DESCRIPTION OF THE NTU SYSTEM USED FOR MET2

Hsin-Hsi Chen, Yung-Wei Ding, Shih-Chung Tsai and Guo-Wei Bian

Natural Language Processing Laboratory
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN
E-mail: hh_chen@csie.ntu.edu.tw
Fax: +886-2-23628167

ABSTRACT

Named entities form the major components in a document. When we catch the fundamental entities, we can understand a document to some degree. This paper employs different types of information from different levels of text to extract named entities, including character conditions, statistic information, titles, punctuation marks, organization and location keywords, speech-act and locative verbs, cache and *n*-gram model. In the formal run of MET-2, the F-measures P&R, 2P&R and P&2R are 79.61%, 77.88% and 81.42%, respectively.

INTRODUCTION

People, affairs, time, places and things are five basic entities in a document. When we catch the fundamental entities, we can understand a document to some degree. Natural Language Processing Laboratory (NLPL) in Department of Computer Science and Information Engineering (CSIE), National Taiwan University (NTU) starts to study named entity extraction problem in 1993. At first, we focus on the extraction of Chinese person names, transliterated person names [1] and organization names [2]. The training data and the testing data in these experiments are selected from three Taiwan newspaper corpora (China Times, Liberty Times News and United Daily News). Chen and Lee [3] reported the precision rate and the recall rate for the extraction of Chinese person names, transliterated person names and organization names are (88.04%, 92.56%), (50.62%, 71.93%) and (61.79%, 54.50%), respectively in the 16th International Conference on Computational Linguistics. We employ these results to several applications. Chen and Wu [4] considered person names as one of clues in sentence alignment. Chen and Lee [3] show its application to anaphora resolution. Chen and Bian [5] proposed a method to construct white pages for Internet/Intranet users automatically. We extract information from World Wide Web documents, including proper nouns, E-mail addresses and home page URLs, and find the relationship among these data. Chen, Ding and Tsai [6,7] dealt with proper noun extraction for information retrieval.

In MUC-7 and MET-2, we attend named entity extraction tasks for both English and Chinese. We extend our previous work on this problem to cover more named entity types such as locations, date/time expressions and monetary and percentage expressions. Several issues have to be addressed during extension. One of the major differences between Chinese and English language processing is that segmentation is required for Chinese. That is, we have to identify word boundary in Chinese sentences beforehand. That makes Chinese named entity extraction tasks more changeable. Besides, the vocabulary set and the Chinese coding set used in Taiwan and in China are not the same. The documents adopted in MET-2 are selected from newspapers in China, thus we have to transform simplified Chinese characters in GB coding set to traditional Chinese characters in Big-5 coding set before testing. A word that is known may become unknown due to transformation. For example, the character "晨" in "凌晨" (early morning) is used in traditional Chinese characters. However, "辰" is used in simplified Chinese characters and it is also a legal traditional Chinese character that denotes another meaning. In other words, the mapping from GB to Big5 is "凌辰", which is an unknown word based on our dictionary. The different vocabulary set between China and Taiwan results in different segmentation.

This paper is organized as follows. Section 2 illustrates the flow of named entity extraction and the summary scores of our team in MET-2 formal run. Sections 3, 4 and 5 propose methods to extract named people, organizations and locations. Section 6 deals with the rest of named entities, i.e., date/time expressions and monetary and percentage expressions. After each section, we discuss the sources of errors in the formal run. Section 7 concludes the remarks.

FLOW OF NAMED ENTITY EXTRACTION

The following shows the flow of named entity extraction in MET-2 formal run.

(1) Transform Chinese texts in GB codes into texts in Big-5 codes.
(2) Segment Chinese texts into a sequence of tokens.

(3) Identify named people.
(4) Identify named organizations.
(5) Identify named locations.
(6) Use n-gram model to identify named organizations/locations.
(7) Identify the rest of named expressions.
(8) Transform the results in Big-5 codes into the results in GB codes.

Steps (1) and (2) form the preprocessing of the named entity extraction tasks. As mentioned in Section 1, Big-5 traditional character set and GB simplified character set are adopted in Taiwan and in China, respectively. Our system is developed on the basis of Big-5 codes, so that the transformation of the official documents in MET-2 into the documents in terms of Big-5 codes is necessary. Characters used both in simplified character set and tradition character set always result in error mapping. For example, 旅遊 vs. 旅游, 報導 vs. 報道, 最後 vs. 最后, 那麼 vs. 那么, 準確 vs. 准確, 並不是 vs. 并不是, 幾十年 vs. 几十年, 長時間裡 vs. 長時間里, 好像 vs. 好象, 由於 vs. 由于, and so on.

A Chinese sentence is composed of a sequence of characters without any word boundary. Step (2) tries to identify words on the basis of a dictionary and segmentation strategies. We list all the possible words by dictionary look-up, and then resolve ambiguities by segmentation strategies. Our dictionary is trained from CKIP corpus [8], of which articles are collected from Taiwan newspapers, magazines, and so on. The vocabulary used in MET-2 documents may be different from the vocabulary trained from Taiwan corpora, so that more unknown words are introduced. For example, "人工智慧" vs. "人工智能", "軟體" vs. "軟件", "紐西蘭" vs. "新西蘭", "肯亞" vs. "肯尼亞", and so on. That will interfere with the named entity extraction because named entities are often unknown words too.

Table 1 summarizes the results of MET-2 formal run of our team. The F-measures in terms of P&R, 2P&R, and P&2R are 79.61%, 77.88% and 81.42%, respectively. The recall rate and the precision rate (object scores) for the extraction of name, time and number expressions are (85%, 79%), (91%, 98%) and (95%, 85%), respectively. We will discuss the major errors for each type of named entities.

NAMED PEOPLE EXTRACTION

The naming methods are totally different for Chinese person names and transliterated person names. The following two subsections deal with each of them.

Identification of Chinese Person Names

Chinese person names are composed of surnames and names. Most Chinese surnames are single character and some rare ones are two characters. The following shows three different types:

(1) Single character like '趙', 錢', '孫' and '李'.
(2) Two characters like '歐陽' and '上官'.
(3) Two surnames together like '蔣宋'.

Most names are two characters and some rare ones are single characters. Theoretically, every character can be considered as names rather than a fixed set. Thus the length of Chinese person names ranges from 2 to 6 characters.

Three kinds of recognition strategies are adopted:

(1) name-formulation rules
(2) context clues, e.g., titles, positions, speech-act verbs, and so on
(3) cache

Name-formulation rules form the baseline model. It proposes possible candidates. The context clues add extra scores to the candidates. Cache records the occurrences of all the possible candidates in a paragraph. If a candidate appears more than once, it has high tendency to be a person name. The following illustrates each strategy in details.

Name-formulation rules are trained from a person name corpus in Taiwan [9]. It contains 1 million Chinese person names. Each contains surname, name and sex. During training, we divide the corpus into two partitions according to sex of persons. In our method, we postulate that the formulation of names is different for male and female. At first, we get 598 surnames from this 1M person name corpus, and then compute the probabilities of these characters to be surnames. Of these, surnames of very low frequency like "是", "那", *etc*., are removed from this set to avoid too much false alarms. Only 541 surnames are left, and are used to trigger the person name identification system. Next, the probability of a Chinese character to be the first character (the second character) of a name is computed for male and female, separately.

	POS	ACT	COR	PAR	INC	MIS	SPU	NON	REC	PRE	UND	OVG	SUB	ERR
SUB TASK SCORE														
enamex														
organization	377	344	293	0	7	77	44	0	78	85	20	13	2	30
person	174	215	159	0	0	15	56	0	91	74	9	26	0	31
location	750	842	583	0	65	102	194	0	78	69	14	23	10	38
other	0	0	0	0	0	0	0	0	0	0	0	0	0	0
timex														
date	380	410	359	0	0	21	51	0	94	88	6	12	0	17
time	43	60	42	0	0	1	18	0	98	70	2	30	0	31
other	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numex														
money	52	52	51	0	0	1	1	0	98	98	2	2	0	4
percent	47	40	39	0	0	8	1	0	83	98	17	3	0	19
other	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SECT SCORE														
DOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AU	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUTHOR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DATE	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HEADLINE	144	138	105	0	7	32	26	0	73	76	22	19	6	38
HL	50	54	38	0	4	8	12	0	76	70	16	22	10	39
ID	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TEXT	3452	3734	2871	0	171	410	692	0	83	77	12	19	6	31
OBJ SCORE														
enamex	1301	1401	1107	0	0	194	294	0	85	79	15	21	0	31
numex	99	92	90	0	0	9	2	0	91	98	9	2	0	11
timex	423	470	401	0	0	22	69	0	95	85	5	15	0	18
SLOT SCORE														
enamex														
text	1301	1401	1039	0	68	194	294	0	80	74	15	21	6	35
type	1301	1401	1035	0	72	194	294	0	80	74	15	21	7	35
numex														
text	99	92	88	0	2	9	2	0	89	96	9	2	2	13
type	99	92	90	0	0	9	2	0	91	98	9	2	0	11
timex														
text	423	470	361	0	40	22	69	0	85	77	5	15	10	27
type	423	470	401	0	0	22	69	0	95	85	5	15	0	18
ALL SLOTS														
	3646	3926	3014	0	182	450	730	0	83	77	12	19	6	31
F-MEASURES									P&R		2P&R		P&2R	
									79.61		77.88		81.42	

Table 1: Summary Scores of NTUNLPL

The following models are adopted to select the possible candidates. We consider the above three types of surnames.

Model 1. Single character

(i) $P(C_1)*P(C_2)*P(C_3)$ using male training table > $Threshold_1$ and
$P(C_2)*P(C_3)$ using male training table > $Threshold_2$, or

(ii) $P(C_1)*P(C_2)*P(C_3)$ using female training table > $Threshold_3$ and
$P(C_2)*P(C_3)$ using female training table > $Threshold_4$

Model 2. Two characters

(i) $P(C_2)*P(C_3)$ using male training table > $Threshold_2$, or

(ii) $P(C_2)*P(C_3)$ using female training table > $Threshold_4$

Model 3. Two surnames together

$P(C_{12})*P(C_2)*P(C_3)$ using female training table > $Threshold_3$,
$P(C_2)*P(C_3)$ using female training table > $Threshold_4$ and
$P(C_{12})*P(C_2)*P(C_3)$ using female training table >
$P(C_{12})*P(C_2)*P(C_3)$ using male training table

where C_1, C_2 and C_3 are a continuous sequence of characters in a sentence, and they denote surname and names, respectively,
C_{11} and C_{12} denote the first and the second surnames,
$P(C_i)$ is the probability of C_i to be a surname or a name.

For different types of surnames, different models are adopted. Because the surnames with two characters are always surnames, Model 2 neglects the score of surname part. Both Models 1 and 3 consider the score of surname. We compute the probabilities using female and male training tables, respectively. In Models (1) and (2), either male score or female score must be greater than thresholds. In Model (3), the person names must denote a female. In this case, the probability to be female must be greater than the probability to be male. The above three models can be extended to the single-character names. When a candidate cannot pass the thresholds, its last character is cut off and the remaining string is tried again. Thresholds are trained from the 1-million person name corpus. We let 99% of the training data pass the thresholds.

Besides the baseline model, titles, positions and special verbs are important local clues. When a title such as '總統' (President) appears before (after) a string, it is probably a person name. There are 476 titles in our database. Person names usually appear at the head or the tail of a sentence. Persons may be accompanied with speech-act verbs like "發言", "說", "提出", *etc*. For these cases, extra scores are added to help strings pass the thresholds.

Finally, we present a global clue. A person name may appear more than once in a document. We use cache to store the identified candidates and reset cache when next document is considered. There are four cases shown below when cache is used:

(1) $C_1C_2C_3$ and $C_1C_2C_4$ are in the cache, and C_1C_2 is correct.
(2) $C_1C_2C_3$ and $C_1C_2C_4$ are in the cache, and both are correct.
(3) $C_1C_2C_3$ and C_1C_2 are in the cache, and $C_1C_2C_3$ is correct.
(4) $C_1C_2C_3$ and C_1C_2 are in the cache, and C_1C_2 is correct.

Cases (1) and (2) (cases (3) and (4)) are contradictory. In our treatment, a weight is assigned to each entry in the cache. The entry that has clear right boundary has a high weight. Titles, positions, and special verbs are clues for boundary. For those similar pairs that have different weights, the entry having high weight is selected. If both have high weights, both are chosen. When both have low weights, the score of the second character of a name part is critical. It determines if the character is kept or deleted.

Identification of Transliterated Person Names

Transliterated person names denote foreigners. Compared with Chinese person names, the length of transliterated names is not restricted to 2 to 6 characters. The following strategies are adopted to recognize transliterated names:

(1) transliterated name set
The transliterated names trained from MET data are regarded as a built-in name set.

(2) character condition
Two special character sets are retrieved from MET training data, Hornby [10] and Huang [11]. The first character of transliterated names must belong to a 280-character set, and the remaining characters must appear in a 411-character set. The character condition is a loose restriction. The string that satisfies the character condition may denote a location, a building, an address, *etc*. It should be employed with other clues (refer to (3)-(5)).

(3) titles
Titles used in Chinese person names are also applicable to transliterated person names. Thus, $\boxed{\text{聖}}$ $\boxed{\text{約翰}}$ 港 will not be recognized as a transliterated person name.

(4) name introducers
Some words like "叫", "叫作", "叫做", "名叫", and "尊稱" can introduce transliterated names when

they are used at the first time.

(5) special verbs

Persons always appear with some special verbs like "發表", "暗示", and so on. Thus the same set of verbs used in Chinese person names are also used for transliterated person names.

Besides the above strategies, a complete transliterated person name is composed of first name, middle name and last name. For example, 阿卜杜勒・巴塞特・阿里・賽格拉西. The first, middle and last names are connected by a dot.

Cache mechanism is also helpful in the identification of transliterated names. A candidate that satisfies the character condition and one of the clues will be placed in the cache. At the second time, the clues may disappear, but we can recover the transliterated person name by checking cache. The following shows an example:

… 米切爾法官 …，… 米切爾因為 …

Title does not show up, when the name is mentioned again.

Discussion

The summary report in Table 1 shows the recall rate and the precision for person names are 91% and 74%, respectively. The major errors are listed below:

(1) segmentation

In our treatment, segmentation is done before named entity extraction. Part of person names may be regarded as words during segmentation. The following show some examples. The characters "黄金", "高潮", and "盛世" are common content words.

黄金富 -> 黄金 富

義高潮 -> 義 高潮

盛世良 -> 盛世 良

In this case, the person name is missed.

(2) surname set and character set

Those characters not listed in surname set are not considered as surnames, so that they cannot trigger our identification system. The characters "肖" and "庄" in person names "肖成林" and "庄霞琴" are typical examples. Similarly, if the character of a transliterated person name does not belong to the predefined character set, the character will be neglected. For example, "捷" in "卡拉捷耶夫" is not listed in the character set, and the scope error happens.

(3) blanks

Blank may appear between surname and name in the original MET-2 documents, e.g., "羅 俏". After segmentation, blanks are also inserted between words. We cannot tell if the blanks exist in the original documents or are inserted by our segmentation system.

(4) boundary errors

Some Chinese person names are mis-regarded as transliterated names, e.g.,

溫克剛 -> 溫克

賈西平 -> 賈西

(5) titles

Titles are important clues for the identification of transliterated person names. Even if a transliterated name satisfies the character condition, it is not identified without title. The name "卡庫" in the string "醫生卡庫" is missed because "醫生" is not listed in our title set.

(6) Japanese names

The current version cannot deal with Japanese names like "田中真紀子".

NAMED ORGANIZATION EXTRACTION

Extraction Algorithm

The structure of organization names is more complex than that of person names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. The following specifies the rules we adopted to formulate its structure.

OrganizationName → OrganizationName OrganizationNameKeyword

e.g., 聯合國 部隊

OrganizationName → CountryName OrganizationNameKeyword

e.g., 美國 大使館

OrganizationName → PersonName OrganizationNameKeyword

e.g., 羅慧夫 基金會

OrganizationName → CountryName {D|DD} OrganizationNameKeyword

where D is a content word.

e.g., 中國 國際 廣播電台

OrganizationName → PersonName {D|D} OrganizationNameKeyword

e.g., 羅慧夫 文教 基金會

OrganizationName → LocationName {D|D} OrganizationNameKeyword

e.g., 台北 國際 廣播電台

OrganizationName → CountryName OrganizationName

e.g., 美國 國防部

OrganizationName → LocationName OrgnizationName

e.g., 伊利諾州 州府

In current version, we collect 776 organization names and 1059 organization name keywords.

Transliterated person names and location names in the above rules still have to satisfy the character condition mentioned in last section. However, the character set is trained from transliterated person name corpus. It may not be suitable for location names. Consider an example "帕鬆錯湖旅遊度假村". "帕鬆錯湖", which is a lake in China, is not a transliterated name. The characters "鬆" and "錯" do not belong to the character set. Here, we utilize the feature of multiple occurrences of organization names in a document and propose n-gram model to deal with this problem. Although cache mechanism and n-gram use the same feature, i.e., multiple occurrences, their concepts are totally different. For organization names, we are not sure when a pattern should be put into cache because its left boundary is hard to decide. In our n-gram model, we select those patterns that meet the following criteria:

(1) It must consist of a name and an organization name keyword.
(2) Its length must be greater than 2 words.
(3) It does not cross sentence boundary and any punctuation marks.
(4) It must occur at lease two times.

Discussion

Table 1 shows the recall rate and the precision rate for the extraction of organization names are 78% and 85%, respectively. The following shows the error analysis.

(1) more than two content words between name and keyword
In current version, we accept only two interference words. Thus, the string "中國 衛星 發射 代理 公司" is not recognized.

(2) absent of keywords
Keywords are important indicators for right boundary. The string "巴解法塔賀武裝" is lack of keyword, so it is missed.

(3) absent of name part
Name part serves as an indicator of left boundary. In the string "亞星公司", we cannot find a name.

(4) n-gram errors
N-gram employs multiple occurrences to find a pattern. It is easy to propose false alarms, e.g., "這家銀行" and "安得拉邦東南部發射基地".

NAMED LOCATION EXTRACTION

Extraction Algorithm

The structure of location names is similar to that of organization names. A complete location name is composed of name part and keyword part. We use the following rule to formulate this structure.

LocationName → PersonName LocationNameKeyword

LocationName → LocationName LocationNameKeyword

Currently, we have 45 location keywords. The following shows some examples:

'山', '中心', '公路', '以北', '以西', '以東', '以南', '半島', '半球', '市', '市中心', *etc.*

There are 16,442 built-in location names in current versions. For the treatment of location names without keywords, we also introduce some locative verbs like '來自', '前往', and so on. The objects following this kind of verbs may be location names. For example, in the string "飛往聖路易斯", "聖路易斯" will be identified. Cache is also useful. For example, assume '巴塞隆納市' is recognized as a location name and placed in cache. When '巴塞隆納' appears, it will be identified as a location name even if the location name keyword is omitted. N-gram model is also employed to recover those names that do not meet the character condition.

Discussion

Table 1 shows the recall rate and the precision rate in this part are 78% and 69%, respectively. The performance is worse than that of named people and named organization. The major types of errors are shown below.

(1) character set
The characters "鹿" and "島" in the string "鹿兒島縣" do not belong to our transliterated character set. Actually, it denotes a Japanese location name.

(2) wrong keyword
The character "部" is an organization keyword. Thus the string "菲律賓馬部" is misregarded as an organization name.

(3) common content words
The words such as "太陽", "土星", *etc*., are common content words. We do not give them special tags.

(4) single-character locations
The single-character locations such as "中", "日", and so on, are missed during recognition. Total 57 errors are of this type.

(5) interference words between name part and keywords
There are words between name part and keywords. For example, "肯尼迪航天中心" and "懷特桑茲導彈發射場". Here the words "航天" and "導彈" are common words in China newspaper, but seldom used in Taiwan.

OTHER ENTITY EXTRACTION

Extraction Algorithm

We use grammar rules to capture the remaining entities, including date/time expressions and monetary and percentage expressions. The following shows the specification of each type of expressions. Each rule is accompanied with an example.

(1) date expressions

DATE → NUMBER YEAR	(三 年)
DATE → NUMBER MTHUNIT	(十 月)
DATE → NUMBER DUNIT	(五 日)
DATE → REGINC	(元旦)
DATE → FSTATE DATE	(今年 三月)
DATE → COMMON DATE	(前 兩年)
DATE → REGINE DATE	(民國 七十八年)
DATE → DATE DMONTH	(今年 三月)
DATE → DATE BSTATE	(去年 初)
DATE → FSTATEDATE DATE	(這年 三月底)
DATE → FSTATEDATE DMONTH	(今年 元月)
DATE → FSTATEDATE FSTATEDATE	(明年 今天)
DATE → DATE YXY DATE	(去年三月 至 今年五月)

(2) time expressions

TIME → NUMBER HUNIT	(五 時)
TIME → NUMBER MUNIT	(三十 分)
TIME → NUMBER SUNIT	(六 秒)
TIME → FSTAETIME TIME	
TIME → FSTATE TIME	
TIME → TIME BSTATE	
TIME → MORN BSTATE	
TIME → TIME TIME	
TIME → TIME YXY TIME	(今天 到 明天)
TIME → NUMBER COLON NUMBER	(03 ： 45)

(3) monetary expressions

DMONEY → MOUNIT NUMBER MOUNIT	(美金 五 元)
DMONEY → NUMBER MOUNIT MOUNIT	(五 元 美金)
DMONEY → NUMBER MOUNIT	(五 元)

DMONEY → MOUNIT MOUNIT NUMBER (美金 $ 5)
DMONEY → MOUNIT NUMBER ($ 5)
DMONEY → NUMBER YXY DMONEY (三 至 五元)
DMONEY → DMONEY YXY DMONEY (三元 至 五元)
DMONEY → DMONEY YXY NUMBER ($200 - 500)

(4) percentage expressions
DPERCENT → PERCENT NUMBER (百分之 十)
DPERCENT → NUMBER PERCENT (3 %)
DPERCENT → DPERCENT YXY DPERCENT (5% 到 8%)
DPERCENT → DPERCENT YXY NUMBE (百分之八 到 十)
DPERCENT → NUMBER YXY DPERCENT (八 到 十百分點)

Rule-based approach is simple. We can add, delete and modify rules quickly without modifying the identification programs. However, the above rules cannot capture ambiguous cases. For example, "號" may mean the address number (e.g., 中山路九號) or the date number (e.g., 三月九號). Augmented grammar rules are needed to introduce constraints to check if the extracted entity can fit into the context.

Discussion

The summary report in Table 1 shows that the recall rate and the precision rate for date expression, time expression, monetary expression and percentage expression are (94%, 88%), (98%, 70%), (98%, 98%) and (83%, 98%), respectively. The major errors are shown as follows:

(1) propagation errors
Because we employ a segmentation system to identify basic tokens before entity extraction, some words like "迄今", "今後", *etc*., are regarded as terms. In this way, "今" is always missed. Similarly, named people are extracted before date expressions. The errors resulting from the previous steps propagate to the next steps. Consider the following example.
自 1998 年阿麗雅納 …
The named people extraction procedure regards "年阿麗雅納" as a transliterated name. After that, "1998 年" is missed because the date unit is absent.

(2) absent date units
In some sentences, the date unit "年" does not appear, so that "一九九六" is missed. In some examples like "九月十", the date unit should appear but it is absent. Thus it is also not captured.

(3) absent keywords
Some keywords are not listed. For example, "莫斯科時間", "至今", and so on. Thus, for "上午莫斯科時間 8 點 58 分" and "1960 年至今", only some fragments, e.g., "上午", "8 點 58 分", and "1960 年" are identified.

(4) rule coverage
Patterns like "今、明兩年" are not considered in this version, thus they are missed. Similarly, the percentage expressions like "2/3", "二十分之一", "百萬分之一", and so on, are not represented in our grammar.

(5) ambiguity
Some characters like "點" can be used in time and monetary expressions. Expression "十二點七七億美元" is divided into two parts: "十二點" and "七七億美元". Similarly, the strings "十分" and "一時" are words. In our pipelined model, "九點十分" and "下午一時" will be missed.

CONCLUDING REMARKS

This paper proposes a pipeline model to extract named entities from Chinese documents. Different types of information from different levels of text are employed, including character conditions, statistic information, titles, punctuation marks, organization and location keywords, speech-act and locative verbs, cache and *n*-gram model. The context ranges from very short to very long. The recall rate (83%) and the precision rate (77%) are achieved. The major errors result from propagation errors, keyword sets, character sets, rule coverage, and so on. How to integrate different modules (including segmentation and recognition) in an interleaving way, and how to learn grammar rules, keyword sets and character sets automatically have to be studied furthermore.

REFERENCES

[1] Jen-Chang Lee, Yue-Shi Lee and Hsin-Hsi Chen, "Identification of Personal Names in Chinese Texts," *Proceedings of ROCLING VII*, Taiwan, August 12-13 1994, pp. 203-222 (in Chinese).

[2] Hsin-Hsi Chen Jen-Chang and Lee, "The Identification of Organization Names in Chinese Texts,"

Communication of Chinese and Oriental Languages Information Processing Society, **4**(2), Singapore, 1994, pp. 131-142 (in Chinese).
[3] Hsin-Hsi Chen and Jen-Chang Lee, "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9 1996, pp. 222-229.
[4] Hsin-Hsi Chen and Yeong-Yui Wu, "Aligning Parallel Chinese-English Texts Using Multiple Clues," *Proceedings of 2nd Pacific Association for Computational Linguistic Conference*, Queensland, Australia, April 19-22 1995, pp. 39-48.
[5] Hsin-Hsi Chen and Guo-Wei Bian, "Proper Name Extraction from Web Pages for Finding People in Internet," *Proceedings of ROCLING X*, Taipei, Taiwan, August 22-24, 1997, pp. 143-158.
[6] Hsin-Hsi Chen, Yung-Wei Ding and Shih-Chung Tsai, "Proper Noun Extraction for Information Retrieval," *International Journal on Computer Processing of Oriental Languages*, Special Issue on Information Retrieval on Oriental Languages, 1998.
[7] Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding and Shih-Chung Tsai, "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of COLING-ACL98*, Montreal, August 10-14, 1998.
[8] Chu-Ren Huang, *et al*., "Introduction to CKIP Balanced Corpus," *Proceedings of ROCLING VIII*, Taiwan, August 18-19 1995, pp. 81-99 (in Chinese).
[9] ROCLING, *ROCLING Text Corpus Exchange*, Association of Computational Linguistics and Chinese Language Processing, 1993.
[10] A.S. Hornby, *Oxford Advanced Learner's Dictionary of Current English*, Revised Third Edition, 1984.
[11] Y.J. Huang, *English Names for You*, Learning Publishing Company, Taiwan, 1992.