Chinese Organization Name Recognition Using Chunk Analysis

Jihao Yin[1], Xiaozhong Fan[1], Kaixuan Zhang[2], Jiangde Yu[1,3]

[1] Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing, 100081, China;

[2]Liaoning Technical University, Fuxin, Liaoning, 123000, China;

[3] Department of Computer Science, Anyang Teachers' College, Anyang, Henan, 455000, China

yin_jhgg@bit.edu.cn

Abstract. A simplified N-best cascade model is put forward about Chinese organization names automatic recognition. This model can do words segmentation, part-of-speech tag, chunks analysis and Chinese organization names recognition. The N-best cascade method can limit not only the errors propagation but also the search space. In the experiments, we integrate heuristic information and Organization name abbreviations processing into the model to achieve the better experiment results. The last precision and recall of Chinese organization names recognition are 92.31% and 81.01% in IEER99 newswire test set.

Keywords: Chunk analysis; Chinese organization names recognition; N-best cascade model; Heuristic information

1 Introduction

Named entity recognition is a foundational job in natural language processing, is an important pretreatment module too. Generally, the assignment of named entity recognition is identifying person names, location names, organization names, digital and time expression[1] from a text. Person names, location names and organization names are not only difficult but also important.

Many researches about person names and location names have been done, and their precision and recall have meet practical need. But organization names recognition cannot satisfy need whether in theory or in practice[2].

Many scholars have done some researches about Chinese chunk analysis. GuoDong Zhou et al identified English named entities using an HMM-based chunk tagger[3]. But the study of Chinese organization names recognition using chunk analysis technology is a little. Noun-chunk is almost like the structure of Chinese organization names, so it is feasible to identify Chinese organization names using Noun-chunk analysis technology.

Based on Chinese chunk analysis technology and name entity recognition, we put forward a simplified N-best cascade model by mining the characters of Chinese organization names. In the experiments, we added heuristic information and organization name abbreviations processing, which improved the precision and recall of Chinese organization names recognition obviously.

2 Noun-chunk Automatic recognition

2.1 Noun-chunk Definition

Noun-chunk usually includes noun and its former modifier. The former modifier includes number, adjective and noun, but excludes “的” structure. The modifier and center word of “的” structure fall into the different chunks separately[4].

(1) Center word of noun-chunk

The center words of noun-chunk must be nouns. According to the part-of-speech tag criterion of Beijing University, the center words include: noun (n), person name (nr), location name (ns), proper noun (nz), gerund (vn), noun element (Ng), noun abbreviation (j) etc. For example:

[NP 中国/ns 宋庆龄/nr 基金会/n];

[NP 中共/j 北京/ns 市委/n].

(2) Modifier

One noun-chunk cannot include the other noun-chunk. The modifier must include: number phrase, adjective phrase, noun, pronoun. For example:

[NP 第一/m 汽车/n 制造/vn 厂/Ng]

2.2 HMM-based Chunk Analysis

Hidden Markov Model (HMM) is simple and feasible, so it is applied in speech recognition and natural language processing generally. HMM can observe the surface information, but cannot observe the language structure. The chunk analysis is surface and linear, and the chunk interior structure doesn't need be recognized, so identifying chunks using HMM is proper[5].

2.2.1 HMM Theory on Chunk Analysis

Given tagged chunk corpus[4], we can mine some useful linguistics knowledge from the chunk corpus. Then chunks can be identified automatically using these linguistics knowledge.

Given a Chinese words sequence $W(w_1, w_2, \cdots w_m)$, and one of possible part-of-speech tag results is $T(T_1, T_2, \cdots T_m)$. We have mined the linguistics rule-set R, so the chunks recognition can be interpreted to recognize all of chunks $C(r_1, r_2, \cdots r_k)$ from segmented and tagged text *S=(W, T)* using rule-set R. According to statistics, our assignment is to search the most optimal chunk sequence from the given text *S=(W, T).*

$$C^* = \arg\max_C P(C \mid W,T) = \arg\max_{r_i \in R} P(r_1 r_2 \cdots r_k \mid W,T) \qquad (1)$$

Equation (1) can be transformed into Equation (2) using the *Bayes* formula.

$$C^* = \arg\max_C P(C \mid W,T) = \arg\max_C \frac{P(W \mid C,T)P(C,T)}{P(W,T)} \qquad (2)$$

P(W,T) is invariable in the process of chunks analysis because word segmentation and part-of-speech tag have been completed before Chinese chunks recognition. So Equation (2) can be transformed into Equation (3).

$$C^* = \arg\max_C P(W \mid C,T)P(C,T) \qquad (3)$$

P(W|C,T) denotes probability of words sequence in a generative chunk, and it corresponds the observation generative probability in HMM model. It can be established using unigram model, as shown in Equation (4).

$$P(W \mid C,T) \approx \prod_{i=1}^{m} P(w_i \mid m_i, x_i, t_i) \qquad (4)$$

where *m* is chunk boundary coding; *x* is chunk type; *t* is part of speech.

P(C, T) corresponds the states transform probability. It can be established using bigram model, as shown in Equation (5).

$$P(C,T) \approx P(r_1)\prod_{i=2}^{k} P(r_i \mid r_{i-1}) \qquad (5)$$

So the chunk analysis model is a typical HMM model. All of part-of-speech rules compose state set, and each word corresponds its state observation.

2.2.2 Model Estimation

We adopt Maximum Likelihood Estimate (*MLE*) method to estimate each probability of Equation (4) and Equation (5) from the training corpus.

$$P(w_i \mid m_i, x_i, t_i) = \frac{count(w_i, m_i, x_i, t_i)}{count(m_i, x_i, t_i)} \qquad (6)$$

$$P(r_i \mid r_{i-1}) = \frac{count(r_{i-1}, r_i)}{count(r_{i-1})} \qquad (7)$$

where *count(.)* denotes appearance times of parameters in bracket in training corpus.

2.2.3 Chunks Recognition

HMM model decoding usually adopts *Viterbi* algorithm, which is a dynamic programming method, including three parts: initialization, recursion and back-off. The best chunks recognition results can be acquired using *Viterbi* algorithm.

3 Chunk Analysis in Chinese Organization Name Recognition

We adopt chunk analysis technology in Chinese organization names recognition because of the comparability between noun-chunk definition and Chinese organization names structure. We propose a simplified N-best cascade model.

3.1 Algorithm Realization

Given a sequence of Chinese characters, the Chinese organization names recognition process consists of the following four steps:

Step 1: Word segmentations. All possible word segmentations are generated using a Chinese lexicon. We don't explain word segmentation processing because its technology is maturate.

Step 2: Part-of-speech tagging. Part-of-speech tagging use HMM model too. As is shown in Equation (8).

$$T^* = \arg\max_T P(T \mid W) = \arg\max_T P(W \mid T)P(T) \qquad (8)$$

P(W|T) denotes the probability of generative words sequence based on part-of-speech sequence, and it corresponds the observation generative probability in HMM model. It can be established using unigram model, as shown in Equation (4).

$$P(W \mid T) \approx \prod_{i=1}^{m} P(w_i \mid t_i) \qquad (9)$$

P(T) corresponds the states transform probability. It can be established using bigram model, as shown in Equation (5).

$$P(T) \approx P(t_1)\prod_{i=2}^{m} P(t_i \mid t_{i-1}) \qquad (10)$$

$P(w_i|t_i)$ and *$P(t_i| t_{i-1})$* can be acquired using Maximum Likelihood Estimation method from training cor-

pus.

Step 3: Noun-Chunk Recognition (Referring to **section 2.2**)

Step 4: Simplified N-best cascade model is established, which can accomplish words segmentation, part-of-speech tag, chunks analysis unto Chinese organization names recognition. This model can be shown in Equation (11).

$$O^* = \arg\max_{O} P(O \mid S) = \arg\max_{W,T,C,O}[P(W \mid S)P(T \mid W)P(C \mid W,T)P(O \mid W,T,C)] \quad (11)$$

where *W* is the words segmentation result, T is the part-of-speech tag result, C is the chunks analysis result, O is Chinese organization names recognition result. Every probability of this model is acquired in various conditions, and its search space is huge, so this model cannot apply in practice. We must predigest this model, reduce search space, and adjust each probability weight. So we adopt logarithm linear model to solve these problems, which is shown in Equation (12):

$$\begin{aligned} O^* = &\arg\max_{O} \log P(O \mid S) = \arg\max_{W,T,C,O}[\lambda_1 \log P(W \mid S) + \\ &\lambda_2 \log P(T \mid W) + \lambda_3 \log P(C \mid W,T) + \lambda_4 \log P(O \mid W,T,C)] \end{aligned} \quad (12)$$

The most direct predigestion is that only one best result is generated from every phase, which includes words segmentation, part-of-speech tag, chunks recognition and Chinese organization names recognition. But this method can lead the errors propagation easily. The errors in words segmentation, part-of-speech tag and chunks recognition can impact Chinese organization names recognition result[6].

We adopt neutral method, i.e. N-best results are acquired from previous phases, only one best result is generated in Chinese organization names recognition phase. We select words segmentation, part-of-speech tag and chunks recognition which generate the best Chinese organization names recognition as the last results. This method is called N-best cascade, as is shown in Equation (13).

$$\begin{aligned} O^* \approx &\arg\mathrm{N}\max_{W} \lambda_1 \log P(W \mid S) + \arg\mathrm{N}\max_{W_i^* \in W^*} \lambda_2 \log P(T \mid W_i^*) + \\ &\arg N \max_{W_i^* \in W^*, T_i^* \in T^*} \lambda_3 \log P(C \mid W_i^*, T_i^*) + \arg\max_{W_i^* \in W^*, T_i^* \in T^*, C_i^* \in C^*} \lambda_4 \log P(O \mid W_i^*, T_i^*, C_i^*) \end{aligned} \quad (13)$$

where W^* is N-best words segmentation results, T^* is N-best part-of-speech tag results based on N-best words segmentation results, C^* is N-best chunks recognition results based on N-best part-of-speech tag results, *O* is the only best Chinese organization names recognition result based on N-best chunks recognition results.

This method's theory is that N-best results excel one best result in identifying Chinese organization names, so it can limit the errors propagation. And the search space can be reduced greatly because we only search N-best results in each phase.

3.2 Improvement

There are some problems only using the simplified N-best cascade model in Chinese organization names recognition. First, many irrelevant organization candidates will be produced, which results in great search space. Second, the abbreviations of organization names cannot be handled effectively. In the following, we provide solutions about these problems.

3.2.1 Heuristic Information

In order to overcome the redundant candidates generation, the heuristic information is introduced into the simplified N-best cascade model. Chinese organization name keywords list includes 1,355 words[7] (e.g. 大学(university), 公司(corporation)). A keyword and its front two to six words can be considered to construct a Chinese organization name. The heuristic information can limit redundant entities generation. If one organization name doesn't include the words in keyword list, this organization name will be eliminated directly.

3.2.2 Dealing with Abbreviation

We found many errors when identifying organization name abbreviations. Therefore, some strategies are adopted to deal with abbreviations. We use the simple rule-based method to identify organization name abbreviations.

4 Experiments and Analysis

4.1 Evaluation Metric and Data Sets

This model was evaluated in terms of precision (*P*), recall (*R*) and F-measure (*F1*).

$$P = \frac{number\ of\ correct\ responses}{number\ of\ responses} \qquad (14)$$

$$R = \frac{number\ of\ correct\ responses}{number\ of\ all\ organization\ names} \qquad (15)$$

$$F = \frac{(\beta^2+1)\times \mathrm{P}\times \mathrm{R}}{(\beta^2\times \mathrm{P})+\mathrm{R}} \quad (\beta = 1) \qquad (16)$$

The training set is chunk corpus of Microsoft Research Asia, including 502,141 words, 259,843 chunks. In the corpus, Chinese organization name chunks were marked especially.

The test sets select IEER99 newswire test set and "Xinhuanet---Chinese simplified webpages" (http://www.xinhuanet.com) tagged texts, as is shown in Table 1.

Table1. Chinese simplified edition of xinhuanet

ID	Domain	Number of ORG	Size
1	Sports	643	117K
2	Economy	356	103K
3	Politics	227	126K
4	Entertainment	152	108K
5	Education	97	145K
	Total	1475	599 K

4.2 Experiments

We conduct incrementally the following three experiments:

(1) Simplified N-best cascade model, we view the results as the baseline performance;

(2) Integrating heuristic information into (1);

(3) Integrating Cache-based model with (2).

4.3 Experiment Results

4.3.1 Simplified N-best Cascade Model

Based on the basic simplified N-best cascade model, we obtained the baseline performance, as is shown in Table 2. We found that the performance of IEER99 newswire test set is better than "Xinhuanet" test set because "Xinhuanet" test set is much bigger than IEER99 newswire test set ("Xinhuanet" test set includes 1475 organization names, but IEER99 test set includes 497 organization names).

Table 2. Test 1 experiment results

Test Set	Precision(%)	Recall(%)	F1(%)
IEER99	75.23	68.86	71.90
Xinhuanet	71.19	59.97	65.10

4.3.2 Integrating Heuristic Information

After heuristic information is added to the experiment, we found that decoding became simple and the precision of organization names recognition was improved, as is shown in Table 3. For example, the precision of organization names recognition in IEER99 test set increases from 75.23% to 87.31%. The reason is that integrating heuristic information reduces the noise influence.

Table 3. Test 2 experiment results

Test Set	Precision(%)	Recall(%)	F1(%)
IEER99	87.31	64.22	74.01
Xinhuanet	89.52	60.37	72.11

However, we noticed that the recall of organization names recognition decreased. The reason is organization names without organization ending keywords were not identified. For example, “中国保险 (China assurance)” is the abbreviation of “中国保险公司 (China assurance corp.)”, so it cannot be identified.

4.3.3 Dealing with Organization Names Abbreviation

In this experiment, we integrate the organization name abbreviations processing, the results are shown in Table 4. Comparing Table 3 and Table 4, we found the recall of organization names recognition in IEER99 test set increases from 64.22% to 81.01%.

Table 4. Test 3 experiment results

Test Set	Precision(%)	Recall(%)	F1(%)
IEER99	92.31	81.01	86.29
Xinhuanet	90.75	76.92	83.26

4.3.4 Experiment Results and Analysis

From above experiments, we can make some conclusions: (1) The simplified N-best cascade model can select the front N-best results as the candidates of the next phase, so the errors propagation will be limited. Moreover, N-best results were searched in each phrase, which can reduce the search space greatly. (2) Integrating heuristic information, i.e. using keywords list and some simply rules, can restrict redundant organization names generation. So the precision of organization names recognition increased obviously. (3) The organization name abbreviations processing can improve the recall of organization names recognition.

5 Conclusions and Future Work

Based on HMM model, Chinese words segmentation, part-of-speech tag, chunks analysis and Chinese organization names recognition have been integrated into a unified framework, then we establish a simplified N-best cascade model. In the experiments, integrating heuristic information increases the precision of organization names recognition; integrating organization name abbreviations processing increases the recall of organization names recognition. In the IEER99 newswire test set, the last precision and recall of Chinese organization names recognition are 92.31% and 81.01%, each of values is higher about 9% than precision and recall of reference [8].

In the process of organization names recognition, one organization name usually appears in the different style. This problem is named entity coreference. So we will focus more on name entity coreference. Furthermore, we intend to extend our model to the specific domain for proper names recognition.

Acknowledgment. This paper is supported by the Doctor Foundation project 20050007023 of China Ministry of Education.

References

1. Sun Bin. A summarization of information extraction[J]. Language Information Processing, 2003, 2(1): 34-35.
2. Yu Hongkui, Zhang Huaping. Recognition of chinese organization name based on role tagging[A]. Proceedings of 20th International Conference on Computer Processing of Oriental Languages[C]. Beijing, China, 2003, 79-87.
3. GuoDong Zhou, Jian Su. Named entity recognition using an HMM-based chunk tagger[A]. Proceedings of the 40th Annual Meeting of ACL[C]. Philadelphia, America, 2002, 473-480.
4. Li Hongqiao. Chinese chunking and its applications[D]. Beijing: Beijing Institute of Technology, 2004.
5. L. R. Rabiner, B. H. Juang. An introduction to Hidden Markov Models[J]. IEEE ASSP Magazine, 1986, 3(1): 4-16.
6. Sabine Buchholz, Jorn Veenstra, Walter Daelemans. Cascaded grammatical relation assignment[A]. Proceedings of EMNLP/VLC-99[C]. University of Maryland, USA, 1999, 239-246.
7. Jian Sun, Jianfeng Gao. Chinese named entity iIdentification using class-based language model[A]. Proceedings of the 19th International Conference on Computational Linguistics[C]. Taipei, China, 2002, 24-25.
8. Youzheng Wu et.al. Chinese named entity recognition combining a statistical model with human knowledge[A]. Proceedings of the Workshop on Multilingual and Mixed-language Named Entity Recognition[C], Japan, 2003, 65-72.