Construction of Grammar-Based Term Extraction Model for Japanese

Koichi Takeuchi
Okayama University
3-1-1, Tsushimanaka,
700-8530, Okayama,
Japan,
koichi@cl.it.okayama-u.ac.jp

Kyo Kageura
Teruo Koyama
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
101-8430, Tokyo,
Japan,
kyo,t_koyama@nii.ac.jp

Béatrice Daille
IRIN - Universite de Nantes
2, rue de la Houssiniere, BP 92208
F44322, Nantes, Cedex 3,
France,
Beatrice.Daille@irin.univ-nantes.fr

Laurent Romary
LORIA
Campus Scientifique, B.P. 239,
54506, VANDOEUVRE-les-NANCY Cedex,
France,
Laurent.Romary@loria.fr

Abstract

In this paper we propose a grammar-based term extraction model for Japanese toward construction of multilingual term dictionary. The proposed approach evaluates termhood using morphological patterns of terms. Most of terms in Japanese consist of compound nouns or simple phrases, but detailed grammatical patterns for term extraction have not been constructed because of their complex compounding mechanisms. Applying detailed morphological investigation to Japanese compounding, we make sure their structures of word formation that can be comparable to those of Indo-European terms such as English and French.

1 Introduction

The difficulty of term extraction is how to evaluate termhood. Since the role of a term is to denote a specific concept in a domain, termhood should be evaluated by the following two sides: the first is the strength of unity that is called as unithood (Nakagawa, 2000) of element words as a term, and the second is the domain specialty of the word.

Most of previous approaches (see (Kageura and Koyama, 2000)) took the latter, but the former, i.e., unithood of a term, is very important because unithood is directly related to the mechanisms of creation of new concept by composing words, which should reflect the integrity of the new concept.

Recently some of term extraction work focus on the unithood of words. Nakagawa (2000) incorporated a statistical method, while Jacquemin (1996), Ananiadou (1994), Daille (2003) successfully applied detailed morpho-syntactic pattern-based approaches to evaluating unithood on term extraction of French and English.

We construct grammatical patterns for Japanese term extraction on the framework of ACABIT system (Daille, 2003) in order to make comparable morpho-syntactic grammar to the Indo-European languages such as English and French. Clarifying the comparable structure of terms enables us to have a basic framework of multilingual term dictionary and multilingual term extraction.

2 Grammatical Patterns

Grammatical patterns are defined by means of three features: unit of integration, grammatical category, and origin of a word. Unit of integration means a word formation class according to morphology. The smallest unit of integration in compounds and phrases is a morpheme. Grammatical category means grammatical function of morpheme like a part-of-speech. Origin of a word means a type of word that has a great influence to compounding rules in Japanese. There are two types of origins, which are a word originated in old Japanese **(OJ)** and a word imported from Chinese or other languages **(IM)**. Each group of words tends to connect the same group of words when it composes a compound noun.

2.1 Grammatical Categories

Grammatical categories are defined according to the functionality of connecting the other morphemes, therefore, the categories do not exactly correspond to conventional categories of grammar. The categories, however, are comparable to Daille's approach from the functional point of view.

⟨**Noun**⟩**(OJ)(IM):** This includes not only a nominal noun but also a deverbal noun,[1] an adjectival noun,[2] a number and a symbol. Morphemes in **Noun** can compose compounds without grammatical limitation.

⟨**AdjStem**⟩**(IM):** This denotes an adjectival stem whose function is to modify the next morpheme. The morphemes in **AdjStem** does not come to head in compound nouns or phrases because it is a stem.

⟨**Prefix**⟩**(IM):** This denotes a prefix in Japanese.

⟨**Suffix**⟩**(IM):** This denotes a nominal suffix that does not change the grammatical category of compounded word.

⟨**AdjStemSuffix**⟩**(IM):** This denotes a suffix that derives a stem of adjective. Because of this function, the morpheme in **AdjStemSuffix** takes two morphemes at the both sides in order to be a word.

⟨**NomSuffix**⟩**(OJ):** This denotes a suffix that derives a noun connecting to a stem of adjective, i.e., **AdjStemOJ** whose type is original Japanese (OJ).

⟨**InfV**⟩**(OJ):** This denotes a verb with inflection.

⟨**InfA**⟩**(OJ):** This denotes an adjective with inflection.

⟨**AdjStemOJ**⟩**(OJ):** This denotes a stem of adjective whose type is original Japanese. The morpheme **AdjStemOJ** needs to be followed by **NomSuffix** in order to be a word.

⟨**ConOf**⟩**(OJ):** This denotes a post positional particle whose function is to compose a phrase connecting two nouns. Roughly speaking, the grammatical function of the particle corresponds to 'of' in English.

[1] Deverbal noun has both characteristics of noun and verb. Basically, it acts as noun but it can be a verb followed by an auxiliary verb.

[2] Adjectival noun has both characteristics of noun and adjective.

2.2 Grammatical Patterns of Japanese terms

After we describe grammatical patterns of words as a stable unit, we show the patterns of compounds and phrases as a term based on the patterns of words.

Patterns of words are denoted as **WordIM** and **WordOJ** according to the types of origins of words. The patterns are described in Backus Naur Form (BNF).

$$
\begin{aligned}
&\langle WordIM\rangle \\
&\quad ::= \langle Noun\rangle \mid \langle AdjStem\rangle\langle Noun\rangle \mid \\
&\qquad \langle Prefix\rangle\langle Noun\rangle \mid \langle Noun\rangle\langle Suffix\rangle \mid \\
&\qquad \langle Noun\rangle\langle AdjStemSuffix\rangle\langle Noun\rangle \mid \\
&\qquad \langle AdjStem\rangle\langle AdjStemSuffix\rangle\langle Noun\rangle \mid \\
&\qquad \langle Prefix\rangle\langle Noun\rangle\langle Suffix\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\langle WordOJ\rangle \\
&\quad ::= \langle InfV\rangle\langle Suffix\rangle \mid \\
&\qquad \langle AdjStemOJ\rangle\langle NomSuffix\rangle \mid \\
&\qquad \langle InfV\rangle\langle Noun\rangle \mid \langle InfV\rangle\langle SuffixV\rangle \mid \\
&\qquad \langle InfA\rangle\langle Noun\rangle
\end{aligned}
$$

Patterns of compounds are also divided into two types that are related to **WordIM** and **WordOJ**, respectively. The former type of words **(IM)** can compose compounds without limitation because of their strong unithood, while the later words **OJ** can compose compounds within limited patterns. Since the restriction is not a grammatical theory but a kind of empirical knowledge.

$$
\begin{aligned}
\langle CompIM\rangle \quad ::= \quad & \langle CompIM\rangle\langle WordIM\rangle \mid \\
& \langle WordIM\rangle\langle CompIM\rangle \mid \\
& \langle WordIM\rangle
\end{aligned}
$$

$$
\begin{aligned}
\langle CompOJ\rangle \quad ::= \quad & \langle WordOJ\rangle \mid \\
& \langle InfA\rangle\langle NomSuffix\rangle\langle Noun+\rangle \mid \\
& \langle Noun\rangle\langle InfV\rangle\langle Suffix\rangle \mid \\
& \langle Noun\rangle\langle InfV\rangle\langle Noun\rangle
\end{aligned}
$$

Where $\langle Noun+\rangle$ means

$$\langle Noun+\rangle ::= \langle Noun\rangle \mid \langle Noun+\rangle\langle Noun\rangle$$

Patterns of terms consist of those of compounds and simple phrases. The grammatical

patterns of simple phrases are limited to only "A no (of) B" since this pattern is the simplest phrase keeping unithood. Therefore grammatical patterns of terms we defined are as follows.

$$\langle Phrase\rangle ::= \langle CompIM\rangle\langle ConOF\rangle\langle CompIM\rangle$$
$$\langle TERM\rangle ::= \langle CompIM\rangle|\langle CompOJ\rangle|\langle Phrase\rangle$$

3 Preliminary Experiment

We made two types of experiments for Japanese ACABIT system. The first evaluation is about the coverage of morphological patterns. We input technical terms to Japanese ACABIT and check the rate of acceptability of the patterns. The second experiment is term extraction performance. For this experiment, we use the set of abstracts and author's keywords distributed by NII(Kageura et al., 2000).

3.1 Overview of ACABIT system

Japanese ACABIT system detects nominal expressions by local grammar rules we elaborated in Section 2.[3] Since the input of ACABIT needs POS tagged information, we apply Japanese morphological analyzer ChaSen (Matsumoto et al., 1999). The output of ACABIT is a list of two-words candidate terms ranked according to the degree of representativeness of the corpus using the log-likelihood statistics (Dunning, 1993).

3.2 Coverage of patterns

We prepare three kinds of technical terms: 1) a technical term dictionary of information processing (ipdic), 2) a term dictionary in computer domain (comdic) and 3) a list of author's keywords in artificial intelligence domain (jsai) (Kageura and Koyama, 2000).[4] All terms are analyzed by ChaSen first. After this process, we evaluate quantitative nature of terms about number of one word terms and number of phrasal terms

Table 1 shows that the ratio of one word term is not small, i.e., more than 10% for every source. So the upper bounds for extracting complex terms are 86.4% (ipdic), 88.4% (comdic) and 84.4% (jsai). Table 2 shows the results of coverage performance of Japanese ACABIT.

[3]One morpheme pattern such as **Noun** in our grammatical category is deleted in this application because one morpheme pattern is too noisy to detect terms.

[4]It is abstracts of Japanese society of artificial intelligence.

Table 1: Statistics of input terms

	One word terms (%)	phrasal terms (%)
ipdic	2207/16275(13.6)	409/16275(2.5)
comdic	4480/38785(11.6)	2366/38785 (6.1)
jsai	658/4206(15.6)	231/4206(5.5)

Table 2: Coverage of Japanese ACABIT

	coverage (%)
ipdic	12195/16275(74.9)
comdic	28162/38785(72.6)
jsai	3056/4206(72.7)

ACABIT works well. The upper bound of this experiment is about from 86% to 89% for these terms. The error types are categorized into two: Variety of term: most errors occur for terms that contain proper nouns. For example, "nyuuton" (Newton) in "nyutonn hou" (Newton method) is annotated proper noun in ChaSen. We just excluded this sort of patterns as they are unlikely to make terms. Errors of annotation of ChaSen: ChaSen makes annotation errors on ambiguous words.

3.3 Keyword extraction

Japanese ACABIT is applied to abstracts in the domain of artificial intelligence in order to show term extraction performance. Assuming that author's keywords are correct terms, we evaluate the performance of ACABIT by comparing extracted terms with author's keywords. Table 3 shows the statistics of author's keywords. According to the table, 68.7% the keywords occur in the abstracts of which 20.1% are one-word keywords. So 2308 words are the upper bound for the extraction experiment.

Table 3: Statistics of author's keywords

	author's keywords (%)
contained in text	2890/4206 (68.7)
one word keyword	582/2890 (20.1)
upper bound	2308/4206 (54.9)

Table 4 shows the results of term extraction comparison to author's keywords. We evaluate precision, and correct rate to upper bound of author's keywords.

All extracted terms are evaluated. In Table 4, ACABIT works well because 71% to upper bound are successfully extracted. The precision

Table 4: Results of term extraction

	author's keywords (%)
precision	1639/ 23494(7.0)
recall to upper bound	1639/2308(71.0)

is, however, poor. When we filter out the candidates with low log-likelihood value terms, precision becomes 20.8% (and recall with respect to upper bound was 25.0%). Example correctly extracted words are "iden-teki-arugorizumu" (generic algorithm), "chishiki-beisu" (knowledge base), and wrong examples are "hon-ronbun" (this paper), "hon-kenkyu" (this research) "hissya-ra" (authors). The words that are extracted wrongly are high frequency words in target text.

4 Discussions

Comparing with the results of French term extraction (Jacqumin et al. 2002) using ACABIT system, termhood precision and recall are 79% and 67.5% in French (Daille, 1996), respectively. A precision in English is 67% (Savary, 2001). The recall rate of Japanese ACABIT is almost the same as the results in French, thus, we can conclude the proposed morpho-syntactic patterns have good coverage for technical terms. The precision of keyword extraction is poor, however, it is not the current target since the proposed approach is intended to evaluate unithood of terms without domain dependence.

The current approach does not extract uniterms that account for a fifth in keywords (Table 3) since uniterms have naturally strong unithood and the proposed approach should not be applied to them. The problems of uniterms and low precision should be solved by investigating approaches how to evaluate specificity of words in a domain.

5 Conclusions

We proposed a pattern-based term extraction method and showed the experimental results. We tried to extract terms focusing on the unity of intra-term structure by morpho-syntactic patterns without domain dependence. From the experimental results, the constructed patterns worked well for coverage. This means that the proposed morpho-syntactic set is enough to be applied to a basic framework to describe grammatical structure of multilingual terms.

So far, we have only discussed the intra-structure of terms at the level of morpho-syntactic patterns of terms. To develop a multilingual term extraction model we need to delve further into two sides: The first is structural analyses of terms in sentences to evaluate specificity of words in a domain. The second is the multilingual correspondence between the terms at the level of individual correspondences of lexical items and POS-categories.

References

S. Ananiadou. 1994. A methodology for automatic term recognition. In *In Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 1034–1038.

Beatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Application to Language*, pages 49–66.

B. Daille. 2003. Terminology Mining. In M.T. Pazienza, editor, *Information Extraction in the Web Ea*, pages 29–44. Springer.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–76.

C. Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In E. Riloff S. Wermter and G. Scheler (Eds.), editors, *In Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.*

K. Kageura and T. Koyama. 2000. *Terminology, vol.6 no.2.* John Benjamins Publishing Company.

K. Kageura, M. Yoshioka, and T. Koyama. 2000. Towards a common testbed for corpus-based terminology: Lexical units, POS information and their utilisation. In *First International Symposium on Advanced Informatics (AdInfo2000).*

Y. Matsumoto, A. Kitauchi, T. Yamashita, and Y. Hirano, 1999. *Japanese Morphological Analysis System ChaSen 2.0 Users Manual.*

H. Nakagawa. 2000. Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2):195–210.

A. Savary. 2001. Etude comparative de deux outils d'acquisition de termes complexes. In *In TIA-2001 (4th meeting Terminology and Artificial Intelligence).*