A Pre-Identification Method for Chinese Named Entity Recognition

Liu Hongjian
Hitachi(China) Research & Development Corporation, Shanghai, China
Email: hjliu@hitachi.cn

Guo Defeng
Hitachi(China) Research & Development Corporation, Shanghai, China
Email: dfguo@hitachi.cn

Zhou Quan
Hitachi(China) Research & Development Corporation, Shanghai, China
Email: quanzhou@hitachi.cn

Nagamatsu Kenji
Hitachi, Ltd., Central Research Laboratory, Tokyo, 185-8601, Japan

Sun Qinghua
Hitachi, Ltd., Central Research Laboratory, Tokyo, 185-8601, Japan

***Abstract*—In this paper, a pre-identification method for Chinese named entity recognition is proposed. Internal information of entity name like family name, first name in person name, feature word in place name and organization name do not needed. Through entity name guessing based on context keywords, pre-identification is realized. Definition of bidirectional potential entity name recognition, rough confirmation of potential entity name, segmentation word is proposed. To solve the possible ambiguity in entity name identification, the degree of segmentation and conjunction is presented as well as cascade recognition and final confirmation. Combining with this pre-processing method, performance will be improved by using internal information of entity name. Experiment proves that the method have a special advantage in recognition special entity name, ambiguity name and irregular name. In this paper, Chinese person name is taken as an example for entity name recognition. Nevertheless, the method is not limit to person name recognition but also a pre-identification method for other entity name.**

***Index Terms*—named entity recognition, potential entity name, context keywords, segmentation degree, conjunction degree**

I. INTRODUCTION

Chinese Named Entity(NE) recognition is to recognize specific entities in text. It is the key technique in many Chinese information processing applications such as information extraction, machine translation, question answering, etc. But because of the property of Chinese language itself, it is difficult to recognize Chinese named entity.

In named entity recognition, person name, place name and organization name are three most important parts[1-3], which play an important role in Chinese lexical and syntactic analysis and widely applied in systems such as Chinese speech synthesis, Chinese speech recognition, machine translation and information retrieval.

Difficulties in Chinese Named EntityRrecognition

Chinese named entity name means a name exists in the form of Chinese language. For person name name, Besides Han People(China's main nationality) name, Chinese name also includes pen name, stage name, minority person name and foreign person translation name etc. In most current literatures, Chinese name is often recognized by using internal name information like family name, single first name, double first name or given name. However, corpus cover only a small scale(about 2%-5%) names while person name is unlimited in practice. Taking "张大海" as an example, quantity of "大海" as a person name is very small while is very large as a non-person name which often lead to recognition by mistake.

For Chinese person name recognition, there are some kind of rules can be utilized such as "Family name + Single first name" or "Family name + Double first name". These kind of information can only be available for some parts of Chinese name[4]. However, there exist large quantities of irregular Chinese name. For example, a name has a first name but no family name such as "大海同学是一位好同学", double family name such as "陈方安生", honorific title such as "郭老，陈总"，pen name and stage name such as "巴金", "三毛", "琼瑶", single

© 2010 ACADEMY PUBLISHER
doi:10.4304/jsw.5.1.73-80

person name such as “刘，关，张桃园三结义”, all of them bring a large difficulties for Chinese name recognition.

For minority person name and foreign translation name, there still exist many non-standard translation name, mis-translation name although there are some translation standards for them. For example, “Bush” is often translated into different names like “布什”, “布希” and even “布殊”. In some other condition, there are also many non-standard or error input in handwriting. For example, in the sentence “日本首相麻生太郎访问中国”, “麻生太郎” is sometimes miswritten into “麻省太郎” which will lead to “麻省太郎” can not be recognized because “麻省” does not exist in corpus at all. Also, in some other condition, there are also many print slur phenomenon in presswork. For example, in above same sentence, “麻生太郎” is mis-print into “麻省大郎” or “麻生□郎” for print problem. All the above problems introduce a large challenges for Chinese name recognition.

For Chinese place name recognition and organization name recognition, the same problems will also occur in the application. The data sparse and ambiguity is much more serious in place name and organization name. For example, in the sentence “他终于来到了西玛特尔”, “西玛特尔” is a translation place name but never appear in the corpus. In another example “一等奖由在华中师范任教的美籍教师获得”, “华中师范” is an organization name. In both “西玛特尔” and “华中师范”, there are no internal information like “村”, “镇”, “大学”, “学院” for reference which will bring the difficulty for recognition.

Current Solutions and Disadvantages

Aim at these problems in Chinese named entity recognition, many literatures are proposed in which three kinds of methods are mainly included which are rules method[5,6], statistic method[7-9] and hybrid method[10].

In literature [4], all words are divided into 15 roles according to their functions. For each roles, information will be statistic from corpus. Viterbi algorithm is adopted for calculated the best role sequence of one sentence. Some rules between family name and first are used to determine final person name.

Foreign person names and Chinese person names recognition method are proposed in literature[11]. In the paper, person name and word segmentation are processed at the same time. Suitable weight of edge in directed graph according to statistic information is calculated to find best segmentation path in the graph and correct name will be recognized at the same time. In the paper, reliability of a name is utilized to describe relationship existing in internal name information of both Chinese and Foreign person name.

No matter rules methods or statistic methods, analysis is still focus on internal name information. These methods will obtain a good performance normally for regular name while bad performance for irregular name. As introduced above, in practice, there are many irregular names, foreign translation names and all kinds of misprint names which will lead to absence of internal name information in corpus. These problems largely limit the practical application of present methods with only using internal name information in Chinese name recognition.

Based on above disadvantage of current research, cascade Chinese named entity recognition is proposed in this paper. Chinese potential entity name is guessed by using context keywords without over dependant internal information like family name, given name or place feature etc. The experiment proves a high precision of the method especially an advantage in foreign translation name, rare name, special name, combinational name, ambiguity name, irregular name etc. which is a effective pre-processing method for Chinese named entity recognition.

In this paper, to describe the method conveniently, person name is taken as an example to explain the whole process. However, it does not means the method is only limited in person name recognition. In application, the method can be easily generalized to place name and organization name.

II. CHINESE POTENTIAL NAMED ENTITY RECOGNITION

Hypothesis for person name keywords

Although there exist large quantity irregular names in our daily life, an important phenomenon can be found that person names are always appear with some special context keywords. In many cases, the role of keyword is even more important than person name itself. For example, in the sentence “国家主席×××致电印度总统×××”, “×××”can be predicted as a person at the first glance without knowing the content. According to this general knowledge, the reason that name given is used for easy to remember and differentiate from other names. The name is seldom given similar with its context keywords. For example, in the sentence “这是张小明记者”, “张小明”is a keywords which is easy to be differentiated from “记者”. A name is seldom given like “张记者”, because “张记者” is easy to confused with its keywords “记者” in the sentence “这是[张记者]记者”. Based on this common knowledge, an important hypothesis is proposed as follows,

Hypothesis: The larger of probability the word is taken as a name keywords, the smaller is taken as a person name itself.

Hypothesis1 reveals that name recognition can be analyzed through its context keywords indirectly. For traditional method, corpus will be divided into name sets and non-name sets through which the information will be statistic from name sets. However, the infinite of person name and finite corpus often lead to difficulties for name recognition especially for irregular name. According to hypothesis 1, the analysis set can be changed into form of Figure 1(b) in which non-name sets and name sets will be transformed into keywords sets and non-keywords sets. Because quantity of name keywords is limited and stable, Figure 1(b) will provide us an useful tools to analyze infinite person name.

© 2010 ACADEMY PUBLISHER

However, only context name is not enough for name recognition. For example, both “他是张小明记者” and “他是人民日报记者” belong to “他是×××记者”, context keywords are the same while “人民日报” is not a person name. In another example, how we distinguish “他是张小明记者”and“他是科切里尔记者”? Both of them are also “他是×××记者” in which “张小明” and “科切里尔” are both person name. However, the name like “科切里尔”maybe does not exist in corpus at all. Here, the conception of potential name is propose to solve such kind of problems.

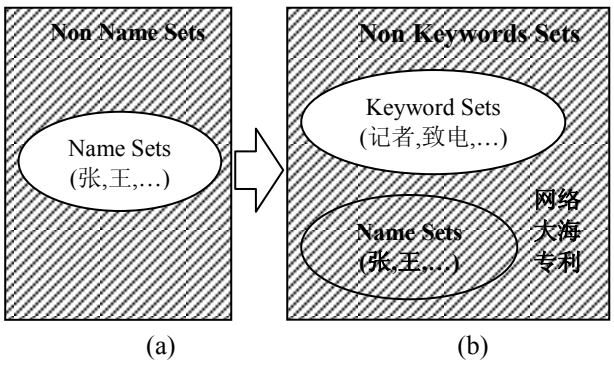

Figure1. Name Analysis Sets

Potential Entity Name

Here, conceptions of potential name is provide.

Definition1：The chunk surrounded by keywords is named a potential name.

Definition2：Keywords before potential name are called predecessor keywords or startup keywords.

Definition3：Keywords after potential name are called successor keywords or end keywords.

Potential name is a chunk which may be a person name. Two information can be find in definition 1. Relative to name recognition, definition 1 provide us a necessary condition. That is to say, the chunk may be a person name if keywords are found. Definition 1 is not a sufficient condition which we can confirm a part must not be a person name if it does not surrounded by keywords. For example, in the sentence “这是百货大楼”, “是” is a startup keyword while “大楼” is not a end word, therefore, “百货” must not be a person name.

According to above definitions, potential name can be guessed at first though which can largely cut down the search scale.

Bidirectional Potential Name Recognition

Forward potential name recognition

Suppose $W = w_1 w_2 \cdots w_n$ is Chinese word string after word segmentation. Suppose w_i is a startup keyword and l is maximum length for potential name analysis. $S = w_{i+1} w_{i+2} \cdots w_{i+j} (j \le l)$ is the part of $W = w_1 w_2 \cdots w_n$ for potential name analysis. For $S = w_{i+1} w_{i+2} \cdots w_{i+j}$, w_i is its predecessor keyword and w_{i+j+1} successor keyword.

Hypothesis 1 can be expressed as follows:

$$j' = \arg\max_{j \in [1,l]} P(K_B \mid w_i) P(K_E \mid w_{i+j+1}) P(\overline{K} \mid S) \quad (1)$$

in which K_B is predecessor keyword and K_E successor keyword. $P(K_B \mid w_i)$ denotes probability of w_i as a predecessor keyword. $P(K_E \mid w_{i+l+2})$ denotes probability of w_{i+l+2} as a successor keyword. $P(\overline{K} \mid S)$ means probability of S not as a keywords. Equation (1) means when S as a potential person name, probability of its context keyword will reach maximum while probability of S will reach minimum as a person name keywords. In equation (1), $P(K_B \mid w_i)$ can be omitted for no related with j.

In equation (1),

$$P(\overline{K} \mid S) = \prod_{j=1}^{l} P(\overline{K}_{i+j} \mid w_{i+j}) = \prod_{j=1}^{l} [1 - P(K_{i+j} \mid w_{i+j})]$$

According to independence hypothesis（1）will be turn into,

$$j' = \arg\max_{j \in [1,l]} P(K_E \mid w_{i+j+1}) \prod_{j=1}^{l} [1 - P(K_{i+j} \mid w_{i+j})] \quad (2)$$

$P(K_{i+j} \mid w_{i+j})$ denote probability of w_{i+j} as a keyword.

Backward potential name recognition

From equation（2）, $P(K_B \mid w_i)$ is omitted for no related with j. However, in practice, predecessor keyword K_B is also very important. To full use the information of K_B, backward potential name recognition is proposed.

According to（1）, the search direction change to backward instead of forward, let $i = i + j + 1$

$$j' = \arg\max_{j \in [0,l]} P(K_B \mid w_{i-j-1}) P(K_E \mid w_i) P(\overline{K} \mid S) \quad (3)$$

In which，$P(K_E \mid w_i)$ can be omitted with no related with j.

Equation (2) will change into：

$$j' = \arg\max_{j \in [0,l]} P(K_B \mid w_{i-j-1}) P(\overline{K} \mid S)$$

Bidirectional potential name recognition

From (2) and (3), two potential name recognition results can be obtained from forward and backward respectively, but which result is better?

Suppose original word string is $W = w_1 w_2 \cdots w_n$, forward recognition result is $W_f = w_1 \cdots [w_i \cdots w_{i+l}] \cdots w_n$ and backward recognition result is $W_b = w_1 \cdots [w_j \cdots w_{j+t}] \cdots w_n$.

To avoid errors introduced from whole sentence, method of maximum chunk analysis is proposed to compare the two results. If $[w_i \cdots w_{i+l}]$ and $[w_j \cdots w_{j+t}]$ overlap, maximum coverage area of $[w_i \cdots w_{i+l}]$ and $[w_j \cdots w_{j+t}]$ is called maximum chunk. Let $i < j < i + l < j + t$, then maximum chunk is $[w_i \cdots w_{j+t}]$.

Here we define：

© 2010 ACADEMY PUBLISHER

$$P=\prod_{h=i}^{j+t}P(\overline{K}\mid w_h\in PR)P(\overline{R}\mid w_h\notin PR) \quad (4)$$

in which $w_h\in PR$ means w_h is tagged as potential name while $w_h\notin PR$ not as a potential name.

in which $P(\overline{K}\mid w_i)=1-P(w_i=K_E\cup K_B)$

$P(\overline{R}\mid w_i)=1-P(w_i=R)$

According to equation (4), suppose W_1 is the tagging result of forward recognition corresponding to maximum chunk and probability corresponding to W_1 is P_f. In the same, W_2 is the tagging result of backward recognition corresponding to maximum chunk and probability corresponding to W_2 is P_b

If $P_f\geq P_b$, W_1 will be regarded as the correct result and W_2 will be regarded as the correct result if $P_f<P_b$.

Take the sentence "这是王小明的父亲王大明" as an example,

Forward recognition result is "这是[王小明]的父亲[王大明]" and backward recognition result is "这是王小[明的父亲王]大明".

We can find that there is an overlap in [明的父亲王] and [王小明] [王大明]. Then the maximum chunk will be "王小明的父亲王大明".

Forward analysis result W_1 corresponding to maximum chunk is "[王小明]的父亲[王大明]"

Backward analysis result W_2 corresponding to maximum chunk is: "王小[明的父亲王]大明"

According to equation (4),

$P_f=P(\overline{K}\mid 王)P(\overline{K}\mid 小)P(\overline{K}\mid 明)P(\overline{R}\mid 的)$

$P(\overline{R}\mid 父亲)P(\overline{K}\mid 王)P(\overline{K}\mid 大)P(\overline{K}\mid 明)$

$P_b=P(\overline{R}\mid 王)P(\overline{R}\mid 小)P(\overline{K}\mid 明)P(\overline{K}\mid 的)$

$P(\overline{K}\mid 父亲)P(\overline{K}\mid 王)P(\overline{R}\mid 大)P(\overline{R}\mid 明)$

In this example $P_f>P_b$, then W_1 is the correct recognition result.

Rough Confirmation of Potential Name

Through bidirectional potential name recognition, the most possible potential name will be found corresponding to context keywords. For example, in the sentence "这是人民日报记者", "人民日报" will be recognized as a potential name. However, this result is only obtained corresponding to its keywords "是" and "记者". To judge the probability between a potential name and non-name, rough confirmation of potential name is needed.

Suppose $W=w_1w_2\cdots w_n$, potential name recognition result is $W_{nr}=w_1\cdots[w_i\cdots w_{i+l}]\cdots w_n$ while non-name result is $W_{\overline{nr}}=w_1\cdots w_i\cdots w_{i+l}\cdots w_n$.

When $w_i\cdots w_{i+l}$ is recognized as a person name, according to hypothesis (1), its probability not taken as a name keyword is:

$$P_{nr}=P(\overline{K}\mid w_j)\cdots P(\overline{K}\mid w_{j+l}) \quad (5)$$

in which $P(\overline{K}\mid w_j)=1-P(w_j=K)$

When $w_i\cdots w_{i+l}$ is recognized as a non-name, its probability as a non-name is:

$$P_{\overline{nr}}=P(\overline{R}\mid w_j)\cdots P(\overline{R}\mid w_{j+l}) \quad (6)$$

in which $P(\overline{R}\mid w_j)=1-P(w_j=R)$

Then $w_i\cdots w_{i+l}$ is regarded as a potential name if $P_{nr}\geq P_{\overline{nr}}$ while not a potential name if $P_{nr}\leq P_{\overline{nr}}$.

Take "人民日报" as an example,

$P_{nr}=P(\overline{K}\mid 人民)P(\overline{K}\mid 日报)$

$P_{\overline{nr}}=P(\overline{R}\mid 人民)P(\overline{R}\mid 日报)$

Here $P_{nr}<P_{\overline{nr}}$, then "人民日报" is not a potential name.

From above analysis, we can find the target of rough confirmation for potential name is to delete parts which must not be a person name while remain parts which possible a person name.

Cascade Entity Name Recognition

For some sentences, some large chunks may be misrecognized using above methods. For example, "国务院/总理/[朱镕基今天主持召开国务院第二十七次常务会议]/nr" (here /nr is a tagging mark for person name) in which "朱镕基今天主持召开国务院第二十七次常务会议" is mis-recognized as a person name for influence of keyword "总理" and space. However, this result in fact is correct in a sense according with our imagination because "×××"will be recognized as a person name by our mind if there is no other information in the sentence "国务院总理×××". The reason that chunk is recognized by mistake because "×××" is more like a person name from whole chunk than from the part. To recognized the detail information hidden in the whole chunk, cascade person name recognition is proposed.

"朱镕基今天主持召开国务院第二十七次常务会议" is extracted as a separated sentence for potential name recognition once again. "[朱镕基]/nr 今天主持召开国务院第二十七次常务会议"can be recognized in further processing because information of "今天" "主持" are full used in this sentence.

"朱镕基" will also continue to be processed as a potential name in circulation but lead to a stop of recursion for invariability of "朱镕基".

Segmentation Word of Potential Name

In some other cases, influence of keyword is much smaller than space at the beginning and end of the sentence which will lead to suspend of cascade potential name recognition. For example, cascade potential name recognition will be suspended in the sentence [胡锦涛积极评价中秘建交以来两国关系的发展]/nr". The whole sentence will be always recognized as a person name without continue to further processing. To solve this problem, segmentation word is introduced.

© 2010 ACADEMY PUBLISHER

Here, a name keyword with maximum probability will be called segmentation word,

$$P(K \mid w_i) = \max_{k=[1,n]} P(K \mid w_k) \quad (7)$$

In above sentence, "评价" is keyword with maximum probability. Then the sentence will be segmented into two parts "胡锦涛积极" and "中秘建交以来两国关系的发展". Based on these two parts, potential name recognition is applied respectively. Then "[胡锦涛积极]/nr" and "中秘建交以来两国关系的发展" will be obtained. "[胡锦涛]/nr" can be finally obtained by using cascade potential name recognition based on"[胡锦涛积极]/nr".

Segmentation Degree and Conjunction Degree

Above method of using segmentation word will be suitable on the condition that a person name in which does not include keywords. However, some words can be taken as both keywords and inside word of a person name. For example, in the sentence "俄罗斯选手[玛别特洛娃和阿吉洪诺夫]/nr 获得冠军". Without prior knowledge, even people will guess "×××" is a person name in "俄罗斯选手×××获得冠军". However, Chinese people often think "玛别特洛娃和阿吉洪诺夫" is two person "玛别特洛娃" and "阿吉洪诺夫". While "徐和轩" in "这是同事[徐和轩]/nr 先生" is regarded as one person name instead of "徐" and "轩".

Here, length is regarded as an important factor for person name recognition. For the word which be taken as both keywords and inside word of a person name, we called ambiguity words (PAM). Based on the conception, the segmentation degree and conjunction degree is proposed.

$PAM_{nr}^{f}(w)$ is number of word before ambiguity word w in a person name.

$PAM_{nr}^{p}(w)$ is number of word after ambiguity word w in a person name.

$PAM_{nr}^{f}(w)$ is number of word before ambiguity word w as a keyword of person name.

$PAM_{nr}^{p}(w)$ is number of word after ambiguity word w as a keyword of person name.

$\vec{v}_j(w) = (PAM_{nr}^{f}(w), PAM_{nr}^{p}(w))$ is defined a conjunction vector for PAM.

$\vec{v}_s(w) = (PAM_{nr}^{f}(w), PAM_{nr}^{p}(w))$ is a segmentation vector of PAM.

$P(\vec{v}_j \mid w)$ is probability of conjunction vector $\vec{v}_j$ of w as a person name.

$P(\vec{v}_s \mid w)$ is probability of segmentation vector $\vec{v}_s$ as a keyword of person name.

Suppose $\vec{v}(w) = (PAM^{f}(w), PAM^{p}(w))$ is a vector which needs to be recognized. In which, $PAM^{f}(w)$ and $PAM^{p}(w)$ is the number of words in potential name before and after w respectively.

Base on above conceptions, the conjunction degree will be defined as,

$$d_j = \cos(\vec{v}(w), \vec{v}_j^{a}(w)) \bullet P^{a}(\vec{v}_j \mid w)$$

in which, $a = \arg\min_{k=[1,n]} \left\| \vec{v}(w) - \vec{v}_j^{k}(w) \right\|$

the segmentation degree will be defined as,

$$d_s = \cos(\vec{v}(w), \vec{v}_s^{b}(w)) \bullet P^{b}(\vec{v}_s \mid w)$$

in which, $b = \arg\min_{t=[1,m]} \left\| \vec{v}(w) - \vec{v}_s^{t}(w) \right\|$

here, $\vec{v}_j^{k}(w)$ denotes k th conjunction vector for PAM as a person name while $\vec{v}_s^{t}(w)$ denotes t th segmentation vector as a keyword of person name.

Then the principle to determine whether a PAM will be segmented or not is proposed as follows:

if $d_j \geq d_s$ potential name will not be segmented.

if $d_j < d_s$ potential name will be segmented into two parts.

Here, it should be noted that only two parts on either side of PAM are the potential name, the degree of segmentation and conjunction will be applied. It will not be applied, if the condition is not satisfied. For example, in the sentence "[玛别特洛娃和朋友们笑了]/nr", "[玛别特洛娃]/nr" and "朋友们笑了" will be segmented based on PAM "和". However, in this sentence, only "[玛别特洛娃]/nr" is a potential name. In this case, degree of segmentation and conjunction is not used while "[玛别特洛娃]/nr" will be processed directly and "朋友们笑了" is regarded as not a person name.

Final person name confirmation

Above description mainly focus on potential name recognition. In potential name recognition, the main idea is to remain possible part while to delete the part which surely a non-name. Process still needs to be done to confirm whether the potential name is a person name or not.

For example, in "[我谨]/nr 代表中国和中国人民", "我谨" is recognized as a potential name which is according with human mind because "×××" is often regarded as a person name in sentence "×××代表中国和中国人民". However, "我谨" will be easily excluded from person names when people find the concrete content of "我谨代表中国和中国人民" while "秋谨"(a famous Chinese person name in history) is easily regarded as a person name. One of most important reason is relationship of "我谨" in non-name corpus is larger than name corpus while relationship of "秋谨" in name corpus is larger than non-name corpus.

Suppose original word string is W, potential recognition result is $W_{nr} = w_1 \cdots [w_i \cdots w_{i+l}] \cdots w_n$, our target is to confirm whether $w_i \cdots w_{i+l}$ is a person name or not.

Suppose $w_i \cdots w_{i+l}$ is recognized as a person name, from 1 order Markov Assumption:

$$P(w_i w_{i+1} \cdots w_{i+l})_{nr} = P(w_i)_{nr} P(w_{i+1})_{nr} \cdots P(w_{i+l})_{nr} \quad (8)$$

Suppose $w_i \cdots w_{i+l}$ is recognized as a non-name，

$$P(w_i w_{i+1} \cdots w_{i+l})_{\overline{nr}} = P(w_i)_{\overline{nr}} P(w_{i+1})_{\overline{nr}} \cdots P(w_{i+l})_{\overline{nr}} \quad (9)$$

© 2010 ACADEMY PUBLISHER

Here, we find the (8)(9) is very similar, but they are quite different. nr denotes statistic result from person name corpus while $\overline{nr}$ from non-name corpus.

Combining with the keywords of $w_i \cdots w_{i+l}$:

$$P_{nr} = P(K_B)_{nr} P(w_i w_{i+1} \cdots w_{i+l})_{nr} P(K_E)_{nr} \quad (10)$$

$$P_{\overline{nr}} = P(K_B)_{\overline{nr}} P(w_i w_{i+1} \cdots w_{i+l})_{\overline{nr}} P(K_E)_{\overline{nr}} \quad (11)$$

in which K_B, K_E is predecessor keyword and successor keyword respectively. $P(K_B)_{nr}$ is probability of K_B as a predecessor keywords of person name and $P(K_B)_{\overline{nr}}$ is probability of K_B as a predecessor keywords of non-person name. Similar, $P(K_E)_{nr}$ is probability of K_E as a successor keywords of person name and $P(K_E)_{\overline{nr}}$ is probability of K_E as a successors keywords of non-person name.

According to (10) and (11), confirm principle is shown as below:

If $P_{nr} \geq P_{\overline{nr}}$, $w_i \cdots w_{i+l}$ is regarded as a person name.

If $P_{nr} < P_{\overline{nr}}$, $w_i \cdots w_{i+l}$ is regarded as a non-person name.

For example, “两国之间” is recognized as a potential name in sentence“[两国之间]/nr 的友好交往源远流长” which only tell us that “×××” more possible is a person name in “××× 的友好交往源远流长”. However, according to (8)-(11), we know that relationship of “两国之间” as a non-name corpus is much larger than in name corpus. “两国之间” will be confirmed as a person name finally. However, in some other cases such as foreign translation name, for example, potential name such as “萨纳布里亚” does not exist in both name corpus and non-name corpus in the sentence “秘鲁大使[萨纳布里亚]/nr 等参加了会见”. However, we still can guess “×××” is a person name in “秘鲁大使××× 等参加了会见” in absence of internal information of “×××” because the keywords like “大使”, “等”, “会见” in contexts provide more information for potential name recognition which also prove advantage of our method.

Research Flow

Figure 2. is the system flow of cascade potential Chinese entity name recognition. From figure2, bidirectional potential will be first applied for identifying the potential name. With rough confirmation of potential name, the part which is surely not a entity name will be excluded first. Segmentation word will be acted on the potential name repeatedly until the circulation process terminated. During the process, if there exist ambiguity words, the cascaded recognition will be processed according to segmentation and conjunction degree. Through above procedure, the part which must not be a entity name will be excluded continuously while the part which maybe a entity name will be kept. Through final confirmation model, entity name will be confirmed to determine whether the part is a real name or not.

III EXPERIMENT AND EVALUATION

Evaluation for Potential Name Recognition

In following experiment, corpus of “2002 People’s Daily” corpus is taken as training samples which totally include 12 month news (about 26 million Chinese word). In the corpus, name of Han people is tagging as “family name /nrf” and “first name /nrg”. For example, “记者/n 罗/nrf 昌爱/nrg 报道/v”. Because in this paper, the topic is to discuss potential name which does not need information of family name and first name, /nrf and /nrg are therefore combined together. For example, the above

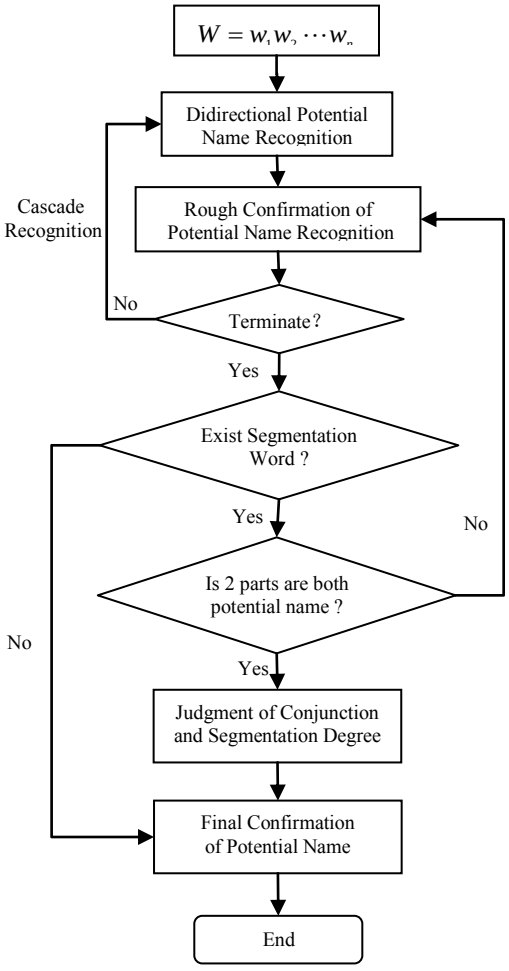

Figure.2 Cascade Potential Name Recognition

sentence will be merged into one name “记者/n 罗昌爱/nr 报道/v”. For foreign name such as “印度共和国/ns 总统/n 科切里尔·拉曼·纳拉亚南/nr 阁下/n ”, the link point is deleted automatically to turn into 印度共和国/ns 总统/n 科切里尔拉曼纳拉亚南/nr 阁下/n ”. The concrete tagging specification can be found in reference the “Specification for Corpus Processing at Peking University” [12].

For open test, January corpus of “1998 People’s Daily” are taken as a test sample.

© 2010 ACADEMY PUBLISHER

For Chinese name recognition, there are three criterions to evaluate performance, precision(P), recalling rate(R) and F-value(F) shown as follows,

$$P=\frac{Number\ of\ correct\ tagged\ person\ name}{Number\ of\ tagged\ person\ name}\times 100\%$$

$$R=\frac{Number\ of\ correct\ tagged\ person\ name}{Number\ of\ person\ name\ in\ corpus}\times 100\%$$

$$F=\frac{(\beta^2+1)PR}{\beta^2 P+R}\times 100\%$$

in which β is an adjustment factor and is 1 in this paper.

TABLE I
EXPERIMENTAL RESULT OF POTENTIAL NAME RECOGNITION

Type	P	R	$F-1$
Open Test	90.12%	81.45%	85.57%
Close Test	93.09%	81.33%	86.81%

From the table, we can find a good performance obtained by using potential name recognition.

To find the performance of degree of segmentation and conjunction, here the experimental result is listed separately. Keep the sample of training and test, here, the function of segmentation and conjunction is added based on above experiment as table 1.

TABLE II
EXPERIMENTAL RESULT OF WITH ESTABLISHING SEGMENTATION AND CONJUNCTION DEGREE

Type	P	R	$F-1$
Open Test	91.70%	84.30%	87.84%
Close Test	94.40%	89.22%	91.74%

From tableII, we can find the precision is increased for about 1%, the recall rate obtain 3-8% and F increase 2-5%.

Evaluation by Adding Internal Name Information Based on Potential Name Recognition

Potential name recognition proposed in the paper can be served as a rough recognition method for other name recognition method using internal name information like family name and first name. For example, internal information of Chinese person name can be utilized. According to statistic, quantity of Chinese person name account for four fifths in whole Chinese name. Therefore, we can imagine that system performance will be better if combining two kinds methods together based on rich internal information existed in Chinese person name.

To explain the performance, information of Chinese family name and first name is also statistic from corpus. HMM is adopted for Chinese person name recognition based on potential name recognition in the paper. Following is the experiment result,

TABLE III
EXPERIMENTAL RESULT ADDING INFORMATION OF CHINESE PERSON NAME

Type	P	R	$F-1$
Open Test	91.23%	82.30%	86.54%
Close Test	93.13%	90.21%	91.65%

From above table, we find the precision is increased 1% or so, recalling rate 1-9% and F 1-5% after using internal information of Chinese person name.

The same to describe the performance of degree of segmentation and conjunction, function of segmentation and conjunction is added based on above experiment as table 3.

TABLE IV
EXPERIMENTAL RESULT ADDING INFORMATION OF CHINESE PERSON NAME WITH ESTABLISHING SEGMENTATION AND CONJUNCTION DEGREE

Type	P	R	$F-1$
Open Test	92.60%	85.10%	88.69%
Close Test	94.30%	92.22%	93.25%

From table IV, we can find the precision obtain the highest performance with using the both internal information and 2-5% and establishing segmentation and conjunction degree.

Here, Chinese person name is only an example. In other examples, there are also exist lots of regular information in many foreign translation names. Combining these kind of information with potential name recognition method, better improvement will be expected which will be discussed in our future work.

IV CONCLUSION

A pre-processing method for Chinese named entity recognition is discussed in the paper without considering internal information of NE. Chinese potential entity name can be guessed guiding by its context keywords. Based on forward potential name recognition and backward potential name recognition, bidirectional potential NE recognition is propose to recognize potential NE. Also rough confirmation of potential entity name, the degree of segmentation and conjunction is presented. Experiment proves a good performance on the NE recognition in foreign translation NE, special NE and irregular NE. The method can be taken as a pre-processing or rough processing method. Better performance can be obtained combining with internal name information like family name in person NE or some feature in place and organization name which from other algorithm. Experiments proves the method is an effective tool for Chinese named entity recognition.

REFERENCE

[1] Sun Maosong, et. Automatic recognition of Chinese personal names. Journal of Chinese Information Processing, 1994,8(2).

© 2010 ACADEMY PUBLISHER

[2] Huang Degen, Sun Yinghong. Automatic Recognition of Chinese Place Names, Computer Engineering, 2006.v32. 219-222.
[3] Zhou Junsheng , Dai Xinyu , Yin Cunyan , Chen Jia Jun, Automatic Recognition of Chinese Organization Name Based on Cascaded Conditional Random Fields. Chinese Journal of Electronics. 2005,34(5), 804-809.
[4] Zhang Huaping, Liu Qun. Automatic Recognition of Chinese Personal Name Based on Role Tagging, CHINESE JOURNAL OF COMPU TERS, Vol.27 No.1, Jan2004, p85-91
[5] D.Farmakiotou, V.Karkaletsis. Rule-based Named Entity Recognition for Greek Financial Texts. COMLEX2000;75-78.
[6] Guodong Zhou, Jian Su. Named Entity Recognition using an HMM-based Chunk Tagger. ACL, PhiladelPhia, USA, 2002:473-480.
[7] A. Borthwick. Maximum Entropy Approach to Named Entity Recognition. PhD. Dissertation, NewYork University, 1999:18-25.
[8] Hai Leong Chieu, Hwee Tou Ng. Named Entity Recognition: A Maximum entropy Approach Using Global Information, COLING, TaiPei, Taiwan, 2002.
[9] K.Takeuchi, N.Collier. Use of Support Vector Machine in Extended Named Entity Recognition. The 6th Conference on Natural Language Learning,2002: 119-125.
[10] Wang Sheng, Huang Degen,Yang Yuansheng. Chinese person name recognition based on mixture of statistic and rules. Corpora of Computational Linguistics. Beijing: Tsinghua University Press, 1999.
[11] Gao Hong, Huang Degen, Yang Yuansheng. Foreign person names and Chinese person names recognition in Chinese texts. Mini-micro System, Vol.27. No.4, Apr.2006,p715-719.
[12] Yu Shiwen, Dua minghui, Zhuo Xuefeng et. Specification for Corpus Processing at Peking University, Journal of Chinese Language and Computing, 13 (2) 121-158.

Dr. Liu Hongjian is a senior researcher in Hitachi (China) Research & Development. He receive his Ph.D. in Pattern Recognition and Intelligent System from Shanghai Jiaotong University, China, 2005. Presently, His research focus on Chinese speech synthesis. Until now, his has published more than 10 papers and applied 10 patents. His research interests include natural language processing, pattern recognition and data mining.

Guo Defeng now work in Hitachi (China) Research & Development as a researcher. He receive his M.S. Degree in Computer Science from Shanghai Jiaotong University, China, 2008. His research interests include natural language processing, pattern recognition and data mining.

Zhou Quan is now working in Hitachi (China) Research & Development as an associate senior researcher. He receive his M.S. Degree in Pattern Recognition and Intelligent System from Shanghai Jiaotong University, China, 2007. His research interests include natural language processing, pattern recognition and data mining.

Nagamatsu Kenji is a senior researcher in Hitachi Ltd. Central Research laboratory. He receive his Ph.D. Degree in Faculty Computer Science from University of Tokyo, Japan, 2008. His research interests include speech processing.

Sun Qinghua now working in Hitachi Ltd. Central Research laboratory. He receive his Ph.D. Degree in Computer Science from University of Tokyo, Japan, 2008. His research interests include speech processing.

© 2010 ACADEMY PUBLISHER