E-20 Research on Rule-based Name Extraction for Organizations in Simplified Chinese Texts

Xinkai Wang[†] , Mazakazu Tateno[‡]

Abstract

Named Entity Extraction (NEE) is an important branch of natural language processing. The correctness of extracted proper names affects the performance of further applications, such as Information Extraction. In this paper, we try to find out the effective ways to extract organizations' names from untagged Simplified Chinese texts using a syntactic parser. The recall rate and precision rate follow respectively.

1. Background

Information Retrieval (IR) now is one of the hot spots in Natural Language Processing (NLP). In most cases, proper names include significant information; and extracting proper names correctly has become the key of whether IR performs well or not. We are concentrated on identification of organizations' names in Chinese. There are two kinds of difficulties that cause it hard to extract proper names for organizations exactly in Chinese.

One is that many words in Chinese play different roles in different sentences, while there are no morphological changes for these words. In the following sentence, for example, "华润创业基金管理公司 12 日宣布… ", which meaning is in 12th Hua Run Career Foundation Management Company announced that… , "华润创业基金" and "管理公司" are thought of as organizations' names by system. In fact, the name is "华润创业基金管理公司". The reason why system makes a mistake is that "管理", generally, is a transitive verb, but here it is a noun. There is no morphological change for "管理" in these two cases.

Another is that POS tagger assigns a wrong POS to some words. In this sentence, for example, "中国积极参与亚太经合组织的活动", which means that China takes an active part in the activities of APEC, "亚太经合组织" is the proper name. However, system finds out "合组织" as the name, because POS for "经" is a preposition, as we know, preposition can never appear in proper names.

These difficulties cause the results of NEE in Chinese to be unable to meet with the requirement of further researches.

2. XIP Rules for Organizations' Names

The organizations' names that we define are:

- Academic organizations' names,
- Commercial organizations' names,
- Governmental organizations' names,
- Military organizations' names, and
- In the case of adjacent organizations' names, all of them are seen as only one name.

XIP (Xerox Incremental Parser) [1] is a powerful and robust system for parsing natural language. Our research is based on it. Figure 1 shows the procedure of our system.

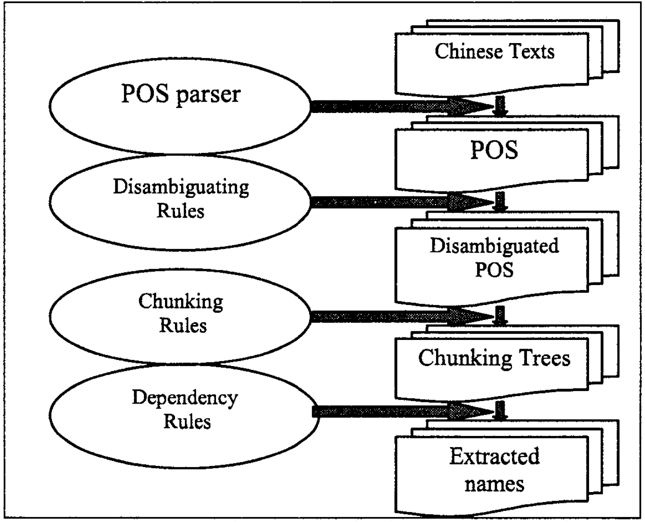


Figure 1 Procedure of organizations' name extraction

In Figure 1, three types of rules, disambiguation rules, chunking rules and dependency rules, are applied in our system. Disambiguation rules are used to prune and correct wrong segmentations and tags produced by any POS tagger. Chunking rules are employed to group words into a chunk tree, which is or includes the proper names for organizations. The function of dependency rules is the extraction of useful information or special structure.

We have found that 83% of names of organizations have suffixes, which is our target now. We use three layers of chunking rules to group names, in which we define some POSs, which are not listed in corpus, and the most important of them is *s_general*, meaning general words, not appearing as names. We also find that they comprise three parts: *head*, *bodies* and *end*.

The Head is the beginning POS of names. And the *head* determines the left boundary of an organization's name. *head* can be one part of an organization's name (We define this type of *head* as **H-1**.) or be a neighbor of names, not including in names (We define this as **H-2**.).

List 1 shows two types of head.

Kind of head	Part of speech
A part of names (H-1)	Proper nouns, English, suffixes for organizations' names, organizations' names
Never a part of names (H-2)	Aspect, comma, conjunction, pronoun, preposition, clitic, verb

List 1 Kinds of head

[†] Fuji Xerox Co. Ltd., Northern Jiao Tong University, P. R. China
[‡] Fuji Xerox Co. Ltd.

In the above two sentences, "华" and "亚太" are the examples of *H-1*. And "基金" is an example of *H-2*.

Bodies are POSs between *head* and *end*, always including in organizations' names, but some categories, namely some POSs, should never become *bodies*. *Bodies* have a hierarchical structure, so *bodies* of rules of different layers may include different categories (We define them as **B-1**, **B-2** or **B-3**.). List 2 shows them.

Layer of rules	POS not included in bodies
Layer 1 (B-1)	Unknown kanji, comma, auxiliary verb, pronoun, clitic, aspect, suffixes for organizations' names, s_general
Layer 2 (B-2)	Unknown kanji, comma, auxiliary verb, pronoun, clitic, aspect, suffixes for organizations' names, s_general, orgnizations' name, title, quantifier, conjunction, preposition.
Layer 3 (B-3)	Same as Layer 2

List 2 Hierarchy of bodies

In the above two sentences, "经", "合" and "创业" are examples of *B-1*. And "管理" is an example of *B-2* because XIP identifies this wrong name, "管理公司", with rules of second layer.

The End is the last POS of an organizations' name; and it determines the right boundary of organizations' names, and it is always one part of names. We list a few organizations' names and most suffixes for these names in the *end* set. In the above sentences, "组织", "基金" and "公司" are all the elements of *end*.

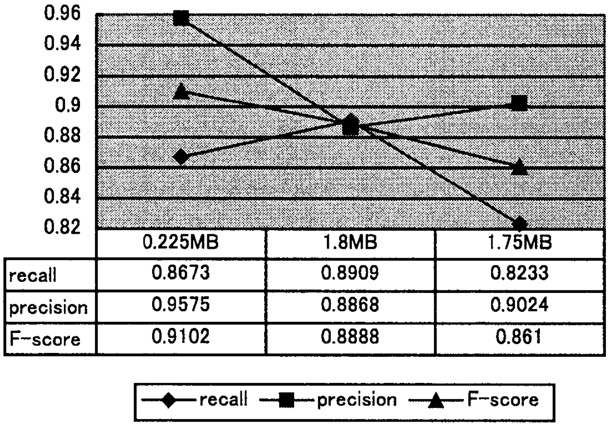
Now, our rules are the meaningful combinations of head, bodies or end. Here are the abstract rules in our system.
1> NAME = H-1 or H-2, +B-1, E.
2> NAME = H-2 , +B-2, E.
3> NAME = H-1 or H-2, +B-3, E.
"*NAME*" is the POS for organizations', which is defined by us. "+", appearing before B-1, B-2 or B-3, means that one or more bodies. The concrete elements for *head*, *bodies* or *end* are changing according to different POS taggers.
For example, the rule of Layer 2 means names of organizations should consist of two parts. One is bodies, which should not have categories listed in B-2; another is ends. The left boundaries of names are determined by the categories listed in H-2, which is not any part of such names of organizations.

3. Results and Conclusions

Our research has used a part of People's Daily tagged corpus (PFP Corpus) **[2]**, the materials of January of 1998, which is made by Beijing University and Fujitsu R&D Center Co. Ltd. The size of training set is 1.80MB, including 5529 organizations' names, while the same things for testing set are respectively 1.75MB and 5716 names.

We selected the first fifty files, about 0.225MB, in the training set and carefully adjusted XIP rules in terms of them. Then we briefly checked the rest part of the training set. Figure 2 shows performance of our system.

Figure 2 Performance of system



	0.225MB	1.8MB	1.75MB
recall	0.8673	0.8909	0.8233
precision	0.9575	0.8868	0.9024
F-score	0.9102	0.8888	0.861

The first set, which we have checked out carefully, performs worse than the second set, which we just browsed. It is caused by the uneven distributions of organizations' names in the training set, which means that some frequent names appear in the hind part of the training set.

4. Future Research

In the future, we are planning to concentrate ourselves on improving the performance of our system, with Machine Learning methods **[3, 4]**. We found that grammatical rules have limits for NEE, and the results are sensitive to the correction of POSs. Although XIP has the mechanism of disambiguation of POS, enumeration of all the incorrect POSs in XIP is not a good way to work around it. Machine Learning methods have good performance **[3]** and have been successfully applied in English **[3]** and in Japanese **[4]**. We now focus on how to convert the problem of NEE in Chinese to a problem of classification.

Reference

1.Salah Ait-Mokhtar, Jean-Pierre Chanod, Claude Roux *Robustness beyond shallowness: incremental deep parsing* In: *Journal of Natural Language Engineering*, Special Issue on Robust Methods in Analysis of Natural Language Data, Afzal Ballim, Vincenzo Pallotta (eds) Cambridge University Press (to appear).
2.Beijing University and Fujitsu R&D Center Co. Ltd., *People's Daily Tagged Corpus (PFP Corpus)*.
3.Tetsuji Nakagawa, Taku Kudoh and Yuji matsumoto, *Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machine*, Proceeding of the Sixth Natural Language Processing Pacific Rim Symposium, Nov. 2001.
4.Satoshi Sekine, Ralph Grishman, Hiroyuki Shinnou, *A Decision Tree Method for Finding and Classifying Names in Japanese Texts,* Sixth Workshop on Very Large Corpora 1998.