNR) plays an important role in the fields of Information Extraction, Question Answering and Text Mining. In this paper we proposed a novel pattern-matching-based method to recognize proper nouns (PNs), which includes person names, location names and organization names, and mainly conducted the following studies: (1) constructing the PNs inner-pattern set; (2) acquiring the PNs outer-patterns by clustering and evaluating automatically; (3) resolving the conflicts of the PNs recognition by computing PNs reliability; (4) conducting the experiment on 1.2M word corpus that are chosen from People Daily corpora. The experimental results, whose recall and precision are 83.4% and 80.1% respectively, indicate that the proposed method is feasible and effective.

Keywords: Natural Language Processing; Proper Nouns Recognition; Pattern-matching; Information Extraction

1. Introduction

The Proper Nouns (PNs) Recognition is that identifying PNs including persons, organizations and locations in the text and labeling the correct semantic categories to them, which is an important part in such applied fields of natural language processing as Information Extraction, Question Answering, Machine Translation, etc.

In present a large number of researches of PNs recognition have been made at home and abroad. For PNs recognition of foreign languages, some methods are put forward such as Hidden Markov Model(HMM)[1-3], Maximun Entropy Model[4-6], Decision Tree[7], and Bootstrapping[8-10]. For Chinese PNs recognition, according to the character information of PNs and their utilization rate in real texts Zheng et al[11] proposed a corpus-based method which is used to recognize person names; Huang et al[12] used statistic information extracted from a training corpus to calculate lexical reliability and context reliability to recognize location names; Zhang et al[13] proposed an automatic recognition method based on role labeling for Chinese person names; Wang et al[14] studied company names recognition by analyzing features of their structure and the outer-field; and Zhu et al[15] put forward an organization names recognition method by co-training from the unlabeled corpus.

But there are some problems in existing approaches. Based on the statistic methods the candidate PNs whose probability is less than the threshold value will be eliminated, which results in a number of PNs lost,

† Corresponding author.
Email addresses: zhanghu@sxu.edu.cn (Hu ZHANG).

December, 2009

while rule-based methods whose expansibility is inferior lack flexibility, and some amendments need be made if the changes appear in the corpus.

According to the analysis of the automatic word segmentation results by the expert team of 863-306 intellectual joint technology[16], the precision of Chinese person names , location names, and foreign translated names are 91.26%, 69.12%, and 82.83%, while their recall are only 68.77%, 60.47%, and 78.29% respectively. Therefore PNs recognition is still an urgent problem to be solved, and after further analyzing the PNs we found there are some difficulties in recognizing them, which can be described as follows: ①The PNs belong to an open category, which cannot be recognized by enumerating, because there are hundreds of millions of names in the location category only. The PNs also belong to an unsteady category in which the old PNs are disappearing while the new ones are appearing; ②It's hard to determine which category a proper noun belongs to. Take "克里姆林宫(*Kremlin*)" for example. It's difficult to identify whether it's an organization name or a location name in the different context; ③There are also alias names and abbreviations recognition of the PNs. Take "上海航空公司(*Shang hai hang kong gong si, Shanghai Airways*)" for instance. It will be "上航(*Shang hang*)" when it appears next in this text; ④Some PNs have not strict naming rules to follow; ⑤Unlike English there are no spaces between words for separation in Chinese, which brings ambiguity conflicts in PNs themselves, between PNs, and between the PNs and their contexts. For example, there are conflicts among "重庆市(*Chongqin*)"，"市长(*mayor*)" and "长寿(*long life*)" in the PN "重庆市长寿县(*Changshou county Chongqin*)".

Based on the preceding analysis we proposed a PN recognition method based on pattern matching to label three type PNs that appear most frequently. In this method PNs are recognized by searching and matching the outer-pattern firstly, and are further judged by the inner-pattern, at the same time during which the conflicts will be resolved by computing and comparing the PNs reliability if there are.

2. The Construction of Knowledge Base

2.1. Indicative Words Base

Indicative words are those indicative words of PN recognition, which appear more frequently in the sentences containing PNs and have semantic association with PNs, such as duty (*chairman, premier, and foreign minister*), titles(*Comrade, Mr., Mrs.*), particular verbs(*say, appoint, meet, garrison, occupy*), etc.

Indicative words with higher reliability have better indication for PNs recognition. The formula of calculating the reliability is showed below:

$$R(W_i) = \frac{N(PN, W_i)}{N(W_i)} \log[N(PN, W_i) + 1]$$

Where $R(W_i)$ is the reliability of W_i as a indicative word; $N(PN, W_i)$ represents the number of times that W_i appears in the sentences containing PN; $N(W_i)$ is the number of times that W_i appears in the corpora.

2.2. Common Location Names Base

Some common location names are collected in common location names base, such as the all nation names, famous cities in the world, provinces and cities of China etc.. The common location names that appear more frequently in real texts are sometimes wrongly recalled, so the reliability given to every name show the probability of the character string as a location name, which is defined as follows:

$$R_{comm}(Cstr) = \frac{Match_{Corr}(Cstr)}{Match(Cstr)}$$

Where *Match(Cstr)* denotes the number of times of the character *Cstr* matching as location names. $Match_{corr}(Cstr)$ expresses the number of times the character *Cstr* correctly matching as location names.

2.3. Border Words Base

The border words are the context borders of the PNs, involving words, punctuation, the beginning and the end of a sentence etc. Take "张正来到四川省汶川县调查 (*Zhang zheng come to Wenchuan county SiChuan province and conduct investigation*)" for instance. The border words of the person name "张正 (*Zhang zheng*)" are the beginning and "来到(*com to*)"，while the border words of the location name "四川省汶川县(*Wenchuan county SiChuan province*)" are "来到(*come to*)" and "调查(*conduct investigation*)".

3. Inner-pattern Structure

The inner-pattern is the formal representation of the inner structure of every type PNs. The inner-patterns of every type are all established based on probability statistics and human analysis, which are expressed in the similar ***BNF****(Backus-Naur Form)*. Here are some examples of inner-pattern structure.

① <Cname_FamilyName>{<Cname_GivenName1>}<Cname_GivenName2>

E.g. 史大桢(*Shi Dazhen*), 罗玉文(*Luo Yuwen*), 祁勇(*Qi Yong*), 高严(*Gao Yan*) etc.

② [<Cplace_Character>]<Location_FeatureWord|Null>

E.g. 安徽省(*Anhui Province*), 日喀则地区(*the Ri Kaze Region*), 寒山寺(*Hanshan Temple*)

③ <Location_Name>{<"人民"|"国民">}<"政府"|"内阁"|"议会"|"国会">

E.g. 中国政府(*the Chinese Government*), 日本内阁(*the Japanese Cabinet*), 美国国会(*the Congress*)

Where Cname_FamilyName denotes Chinese family name; Cname_GivenName1 is the first given name of a Chinese name; Cname_GivenName2 is the last given name of a Chinese name; Cplace_Character means the usage-character of Chinese location name; Location_FeatureWord is the feature words of the location name such as Province, Municipality, Channel, Road, Mountain, etc..

4. Outer-pattern Structure

4.1. Definition of the Outer-pattern

Generally the outer-pattern is regarded as a sequence made up of items arrayed in proper order, and each item is corresponded to a set of words or phrases that have the same or similar meaning in the present information field. The items can be classified into three types according to their different roles:

① **Feature:** Feature item is used to mark the different outer-patterns, and by which we can search and match the pattern from the pattern base.

② **Extraction:** Extraction item is the extraction information defined in the pattern.

③ **Alternation:** Alternation item can further ensure the accuracy of the information extraction, which is not necessary in the pattern-matching.

E.g. in the pattern <PER>（"还"）["指出"](*PER also point out*), ["指出"]is feature item, <PER> is extraction item,（"还"）is alternation item. The outer-patterns are collected in the outer-pattern base.

4.2. Production of the Pattern Examples

The pattern examples can be generated according to the size of the window of the PNs contexts. Here is the example sentence:

$$W_{-n}\ldots\ldots W_{-2}W_{-1}<PN>W_1W_2\ldots\ldots W_n$$

Where <PN>is the PN in the sentence, W_i is the outer word of the proper noun <PN>, and n is the size of the window, and set n=3. Two pattern examples on the <PN> left and right can be generated for every PN, i.e. E_l=$W_{-3}W_{-2}W_{-1}$<PN> and E_r=<PN>W_1W_2 W_3.

4.3. Clustering of the Pattern Examples

The pattern examples can be viewed as ordered sequence of items, and the items are basic linguistic elements (characters, words or phrases) in pattern example. So we define the similarity according to common subsequence of the item sequences corresponding to the pattern examples. Suppose that there are pattern examples E_i and E_j, and the similarity between them is $Sim(E_i, E_j)$:

$$Sim(E_i, E_j) = max(Score(Comm(E_i, E_j))$$

Where $Comm(E_i,E_j)$ is one of the common subsequences of E_i and E_j, $Score(Comm(E_i,E_j))$ denotes the score of the common subsequence $Comm(E_i,E_j)$. The definition of the common subsequence score is:

$$Score(Comm(E_i, E_j)) = \frac{\sum t(Comm_k(E_i, E_j)) \times R(Comm_k(E_i, E_j))}{|E_i| + |E_j| - Num(Comm(E_i, E_j))}$$

Where $Comm_k(E_i,E_j)$ denotes the k item in the common subsequence; $t(Comm_k(E_i,E_j))$ is the position weight of $Comm_k(E_i,E_j)$; $R(Comm_k(E_i,E_j))$ is reability of $Comm_k(E_i,E_j)$ as indicative word; $|E_i|$ and $|E_j|$ are the length of E_i and E_j, respectively; $Num(Comm_k(E_i,E_j))$ denotes the number of items in $Comm(E_i,E_j)$.

The position weight of $Comm_k(E_i,E_j)$ is formulated as follows:

$$t(Comm_k(E_i,E_j)) = W(min(|pos(Comm_k(E_i,E_j), E_i)|, |pos(Comm_k(E_i,E_j), E_j)|))$$

Where $pos(Comm_k(E_i,E_j), E_i)$, $pos(Comm_k(E_i,E_j), E_j)$ are the position of $Comm_k(E_i,E_j)$ in pattern example E_i and E_j, respectively. The position weight $W(i)$ are separately 0.5, 0.3 and 0.2 when i=1, i=2 and i=3.

4.4. Evaluation of the Outer-pattern

By comparing PNs recognition results and the standard corpus, the PNs recognition can be classified into correctly recognized PNs and erroneously recognized PNs. Then the reliability of $canP_i$ can be calculated by this following formula:

$$R(canP_i) = \frac{Corr(canP_i)}{Corr(canP_i) + Err(canP_i)}$$

Where $R(canP_i)$ is the reliability of the candidate pattern $canP_i$; $Corr(canP_i)$ is the number of correctly recognized PNs by candidate pattern $canP_i$, while Err(canPi) is the number of erroneously recognized PNs.

Based on the computing results by the above formula the candidate pattern whose reliability is bigger than the threshold θ_{prec} is accepted into the outer-pattern base.

5. PNs Recognition Based on Pattern-matching

5.1. PNs Recognition Model

By analyzing the PNs inner structure and outer information in Chinese texts, we proposed a novel PNs recognition method based on pattern-matching, which is showed in the Fig.1.

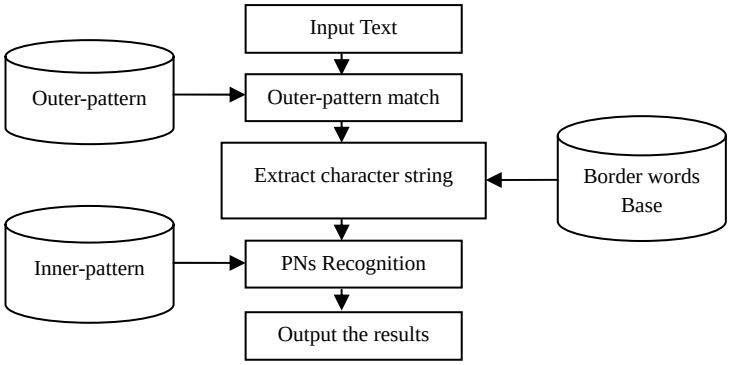

Fig.1 PNs Recognition Model

5.2. Computing the Reliability of the Candidate PNs

For computing the reliability of the candidate PNs we take a thorough account of the reliability of the PNs outer-pattern and inner-pattern, and whether the PNs are common and appear in front or not. That is defined as follows:

$$R(canPN) = R_{comm}(canPN)*t_{comm} + R_{out}(canPN)*t_{out} + R_{in}(canPN)*t_{in} + R_{temp}(canPN)*t_{temp} + FCinflu$$

Where *R(canPN)* is the reliability of a candidate PN; $R_{comm}(canPN)$ is the reliability of the *canPN* as a common one; $R_{out}(canPN)$ and $R_{in}(canPN)$ are the reliability of its outer-pattern and inner-pattern respectively; $R_{temp}(canPn)$ is the reliability of the *canPN* as a temporary one; *FCinflu* is the effect of word segmentation on *canPN*; and t_{comm}, t_{out}, t_{temp} and t_{in} are the weight of different reliability in the formula.

① Computing $R_{comm}(canPN)$

If the candidate proper noun is common, the given value of $R_{comm}(canPN)$ is the reliability that is in the common location names base, otherwise the value is 0.

② Computing $R_{out}(canPN)$

Taking the reliability of PNs pattern and border into consideration, the reliability of the canPN outer-pattern can be computed by the following formula:

$$R_{out}(canPN) = R_{l/r\_p}(canPN)*t_p + R_{l/r\_b}(canPN)*t_b$$

Where $R_{1/r\_p}(canPN)$ is the reliability of *canPN*'s trigger-pattern, R_{1/r_b} is that of its border, and t_p and t_b are their weight, respectively.

③ Computing $R_{temp}(canPN)$

If the recognized PNs are not in the common PNs base, it will be added into the temporary PNs table which can be used to acquire the PNs appearing repeatedly by directly matching. The reliability of the PNs

that are in the temporary table is *0.5lg(Len(canPN))* that is the logarithm of their length. That is to say, the value of $R_{temp}(canPN)$ is the given reliability if the *canPN* appears in the temporary table, otherwise it is 0.

④ Computing *FCinlfu*

By using the word segmentation (WS) information, the parameter of *FCinlfu* exerts effect on PNs recognition by the reward and penalty value.

5.3. Conflict Detection and Resolution of the Candidate PNs

To add the candidate PNs whose reliability is higher than the threshold into the PNs table, we firstly should detect the conflicts between the PNs and other names, and further resolve them if there are. Generally the conflicts can be classified into three categories:

① **Overlapping:** There are overlapping among the recognized PNs. For example, there are overlapping parts "黑龙江(*HeilLongjiang*)" and "江西(*Jiangxi*)" in "黑龙江西北部(*north-west of Hei longjiang Xibei region*)".

② **Inclusion:** The recognized PN is inclusive of other names. It can be classified further:

i、An organization name contains a person name, location name or organization name.

For example, the person name "宋庆龄(*Song Qingling*)" is contained in the organization name "宋庆龄儿童基金(*Children's Foundation of Song Qingling)*"**;** and the location name "白沙县(*Baisha County*)" is included in the organization name "白沙县中等职业技术学校(*Secondary Vocational and Technical School of Baisha County*)".

ii、A location name includes a person name, location name or organization name.

For example, the location name "四川省凉山(*Liangshan in Sichuan Province*)" is included in "四川省凉山彝族自治州(*Sichuan Liangshan Yi Autonomous Region*)"*;* and "周恩来邓颖超纪念馆(*Zhou Enlai & Deng Yingchao Museum*)" is inclusive of the person names "周恩来(*Zhou Enlai*)" and "邓颖超(*Deng Yingchao*)".

iii、A person name includes a person name or a location name

For example, the person name "孟西安(*Meng Xi'an*)" and the location name "西安(*Xi'an*)" are both contained in the "孟西安(*Meng Xi'an)*".

③ **Simultaneous statement**: A character string is recognized for the different type PNs at the same time. Take "阳安江(*yang an jiang)*" for instance. It is recognized for a person name (Yang Anjiang) and a location name (Yang'an River) on the first recognition stage, therefore there must be recognition errors that need be solved if the conflicts appear.

When the conflicts of overlapping and simultaneous statement appear, the PN with lower reliability will be refused by comparing the reliability. And when the conflict of inclusion happens, it should be judged its' realness, and be resolved in the same way if it is real. Moreover, if an inclusion conflict is justified to be reasonable, such as an organization name containing a location name, both of the recognized PNs will be accepted.

6. Experimental Results and Analysis

The experimental corpora come from People Daily corpora in January 1998, in which 0.6 million corpora selected from the first to the tenth day are used to produce the outer-pattern and make the close test, and 0.6

million corpora chosen from the eleventh to the twentieth day are used to conduct the open test.

6.1. PNs Recognition Experimental Results

By using the pattern-matching method we conducted the close and open test respectively, and the experimental results are showed in Table 1:

Table 1 The Experimental Results of the PNs recognition

Type	Close test			Open test		
	Recall(%)	**Precision**(%)	**F**(%)	**Recall**(%)	**Precision**(%)	**F**(%)
PER	94.4	90.4	92.4	90.9	86.5	88.6
LOC	90.0	85.4	87.6	85.5	82.9	84.2
ORG	84.2	82.4	83.3	73.7	70.8	72.8
Average	89.5	86.1	87.8	83.4	80.1	81.9

Based on the above experimental results two interesting trends can be seen. First, the PNs recognition achieves better results in the close test than in the open test, which reason is that some outer-patterns in the open test are not included in the outer-pattern base. Second, the recall and precision of the organization names recognition are the lowest in the three types PNs recognitions.

6.2. The Experiment on the Separate and Unified PNs Recognition

In this experiment we choose 0.6M word corpus to conduct the open test, which includes the separate test and unified test of person names, location names and organization names.

Table 2 The Experimental Results of the Separate and United Recognition

Type	Separate recognition			Unified recognition		
	Recall(%)	**Precision**(%)	**F**(%)	**Recall**(%)	**Precision**(%)	**F**(%)
PER	87.8	84.4	86.1	90.9	86.5	88.6
LOC	80.6	76.5	78.5	85.5	82.9	84.3
ORG	62.4	60.8	61.5	73.7	70.8	72.8

From the above table it can be seen that the experiment acquired better Recall, Precision and F-value by using the unified PNs recognition. The main reasons are showed as follows:

① The recognition of one type PNs offers useful information to other types.

For instance, when the candidate organization name "德阳市恒达企业集团(*Hengda Enterprise of the City Deyang*)" is judged by inner-pattern, the correct recognition of the location name "德阳市(the *City Deyang*)" will improve the reliability of the inner-pattern.

② The unified recognition of several types PNs results in the competitive recognition.

If a candidate PN is correctly recognized, it's avoidable to mistake it to be other types, which decreases the possibility of making a wrong PNs recognition. For example, if "下岗职工生产自救服务中心负责人[王柏林]PER ([Wang Bolin]PER，*the Manager of the Labor Service Center of the Rising Production for the Lay-off*s *Self-reliance*)" is correctly recognized by the unified recognition, it's impossible to make such a mistake as "王 柏林[LOC](Wang [Berlin]LOC)".

7. Conclusions and Future Research

The experimental results show the pattern-matching-based PNs recognition method, who makes full use of both the outer enlightening data and inner indicative data, is feasible and effective, but there are still some researches to be enhanced: ①the outer and inner patterns need be further improved; ②the scale of the experimental corpus need be enlarged yet.

Acknowledgement

This work is partially supported by Natural Science Foundation of China (No.60775041) and Science and Technology Development Project of Shanxi Province (No.20091001).

References

[1] Daniel M. Bikel, et al. Nymble: a High-Performance Learning Name-finder. In Proc 5th Conf on Applied Natural Language Processing, Washinton DC, 1997, pp.194-201
[2] Scott Miller, Michael Crystal, Heidi Fox, et al. Algorithms that Learn to Extract Information BBN: Description of the SIFT SYSTEM as used for MUC-7. In Proc of 7th Message Understanding Conference, 1998, pp.75-89
[3] George R, Krupka. SRA: Description of the SRA system as used for MUC-6. In Proc 6th Message Understanding Conference, 1995, pp.221–235
[4] Mikheev A., Grover C. LTG: Description of the NE recognition system as used for MUC-7. In Proc 7th Message Understanding Conference(MUC-7), 1998, pp.1-12
[5] Borthwich. A. A Maximum Entropy Approach to Named Entity Recognition. PhD. thesis, Computer Science Department, New York University, 1999
[6] Hai Leong Chieu, Hwee Tou Ng. Named Entity Recognition:A Maximum Entropy Approach Using Global Information. In Proc of the 19th International Conference on Computational Linguistics, 2002, pp.190～196
[7] Sekine S, Grishman R., et al. A decision tree method for finding and classifying names in Japanese texts. In Proc 6th Workshop on Very Large Corpora, Montreal, Canada, 1998, pp.171-177
[8] Collins, Singer. Unsupervised Models for Named Entity Classification. Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, College Park, Maryland, 1999, pp.100-110
[9] Strzalkowski T, Wang J. A self-learning universal concept spoteer. In Pro. 16th Int'l Conf. On Computational Linguistics (COLING 96), Copenhagen, Denmark, 1999, pp. 931-936
[10] Cucerzan S, Yarowsky D. Language independent named entity recognition combining morphological and contextual evidenence. Proceedings of 1999 Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, Maryland,1999, pp.90-99
[11] Zheng Jiaheng, Li Xin, Tan Hongye. The Research of Chinese Names Recognition Method Based on Corpus. Journal of Chinese Information Processing, 2000, 14(1):163-168
[12] Huang Degen, Yue Guangling, Yang Yuanshen. Identification of Chinese Place Names Based on Statistics. Journal of Chinese Information Processing, 2003, 17(2):36-41
[13] Zhang Huapin, Liu Qun. Automatic Recognition of Chinese Personal Name Based on Role Tagging. Chinese Journal of Computer. 2004, 27(1):85-91
[14] Wang Ning, Ge Ruifang, Yuan Chunfa. Company NE identification in Chinese finance news, Journal of Chinese Information Processing, 2002, 16(2):1-6
[15] Zhu Jingbo. Using Co-Trianing for Chinese Organization NE Identification. The 1st International Joint Conference on Natural Language Processing (IJCNLP-04)), Hainan, China, 2003
[16] Lu Yajuan, Zhao Tiejun, et al. Leveled Unknown Chinese Words Resolution by Dynamic Programming. Journal of Chinese Information Processing, 2001,15 (1):123～128