自 然 言 語 処 理 152－3
（２００２．１１．１２）

Recognizing Named Entities in Japanese Corpus by Incremental Deep Parsing (1)

Li Liyang† Masakazu Tateno ‡

†Northern Jiao Tong University, Beijing, P. R. China and Fuji Xerox Co., Ltd.
Email: leeliyang@yahoo.com.cn

‡Fuji Xerox Co., Ltd.
Email: Tateno.Masakazu@fujixerox.co.jp

Abstract
Extracting person names in Japanese corpus is a crucial component of Named Entity Extraction. This paper proposes a method to recognize person names in Japanese corpus. We collected necessary syntactic information and wrote rules for the parser to recognize person names. The basic method is if the Kanji character can be used in person names or not to help to find a Japanese person name. Besides Japanese names, this method also can be used for European and American names written in Katakana and other Asian names written in Chinese characters. The F score got from an open set is 89.95%.

Keywords
named entity recognition, word segmentation, incremental deep parsing

増深解析法による 日本語の固有名の認識(1)

李黎陽† 舘野昌一‡

†中国北方交通大学、富士ゼロックス株式会社
Email: leeliyang@yahoo.com.cn
‡富士ゼロックス株式会社
Email: Tateno.Masakazu@fujixerox.co.jp

あらまし
本稿は日本語の文書から人名を認識する方法を提案する。人名の抽出は固有名抽出の重要な構成要素である。本方法は日本人の人名用字の統計情報を導入して、日本人名抽出を行う。さらに、本方法によって、日本語文書中のアメリカ、イギリスなどの片仮名で表される人名と中国、韓国などの漢字で表される人名を抽出することができる。評価データにより得た F スコアは 89.95%であった。

1. Introduction

Named Entity (NE) Extraction is one of the subtasks of the information extraction defined by Message Understanding Conference (MUC). Proper nouns have been defined as the seven classes of Organization, Person, Location, Date, Time, Money and Percent in the category of NE. According to the recent research [1], 20.6% of all named entities in Japanese documents are the person names that are just less than the location names. It means to extract person names from Japanese documents is a very important task in the field of Japanese named entity extraction.

We found that there are mainly three types of person names often appear in modern Japanese documents as follows:

1. Japanese person names:

山前譲 さん
稲田重年 さん
額賀福志郎 幹事長
佐々木順一朗 監督

2. Other Asian names except for Japanese written in Chinese characters.

朱邦造 報道局長
唐 外相の訪米

3. Person names written in Katakana, such as American or European names.

イタリアの **ベルルスコーニ** 首相
イリッチ・ラミレス・サンチェス 服役囚
ブラジル代表ＦＷ **ロナウド**
男優賞 **ルイジ・ロカシオ**

(The bold characters represent person names in the examples above.)

We can find that the patterns of names in the Japanese text such as the examples above are so variable and diversiform and it shows the difficulty of this work. There are a lot of family names (including very rare names such as only one or a few families have such names) and a lot of given names (including very new names) in Japanese. So it is impossible to list all names. At the same time, names of foreigners such as types 2 or 3, which are even more various, also should be taken into account.

There are only about 30 thousands of family names and given names (including Japanese and foreigner's names) registered in the vocabulary of Chasen2.0. Obviously it is not enough, so we can often see such results:

柳沢伯夫金融担当相

⇩

柳沢　+ 詞-固有名詞-地域
伯　+名詞-一般
夫　+名詞-一般
金融　+名詞-一般
担当　+名詞-サ変接続
相　+名詞-接尾-一般

We extracted the person names from the part of the corpus of Mainichi Newspaper '95 by Chasen2.0. The result is as follows.

Truth	Result	Correct	Precision	Recall
2,062	1,834	1,429	77.92%	69.30%

This result shows that we can't get a very good result only by morphological analysis and a compensatory work is necessary.

There are statistic methods such as Maximum Entropy, Decision Tree Learning [1] and Support Vector Machines [2] to extract person names from a corpus. Some methods based on handcrafted rules also have appeared besides these. In this paper, we introduce a new method based on the Incremental Deep Parser using syntactic rules to extract person names from the result of Japanese morphological analysis. We use different rules to extract person names for those three different types.

2. Tools

a. Morphological analyzer

We use Chasen2.0 to segment Japanese text into words.

b. Recognizing NE

We use Xerox Incremental Parser (XIP, [3]) to recognize NE by applying syntactic rules and lexicon registered in its vocabulary. The emphasis of our work is to write proper rules for XIP.

c. Post process for unsure results

We check the unsure result after the recognition of NE with Perl scripts.

3. Boundary information of name

The family names of Japanese are often made up of one to three Chinese characters (Kanji) except for some very special cases that have 4 or even more characters. On the other hand, Japanese given names also often include one to three Kanji, except that some use Hiragana instead of Kanji in their given name.

	Family name	Given name
1 Kanji	森	博
2 Kanji	伊藤	一郎
3 Kanji	佐々木	美智子
4 kanji	勅使河原	-
Hiragana	-	みどり

The length of a Japanese person name is from 2 to 7 or even more, so one of the natural ways to find person names is to find out the left boundaries and the right boundaries of family and given names. We introduce what can be boundaries of a person name in Japanese text as follows:

A. Nouns that often appear around a person name to stand for the status of the person such as 社長 or 首相 are efficient hints of person names. So we attach feature "title" for such nouns.

e.g. *依田紀基*名人

In this example, 名人 that has a feature "title" can be a indicator that 依田紀基 is a person name.

B. Some titles such as 主席 or 書記 can follow some other words to combine a longer title,

e.g. 国家主席

Some single Kanji often can be a prefix (such as 元, 前) or a suffix (such as 官 or 相) of a name.

e.g. **元**労働省職業安定局長

国土 交通 **相**

We attach a feature "title-prefix" or "title-suffix" for the word, respectively. The following table shows examples of words that have such features.

title-prefix	title-suffix	title
前	家	社長
副	長	会長
在	相	助教授
現	者	所長
旧	員	特別顧問
助	官	顧問
元	手	首相
…	…	…

e.g: **HEAD ORGN NOUN NOUN NOUN**

元 労働省 職業 安定 局長

The example above is the result of a long title after word segmentation, from which we can find that noun, organization, nation , or place can be components of a long title, so we attach the feature "title-body" for these kinds of words.

We have rules to extract the long title that are also effective to define boundaries of a name with all the features introduced above. Next example shows such a rule.

e.g. **元法政大教授**駒尺喜美さん

The result of word segmentation is:

元 +0+2+接頭詞-名詞接続
+title-prefix
法政大 +2+8+名詞-固有名詞-組織
+organization
教授 +8+12+名詞-一般
+title
駒 +14+16+名詞-一般
+name-body
尺 +16+18+名詞-固有名詞-人名-姓
+name-body
喜美 +18+22+名詞-固有名詞-人名-名
+given-name
さん +22+26+名詞-接尾-一般+
+nameid

The rule to construct the title is as follows:

longtitle = head[title-prefix:+] , ?*[titlebody] , noun[title]; noun[title-suffix].

If one title-prefix is followed by zero or more title-body and a title or a title-suffix, the sequence is a long title. Using this rule, 元法政大教授 can be extracted and it also means that the left boundary of the name 駒尺喜美 is decided.

Some special characters besides titles usually can be the right boundary of person names in Japanese. The following table shows the list of such words.

Nouns	Auxiliary	Punctuation
氏	は	、
さん	が	。
様	も	・
君	と	（
ちゃん		―
クン		
殿		
…	…	…

e.g. 室町鐘緒**氏**

横綱の 武蔵丸 **は**

竹中 平蔵・経済財政担当相

The following elements can be a left boundary of a person name

a) Auxiliary words, such as の or と
b) Punctuation.
   e.g: 母 、 嶋山ハル子さん
c) The beginning of a sentence (Represented by "BOS")
   e.g: **BOS** 野武義男さんと加代子さんの中国籍を持つ夫婦２人
   Symbol "BOS" will be attached in front of every sentence during pre-process.
d) Some common words composed by two or more Kanji.
   e.g: 酒店経営野勢公夫 さん
   here noun 経営 is the left boundary of a person name 野勢公夫

We attached "name-lb" to all words that can be a left boundary while "name-rb" for a right boundary

4. How to recognize person names

4.1 Recognizing Japanese names

As the word segmentation software only has a part of family names and given names in its vocabularies as described earlier, many family names and given names will be divide into pieces if they are not collected in the vocabularies. The followings are some cases:

e.g:

FNAME	GNAME	NOUN	NOUN
依田	紀	基	名人

FNAME	NOUN	GNAME
嶋	山	ハル子

In the first case, given name 紀基 was segmented in error while in second case, family name 嶋山 was also segmented in error.

So we must judge if a single Kanji is a part of a person name or not. In the rules, we use "name-body" to stand for a single Kanji.

It is difficult to cover all the possible family and given names in Japanese, but the number of single Kanji that can be an element of a family or a given name is not so large. From 24037 Japanese family names, we extracted 1,990 Kanji characters used for the first character of family names and 1,600 for the last character. However, from 59,183 Japanese given names, we extracted 5,550 characters for the first character of given names and 4,030 for the last character. We assigned features "flb", "frb", "glb" and "grb" respectively for the segmented word as a single Kanji character after word segmentation. They mean that one Kanji can be the first or last character of family or given name. Of course it is possible that one kanji has more than one of these features.

If a word has one of these features, it can be a component of a name. Because 紀 has feature "glb" while 基 has "grb" in the first case with the assistance of title 名人, we can decide 紀基 is a given name. However, in another example,

FNAME	GNAME	NFCOM	PUNC
金子	金五郎	戦	。

戦 has no feature "grb", we can decide it should be excluded. The abstract of the rule for this is as follows:

name = | left-boundary| name-body [flb:+], name-body+, given-name | right-boundary |.

• "| left-boundary |" and "| right-boundary |" mean the left and right context , they should not be included in the recognized word.

• "name-body+" means one or more single Kanji characters

• If there is a single Kanji with feature "flb", and after it there is one or several single Kanji followed by a given name between left and right boundary, this sequence can be recognized as a name.

Another important point we should notice is that because many Japanese family name are same as Japanese place names or nouns, many family names are often segmented as place names or nouns by word segmentation such as:

宮園佳征被告

宮園　+0+4+名詞-固有名詞-地域-一般
佳　+4+6+名詞-固有名詞-人名-名
征　+6+8+名詞-サ変接続
被告　+8+12+名詞-一般+L2+FRB

A family name 宮園 was tagged as a place name. So we must have the rules to correct those kinds of errors as follows:

name = | left-boundary| family-name (place-name) , name-body+, name-body[grb:+] | right boundary |.

• "family-name (place-name)" means either a family name or a place name is possible

We got the following results by the rules above:

SYMBOL	NOUN			NOUN	END
BOS	PLACE	GNAME	NOUN	NOUN	EOS
	宮園	佳	征	被告	

4.2 Extracting a foreigner's name written in Katakana

In Japanese, a foreigner's name is written in Katakana, which is one of three kinds of characters in Japanese such as:

e.g. 経営者**バルビル・シン・ソディ**さん

It is more difficult to collect the family and the given names compared with Japanese names because the length and format are more various like the example above. So we suppose that the result of word segmentation is even worse.

経営　　+名詞-サ変接続
者　　+名詞-接尾-一般
バル　　+名詞-一般
ビル　　+名詞-一般
・　　+記号-一般
シン　　+名詞-固有名詞-人名-姓
・　　+記号-一般
ソディ　+未知語
さん　　+名詞-接尾-一般

The first name was segmented into two parts and middle part is tagged as surname while the true surname was tagged as "unknown".

According to our method, the first step to find this kinds of names is to find the first name and the middle part if it exists, then go on to find the last name.

If a first or a last name of a foreigner's name can't be recognized, there will be two results:

The first one is tagged with "未知語", which is represented as "unknown" in the rules.

The second one is to be segmented into pieces. We found these pieces were smaller than 4 Katakana characters in many cases just like the example above. The first name is divided into バル and ビル. So we attach a feature "foreign-body" to every Katakana word shorter than 4.

We can use the following rule to extract first name and the middle part for the example above:

Foreign-left= |left-boundary| foreign-body+ | fc |.

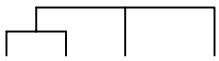
- "foreign-left" refers to the left and the middle part of the Katakana name
- "fc" stands for the connector of name "・"
- "fc" also can be the left boundary of the name, so this rule can be used to extract the middle part of the Katakana name.

The following rule can be applied to recognize the last name of the example above:

Foreign-right = |foreign-left, fc| unknown |right-boundary|.

- "unknown" stand for "未知語"
- The right boundary can not be "fc"

There should be "foreign-left" and "fc" on the left. It means that definition of the first name is crucial for finding last name.



経営者**バル ビル・シン・ソディ**さん

4.3 Recognizing a foreigner's name written in Kanji

Foreigner's names such as the Chinese, the Korean and the Vietnamese are represented by Kanji in Japanese text. Because the format of the Korean and the Vietnamese are similar with the Chinese, we only use Chinese to explain.

Three-character names and two-character names for the Chinese are popular. Both of them only have one Kanji surname in the beginning such as:

李姬鎬 夫人

李 is a popular family name in the eastern Asia, and it is the surname of 李姬鎬**.**

We have the vocabulary to include 365 Chinese family names that cover 99% of family names used in Chinese, and they also cover all the family names from other Asian countries. We use a feature "ch-fname" to tag these 365 Chinese family names.

Because it is almost impossible for the given name in a three-character Asian name to be a meaningful word in Japanese, it would be usually segmented into two single Kanji characters during word segmentation.

李　　+名詞-固有名詞-人名-姓+L1
姬　　+名詞-一般+L1
鎬　　+名詞-一般+L1
夫人　　+名詞-一般+L2

Because the single Kanji in the result of word segmentation is very similar to the result of Japanese names, the feature "name-body" introduced in 4.1 also can be used to represent them.

So we can introduce one of rules for Asian names:

cname= | left-boundary | ch-fname, name-body, name-body | right-boundary |.

We can extract the name 李姬鎬 in the above example by this rule. (夫人 is a title for Mrs.)

5. Experiment and result

During our study, we used the corpus of Mainichi Newspaper to train and test our system. After correcting the errors of tagging by hand in four documents 950101, 950115, 950110 and 950114, we chose the first two for training corpus and used the rest for testing. The results of them is listed in the following table:

		Result	Correct	Truth	P(%)	R(%)
Train	J	542	504	528	93.0	95.5
	K	253	216	261	85.3	82.8
	A	45	35	44	77.8	79.5
Test	J	852	762	809	89.3	94.1
	K	369	313	398	84.8	78.6
	A	28	17	22	60.7	77.3

- J : Japanese names
- K: Names in Katakana
- A: Asian names in Kanji

So the ground total of Mainichi Newspaper 1995 is as follows:

	Result	Correct	Truth	P(%)	R(%)	F(%)
Train	844	755	833	89.45	90.64	90.04
Test	1,249	1,092	1,229	87.43	88.85	88.2

We tested our rules with the corpus of IREX whose content is Mainichi Newspaper 1999. The person names appeared in it would be much different from that in Mainichi Newspaper 1995, so it is very suitable for the test. We selected 69 articles for the test, and the result is as follows:

	Result	Correct	Truth	P(%)	R(%)	F(%)
Test	373	340	383	91.2	88.8	89.95

6. Comparison

The following table shows some of the best results done by Isozaki [2] and Utsuro, Sassano and Uchimoto [1] with our result. Our result seems to be comparable with their statistical methods.

Methods	Done by	F-score (%)
Linguistic method with Handcrafted rules	Li, Tateno	89.95
Support Vector Machine	Isozaki	89.65
Decision List Learning and Maximum Entropy	Utsuro, Sassano, Uchimoto	86.15

7. Conclusion and Future work

We illustrated our work for person names that uses the handcrafted rules on the Incremental Deep Parser. The merits to use the linguistic methods such as handcrafted rules are

(1) The rules capture the linguistic phenomena that are generic in the language.
(2) The rules are understandable, reasonable and editable for human
(3) The errors are obvious and good hints to improve the rules

So we will identify errors that are not fixed yet and improve the rules to reach better precisions and recalls. It would be the good way to extend this method for other named entities such as organization names.

References

[1] 宇津呂 武仁, 颯々野 学, 内元 清貴, "正誤判別規則学習を用いた複数の日本語固有表現抽出システムの出力の混合", 自然言語処理, 第 9 巻, 第 1 号, pp.65-100, January 2002

[2] Hideki Isozaki , Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning, Proceedings of ACL-2001, pp.306--313, 2001

[3] Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. "Robustness beyond shallowness: incremental deep parsing". In Natural Language Engineering, 8(2):121--144, 2002.