Compound Noun Based System for Automatic Term Recognition Task

Hiroshi Nakagawa

Information Technology Center, The University of Tokyo
nakagawa@naklab.dnj.ynu.ac.jp

Abstract

This paper describes the overview of our system and evaluation of the term recognition task and the role analysis task. We used two methods, Compound Noun based ranking method and Nested Collocation based ranking method, in term recognition. In the role analysis task, we use a pattern driven information extraction method.

1 Introduction

Many works have been done on automatic term recognition, and shown amount of results being improved year by year. In this situation, one promising way to develop a new and better method is to combine several methods to utilize all characteristic and/or strong points of them. However, there are countless ways of possible combinations. Thus, for effective combinations, we need a certain kind of categorization among those proposed methods. In this sense, the dichotomy which (Kageura and Umino1996, Kageura and Umino1996) makes is crucial. They divided the all proposed methods into two categories, namely unithood based method and termhood based method. These two notions are stated as follows.

Unithood refers to the degree of strength or stability of syntagmatic combinations or collocations. For instance, a word has very solid unithood. Other linguistic units having high unithood are compound word, collocation, and so forth.

Termhood refers to the degree that a linguistic unit is related to domain-specific concepts. Termhood is usually calculated based on term frequency and bias of frequency(so called Inverse Document Frequency). Even though these calculations give a good approximation of termhood, they do not directly reflect termhood because these calculations are based on superficial statistics.

We combine two methods based on unithood and termhood respectively in this task. Unithood based one we use is (Frantzi and Ananiadou1996, Frantzi and Ananiadou1996), and termhood based one we use is compound noun based method (Nakagawa1997, Nakagawa1997).

We also participate the role analysis task. Our method for role analysis task is basically a hand-coded pattern driven method for extracting three roles defined in the task. We assign the weight to each triplet according to the weight assigned to the corresponding noun of each role by our compound noun based ranking method(Nakagawa1997, Nakagawa1997).

In section 2, we describe our term recognition system. We briefly describe our role analysis system in section 3. Section 4 is our conclusions.

2 Automatic Term Recognition Task

2.1 Overview of Term Recognition System

A term recognition system, in general, consists of three sub-systems, namely 1)candidate picking up,

2)ranking, and 3)selection, as shown in Figure 1.

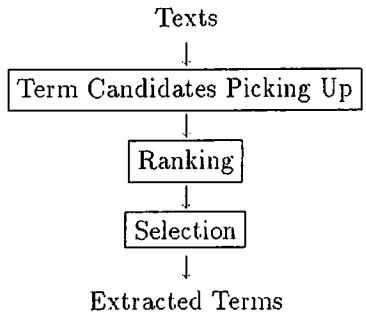


Figure 1: Structure of Term Recognition System

In the following, we sketch each of these three subsystems along with the previous works.

Term Candidates Picking Up Sub-system

A word based candidate of term has been a noun or a compound noun. In these days, more complex structures like noun phrase, collocation consisting of noun, verb, preposition, determiner, and so on, become focused on(Smadja and Mckeown1990, Smadja and Mckeown1990; Frantzi and Ananiadou1996, Frantzi and Ananiadou1996; Evans1996, Evans1996; Hisamitsu and Nitta1996, Hisamitsu and Nitta1996; Shimohata et al.1997, Shimohata et al.1997). All of these are good candidates of terms in a document or a specific domain because all of them have a strong unithood. Needless to say, but as for complex terms like compound words or collocation, we put the following basic assumption:

Assumption 1 *Complex terms are to be made from existing simple terms.*

After POS tagging done by morphological analyzers, the above mentioned complex structure is extracted as a candidate of term. (Ananiadou1994, Ananiadou1994) proposes the way to extract word compounds as terms. (Hisamitsu and Nitta1996, Hisamitsu and Nitta1996) and (Nakagawa1997, Nakagawa1997) concentrate their efforts on compound nouns.

Ranking Sub-system

In order to extract domain specific terms from candidates of term extracted in Term Candidates Picking Up Sub-system, we have to rank them. This ranking has been developed as key word weighting like tf·idf which is widely used in IR. According to (Kageura and Umino1996, Kageura and Umino1996), the frequency information related to the word, like tf·idf, is an approximation of termhood. Obviously, termhood implies semantic weight, and the basic idea is that the frequency information about the word reflects on the semantic importance of the word. On the other hand, ranking methods based on unithood are also intensively studied. For instance, various kinds of statistic information about words co-occurrences which are used to extract promising candidate terms that are in the form of collocation(Smadja and Mckeown1990, Smadja and Mckeown1990; Frantzi and Ananiadou1996, Frantzi and Ananiadou1996; Shimohata et al.1997, Shimohata et al.1997), are of this type. Among them, C-value(Frantzi and Ananiadou1996, Frantzi and Ananiadou1996), entropy(Shimohata et al.1997, Shimohata et al.1997), mutual information(Church and Hanks1990, Church and Hanks1990), etc. are promising.

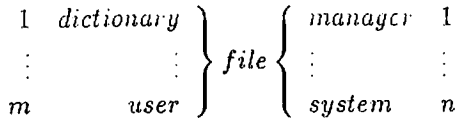
Selection Sub-system

As for selection from ranked candidates, we find very general scheme such as likelihood test(Dunning1993, Dunning1993). However, very few work has been done to directly target genuine term selection process. At the first glance, a selection by the predetermined threshold is, seemingly, simple and powerful. However, the problem is the way to determine the threshold which works equally well on unseen documents. Then, to find another selection method different from simple thresholding is also a challenging problem.

2.2 Compound Noun Based System

Obviously, the relation between the simple term and complex term in which the simple term is included is very important. In my knowledge, this relation has not been paid enough attention so far. (Nakagawa1997, Nakagawa1997) is the method to use this relation.

Here, we focus on compound nouns among various types of complex terms. In technical documents, the



$Pre(\text{"}file\text{"}) = m$ and $Post(\text{"}file\text{"}) = n$

Figure 2: An example of *Pre* and *Post*

majority of domain specific terms are complex terms, more precisely, compound nouns. In spite of huge number of technical terms being compound nouns, not so many number of simple nouns contribute to make these compound nouns. Considering this fact, a new scoring method which measures the importance of each simple noun is proposed. This scoring method for a simple noun measures how many distinct compound nouns use the simple noun as their parts in a given document or a set of documents. *Pre*(simple noun) and *Post*(simple noun) are introduced for this purpose, and defined as follows.

Definition 1 *In the given volume of text, $Pre(N)$, where N is a noun appeared in the text, is the number of distinct nouns that come just before N and make compound nouns with N, and $Post(N)$ is the number of distinct nouns that come just after N and make compound nouns with N.*

The key point of this definition is that $Pre(N)$ and $Post(N)$ count not the number of occurrences of word coming just before or after N, but the number of distinct words coming just before or after N. That means that *Pre* and *Post* don't measure mere surface statistics of compound nouns, but do measure how the writer of the technical text understands and expresses the contents of the target system or domain. In this sense, *Pre* and *Post* are basically based on termhood. Figure 2 shows an example of *Pre* and *Post*.

Next, this scoring method is to be extended to score compound nouns. For the given compound noun $N_1N_2\cdots N_k$ where N_is are simple nouns, the score of importance of $N_1N_2\cdots N_k$ namely $Imp(N_1N_2\cdots N_k)$ can be defined here as follows.

$$Imp(N_1N_2\cdots N_k) = (\prod_{i=1}^{k}((Pre(N_i)+1)\cdot(Post(N_i)+1)))^{\frac{1}{2k}}$$

The powering factor $\frac{1}{2k}$ makes *Imp* less dependent on the length of the compound noun: $N_1N_2\cdots N_k$.

2.3 Nested Collocation System

One of the famous approach based on statistics about linguistic structure is the ranking method based on nested collocation(Frantzi and Ananiadou1996, Frantzi and Ananiadou1996). It first extracts all candidates of collocation. Then, it uses the measure they call **C-value** defined by the following formula:

$$\text{C-value(a)} = (\text{length(a)} - 1)(\text{freq.(a)} - \frac{\text{t(a)}}{\text{c(a)}})$$

where, "a" is a collocation, t(a) is frequency of "a" in longer candidates of collocations, and c(a) is number of longer candidates of collocations including "a".

In this method, generally speaking, collocations stably used in documents have a high C-value and are ranked high. But, in fact, things are more complicated. For instance, collocation "Wall Street" seems to be ranked high. However, if "Wall Street" almost always appears as a part of " Wall Street Journal", the latter should be ranked higher and the former should rather be ranked much lower. C-value is devised to rank candidates of collocation according to this intuitive idea. In other words, C-value depends upon how stable the given collocation is used. Therefore, it is a unithood oriented approach.

2.4 Window Method

As for selection sub-systems, we focus on the statistical value in the window on ranked candidates as local statistics. In this method, which we call *window method* henceforth, a window with a certain width is moving from the position of the highest ranked candidate term down to the position of the lowest ranked candidate term. For instance, a window with width=3 is depicted in Figure 3.

A window's position is characterized by the highest *Imp* value or C-value of compound noun or simple noun within the window. For instance, in Figure 3, the window's position corresponds to 17.18.

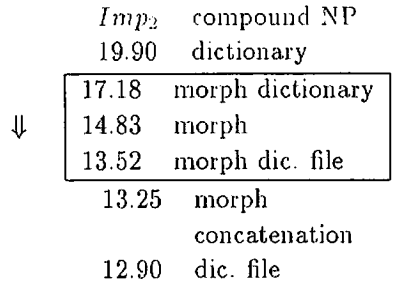


Figure 3: Window with width=3

Now we use some statistical values we obtain from the contents of window along with the window moving downward, to decide whether the nouns in the window is selected as a genuine term or not. Here, a genuine term is defined as a term picked up by hand. Among several kinds of statistical value, we pay our attention to the genuine term ratio in the window, GTR in short, which is defined as follows.

$$GTR = \frac{\sharp(\text{genuine term in the window})}{\text{window width}}$$

where $\sharp X$ means the number of members in the set denoted by X.

The reason why we pay our attention to GTR is that GTR is, in fact, high in the windows of high Imp value. Moreover, a number of genuine terms seems to increase as the length of text increases. In addition, a number of all simple and compound nouns in a text also increases as the text becomes longer. Therefore, GTR is likely to be less dependent on the length of text. We also pay our attention to the compound noun ratio in a window, COMPWR in short, defined as follows.

$$COMPWR = \frac{\sharp(\text{compound nouns})}{\text{window width}}$$

The reason why we pay our attention to COMPWR is that the majority of genuine terms in technical texts are usually compound nouns in the technical texts we investigated. By considering the nature of GTR and COMPWR, we reach the following expectation. In the window whose corresponding Imp value is high, the majority of simple and compound nouns within in the window are genuine terms, and at the same time, the majority of them are compound nouns, too. Therefore, we expect high relevance between GTR and COMPWR, and that has already been experimentally proven(Nakagawa1997, Nakagawa1997). Moreover, in our experiments, we confirm that among simple and compound nouns having a high Imp value, the majority of terms are compound nouns. Thus, it is reasonable to use COMPWR value instead of Imp values themselves for selection by the given threshold. Therefore, what we have to do is to find an optimum, or at least a sub-optimum, threshold of COMPWR to select the genuine terms. In a selection process, the candidate of term which is located at the center of the window is selected if COMPWR of the window is larger than the pre-determined threshold, otherwise that candidate is not selected. In our experience of using our window method, the optimum threshold heavily depends on the academic area treated in texts, or even individual text. Then, we adopt here the averagely well working value, say window width of 10 and COMPWR of 0.3.

2.5 Combination

The next step is the combination of the above described two methods. In fact, there are a number of ways to combine two methods into one method, and it is hard to find the best combined method. Thus, we choose the simplest way which is to combine the result of each method. Still we have several variations to combine the terms extracted by each method. We have already examined these two methods on several Japanese corpora and an English corpus in a certain technical and/or academic domain. In that experimentation, we found that the extracted terms by nested collocation method and those by compound noun based method are not much different, rather similar as a whole, even though the characteristics of extracted terms are little bit different. Then, in this task, we select terms that are extracted by both of these two method as the final result of our system.

2.6 Evaluation

Here we show the results of experimental evaluation of our system, and discuss the characteristics of our

File	tagged	untagged
Total	18608	16764
AF	5554	4013
AA	16673	14596
AI	10484	9580
AP	355	476
AB	280	527

Table 1: Number of Elements Matched against Manual Candidates

File	tagged	untagged
AFP	29.85	23.94
AAP	89.60	87.07
AIP	56.34	57.15
APP	1.91	2.84
ABP	1.50	3.14
AFR	62.86	45.42

```
1st char - A : Handmade candidates
2nd char - F : Full much ;
A :All inclusive ;
I :Result Include Term Candidates ;
P :Result is a Part of Term Candidates ;
B :Result is both I & P
3rd char - P : Precision ; R : Recall
```

Table 2: Recall and Precision based on Manual Candidates

system.

In our system, values of AFP are lower, but values of AIP are higher. The reason is that our system tends to extract longer terms because of its scoring scheme. The result of AF and AI don't include 1837 candidates of handmade candidates. 1837 candidates which couldn't be extracted are mainly KATAKANA words, original English words, and adjective+noun type collocations. The next point is that our extracted terms are longer, because both of two ranking methods we adopted prefer longer terms. If we adopt other ranking methods which prefer short compound noun or simple noun, the result would become totally different. In compound noun based ranking, it could be done through the definition of *Imp* function.

3 Pattern Driven System for Role Analysis Task

3.1 Problem and Our Principle

NTCIR's role analysis task seems to be simple at the first glance. However, in developing the role analysis system, we realized its definition was quite tough to encode as a computer program. Its definition of subject, method and action/process are exemplified as variety of linguistic forms. Of course, it is possible to acquire these linguistic forms with machine learning technologies which have been rapidly developed in these years. But the real difficulty is residing in the vagueness of these definitions. For instance, in the following sentence:

"この理論が技術論文からの情報抽出を可能にするシステムの実装方法を示唆した"

(This theory suggested how we implement the system which can extract information from technical papers.)

the first possibility of extracted roles is

(1)
subject = "how to implement ....."
method = "this theory"
action = "suggest"

Obviously these extracted roles are, at least as technical contents, almost vacant. Then the following roles have much more relevant than the previous ones.

(2)
subject = "the system which can extract information from technical papers"
method = "this theory"
action = "implement" or "suggesting implementation"

We need variety of common sense and knowledge about a target academic field to successfully select (2). Even if we use machine learning technologies, a great amount of human encoded teaching data is needed to learn meaningful results. Unfortunately, we didn't have enough time to do it. Then, in this time, we simply gathered hand-coded linguistic patterns from the given texts. In gathering process of linguistic patterns, we have always been annoyed by the vagueness of definition of roles. We really think

this type of role analysis task is useful to extract academic and/or technical information in this granularity. We have to avoid discussions for too sophisticated definition. But it does not mean it is useless to discuss the definition of each role. In my opinion, usefulness of each role definitely depends on what kind of purpose the extracted information is for.

Anyway it is hard to describe the principle on which we develop linguistic patterns to extract roles because of vagueness of definition of roles. However, roughly speaking, our principle is to extract content oriented information. For instance, in the previous example, we prefer (2) to (1). Actually, we did not exclude linguistic patterns which pick up "提案する (propose)", "述べる (describe)", etc that are regarded not as essential content information but as a formal way of description in writing academic papers, because we are afraid of loosing many important information by restricting linguistic patterns too much. To cope with this problem, We try to give these formal descriptions low weight, as described later.

We show linguistic patterns we developed to extract two or three kinds of roles from the given sentence in 3.2. Then, we describe how to rank extracted roles in 3.3.

3.2 Patterns

3.2.1 Pre-processing

Before applying linguistic patterns we describe later, we pre-process sentences in the following manner.

Sa-Hen-noun

If we have a verb phrase:" の + サ変名詞 (*no*(of) + sa-hen-noun)" in the title, the "サ変名詞 (sa-hen-noun)" becomes categorized as a verb. By this, we can pick up sa-hen-noun which is a very content oriented word, as action/process.

Elimination of non-action verb

We omit verbs that are not to be regarded as action/process in our sense, such as "N という (call as)" , "調査する (investigate), "試みる (try)", etc.

Connecting noun phrases

If we have JNA N pattern, we unify them as a single N. If we have N SCC("の"(of)) N pattern, we also unify them as a single N. These connecting procedures are applied recursively.

For instance,
直観的な (intuitive):JNA, 推論 (inference):N,
の (of):SCC, 枠組 (scheme):N
becomes
直観的な推論の枠組 (intuitive inference scheme):N

3.2.2 Linguistic Patterns

Firstly, we define the linguistic patterns for extracting Method, Action/Process and subject, respectively. Words or phrases extracted by these patterns are candidates of contents of the corresponding role.

Method is defined as "N(noun) + に +(基づき | 基づく | より | よって)",
or "N + を +(用いた | 用いて | 採用して | 利用して | 使用して)"
or "N + (から | で | を | は)"
or "*(wild card) + (法 | モデル | 説 | 論 | システム | 式)

subject is defined as "N + (を | で | が | は)"
or "N + に + (対して | 対し | ついて)"

Action/Process is defined as "V(verb)".

Secondly, we show patterns including Method, Action/Process and subject defined above, which finally extract the contents word or phrase that corresponds to each role of Method, Action/Process or subject.

- 'Method+(により | によって | を用いて)+Action/Process+subject"
- "subject+(は | が | を)+Method+により+Action/Process"
- "Method+により+subject+(が | を)+Action/Process"
- "subject+ を +Method+により+Action/Process"
- "Method+による+subject+(について | が | を | に)+Action/Process"
- "Method+(を用いた | を用いて)+subject+(が | を)+Action/Process"

- "subject+(が | に対して)+Method+から+Action/Process"
- "subject+(は | が)+Method+によって+Action/Process"
- "Method+によって+subject+(について | が | を | に | は)+Action/Process"
- "Method+から+subject+(が | を)+Action/Process"
- "subject+(に対し | に対して | を | について)+Method+を用いて +Action/Process"
- "Method+に基づく+subject+(を | は | に)+Action/Process"
- "Method+に基づき +subject+(を | について)+Action/Process"
- "Method+(で | を採用して | を利用して | を使用して)+subject+ を +Action/Process"
- "subject+ を +Method+ で +Action/Process"
- "subject+ を +Method+から+Action/Process"

3.3 Ranking System

Since we are required to select five best triplets, we have to assign a weight to each extracted triplet. For this, we use *Imp* value of component noun, that is calculated with the method described in section 2.2. Then, the weight of triplet is the sum of the weight of each component noun. As already said, high *Imp* value of compound noun CN means that CN is an important technical term. Thus, this weighting scheme gives high weights to content oriented triplets. Of course this weighting method is too naive to extract semantically important triplets. The more concrete definition of desired characteristics of triplets would give us more valid weighting scheme.

4 Conclusions

As for automatic term recognition task, our system shows a strong performance in extracting long compound terms because our method counts heavily the length of compound terms. However, we need more sophistication on extracting scheme of short or simple terms.

As for role analysis task, our system is very naive and primitive one, and does not show satisfactory results. Our system could be improved by expanding and enriching the linguistic patterns. However, this improvement should be done with machine learning technology in terms of the state of the art.

Acknowledgment

Two of my students at Yokohama National University, Takaya Saito and Hirokazu Ohata, have done a lot of work to implement our term extraction system and role analysis system. Our experiment is totally depends on their effort.

Reference

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *COLING'94*, pages 1034 – 1038.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22 – 29.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):62 – 74.

C. Zhai Evans, D.A. 1996. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of 34th ACL*, pages 17 – 23.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *COLING'96*, pages 41 – 46.

Toru Hisamitsu and Yoshihiko Nitta. 1996. Analysis of japanese compound nouns by direct text scanning. In *Proceedings of COLING'96*, pages 550 – 555.

K. Kageura and B. Umino. 1996. Methods of automatic term recognition:a review. *Terminology*, 3(2):259 – 289.

H. Nakagawa. 1997. Extraction of index words from manuals. In *Proceedings of RIAO'97*, pages 598 – 611.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. 1997. Retrieving collocations by co-occurences and word order constraints. In *Proceedings of 35th ACL*, pages 476 – 481.

Frank A. Smadja and Kathleen R. Mckeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th ACL*, pages 252 – 259.