Term Recognition by Using Different Field Corpora

Kiyotaka UCHIMOTO[†] Satoshi SEKINE[‡] Masaki MURATA[†]

Hiromi OZAKU[†] Hitoshi ISAHARA[†]

[†]Communications Research Laboratory, M. P. T. [‡]New York University

{uchimoto|murata|romi|isahara}@crl.go.jp sekine@cs.nyu.edu

Abstract

We participated in the term recognition task, one of the subtasks covered by the NTCIR tmrec group. In this paper, we present a system used in this task and evaluate the term recognition results of this system. We believe that terms could be words that characterize the field's data and have the following three features: (1) They frequently appear in the target field's corpus. (2) They are not common terms in the target field. (3) They less frequently appear in the other fields' corpora. Our system uses different field corpora and recognizes these features as terms. We extracted a term list by using two kinds of field corpora, the NACSIS Academic Conference Database and the MAINICHI newspaper database. We then analyzed the difference between our term list and Manual-Candidates made by the NTCIR tmrec group. In this paper, we clarify what should be considered when recognizing terms. Furthermore, through comparative experiments based on Manual-Candidates, we verify the importance of indices which are used to extract a term list.

key words: term recognition, field, corpora, term frequency, document frequency

異分野コーパスを用いた用語抽出

内元 清貴[†]　関根 聡[‡]　村田 真樹[†]　小作 浩美[†]　井佐原 均[†]

[†] 郵政省通信総合研究所　[‡] ニューヨーク大学

{uchimoto|murata|romi|isahara}@crl.go.jp　sekine@cs.nyu.edu

概 要

我々は 1999 年 3 月に NACSIS により主催された NTCIR コンテストの用語抽出タスクに参加した。このタスクは NACSIS 論文データベースから取り出された 1,870 の抄録を対象に用語を抽出するというものである。本稿では我々が用いたシステムの評価結果について報告する。

我々は用語を分野に特有の語であり、以下の三つの性質を持つものと考えている。(1) 対象の分野に多く出現する。(2) 対象の分野の一般的な語ではない。(3) 他の分野にあまり出現しない。我々のシステムは、異なる分野のコーパスを利用し、このような性質を持つ形態素あるいは形態素の連接を用語として抽出する。コンテストでは NACSIS 論文データベースと毎日新聞 (94、95 年) データベースの二つの分野のコーパスを用いて抽出したものを用語リストとして提出した。その後主催者側より用意された手作業正解候補と我々の提出したリストとを比較して違いを分析することによって以下の点が明らかとなった。

品詞について 正解候補はすべて名詞である。一方、我々のシステムでは品詞に制限を設けずに用語を抽出した。動詞の候補はほとんどが「離散化する」などのサ変動詞であり、「離散化」のように名詞化したものが正解候補となっている場合が多かった。

名詞句について 正解候補には「一階の論理」、「統合的なプラニング」、「説明に基づく類推」など「A の B」「形容詞＋名詞」「修飾句＋名詞」の形の名詞句が含まれているが、我々の定義では除外していた。このような形のものも十分用語となり得るものである。しかしながら、このような形の名詞句をすべて用語候補として抽出するようにすると、余計なものまで抽出してしまう危険性が高くなる。用語となる名詞句の受け側の名詞の性質、係り側と受け側との関係に規則性があるかどうかを調べる必要がありそうである。

複合語について 我々の定義では複合語あるいは、名詞、カタカナ、アルファベット、未定義語、接頭辞、接尾辞の連接を用語の候補とすることにしている。この定義と評価関数の性質から「未知複合語意味推定システム」や「アブダクティブ論理プログラム処理系」のように長い複合語も用語として抽出する傾向がある。このような複合語はどの分野のコーパスにもあまり出現せず、大抵は一つの分野に一回ないし二回程度しか現れないため、用語かそうでなかの判断が難しい。このような場合、人間は前後の文脈から判断していることが多い。こういった複合語も用語候補として扱う場合、前後の文脈から用語かどうかを判断できる機構が必要となる。

また、手作業正解候補を正解とみなした比較実験で、用語を抽出する際には対象分野における出現頻度だけでなく、対象分野における文書頻度、他の分野における出現頻度が重要な役割を果たしていることが分かった。しかし、実験で用いたコーパスは NACSIS 論文データベースと毎日新聞 (94、95 年) データベースの二種類と少なかったため、分野の種類の違いが重要であるかどうかまでは分からなかった。これについては今後より多くの分野のコーパスを用いる、あるいは新聞データベースを分野ごとに分割して検証したい。

キーワード　用語抽出、分野、コーパス、出現頻度、文書頻度

1 Introduction

We participated in the term recognition task, one of the subtasks covered by the NTCIR tmrec group, in March, 1999. The goal of this task is to recognize terms which characterize data collection on the subject of artificial intelligence. The data collection consists of 1,870 abstracts extracted from the NACSIS Academic Conference Database. In this paper, we present a system to perform this task and the term recognition results of this system.

We believe that terms could be words that characterize the field's data and have the following three features:

1. They frequently appear in the target field's corpus.
2. They are not common terms in the target field.
3. They less frequently appear in the other fields' corpora.

Our system uses different field corpora and recognizes these features as terms. We extracted a term list by using two different field corpora: the NACSIS Academic Conference Database and the MAINICHI newspaper database. The NTCIR tmrec group made two term candidates, that is manually extracted term candidates (Manual-Candidates) and elements listed in the index part of an encyclopedia on artificial intelligence (Index-Candidates). In this paper, we analyze the difference between the Manual-Candidates and our term list, and we clarify what should be considered when recognizing terms. Furthermore, we verify the importance of indices which are used to extract a term list by doing comparative experiments based on Manual-Candidates.

2 Term recognition model

This model recognize terms in tagged-data and untagged-data. We define that a term consists of a morpheme or several morphemes. We also define a morpheme to be the same as those defined in NACSIS tagged-data and JUMAN. We extract morphemes and compound words as term candidates, and we judge whether they are terms or not by using an evaluation function. A compound word is defined to be a conjunction of nouns, katakana strings, letters, unknown words, a prefix word, and a suffix word. Of course, the other conjunctions could be terms. We will discuss what could be terms in Chapter 4.

There are many candidates which meet the above definition. In our model, the candidates that satisfy the following features are recognized as terms:

1. They frequently appear in the target fields's corpus.
2. They are not common terms in the target field.
3. They less frequently appear in the other fields' corpora.

We define the following evaluation function for recognizing these features. Term candidate t_i is recognized as a term when the value estimated by the function f_{ij} is over the threshold. In the following equation, items TF_{ij}, IDF_{ij}, and IFF_i correspond to features 1., 2., and 3., respectively.

$$\begin{aligned} f_{ij} &= TF_{ij} \times IDF_{ij} \times IFF_i \\ &= TF_{ij} \times log\left(\frac{N_j}{DF_{ij}}\right) \times log\left(\frac{N}{FF_i}\right) \end{aligned} \quad (1)$$

Each term in Eq. (1) is given as follows:

- $IDF_{ij} = log\left(\frac{N_j}{DF_{ij}}\right)$.
- N_j: Number of documents included in the corpus of field F_j.
- DF_{ij}: Number of documents which contain term candidate t_i in the corpus of field F_j. (document frequency)
- TF_{F_j}: Number of occurrences of term candidate t_i in the corpus of field F_j. (term frequency)
- $IFF_i = log\left(\frac{N}{FF_i}\right)$.
- N: Number of fields.
- FF_i: Number of fields which contain term candidate t_i.

3 Term recognition algorithm

The algorithm goes through the following steps in order to recognize terms.

1. Text is morphologically analyzed.
   Morphological information, attached to the NACSIS tagged-data and given by JUMAN, is used for tagged-data and untagged-data respectively.
2. Term candidates are extracted.
   Morphemes and conjunctions of morphemes are extracted as term candidates. The conjunctions of morphemes are restricted to compound words when extracting from tagged-data, and they are restricted to conjunctions of nouns, katakana strings, letters, unknown words, a prefix word, and a suffix word when extracting from untagged-data. For example, 10 candidates as shown in Table 1 are extracted from the title "直交型推論に基づく問題解決機構 (A problem solving system based on orthogonal-type reasoning)." In this table, NACSIS(TF) represents the total number of occurrences of each term candidate and NACSIS(DF) represents the number of documents which contain the term candidate in the NACSIS database. MAINICHI represents the total number of occurrences of each term candidate in the MAINICHI database.
3. The frequency of each candidate in each field corpus is counted. If the value estimated by the evaluation function is over the threshold, the candidate is recognized as a term.
   The following function is derived from Eq. (1).

$$f_{iNa} = TF_{iNa} \times log\left(\frac{N_{Na}}{DF_{iNa}}\right) \times log\left(\frac{2}{FF_i}\right) \quad (2)$$

   We used two kinds of field corpora, the NACSIS Academic Conference Database and the MAINICHI newspaper database, which includes articles published in 1994 and 1995. Therefore, the value of item FF_i in Eq. (2) can be 1 or 2, which shows that the rightmost item $log\left(\frac{2}{FF_i}\right)$ does not reflect feature 3., which we mentioned above; namely, "Terms less frequently appear in the other fields' corpora." Fortunately however, the MAINICHI database includes articles in several kinds of fields, so we assumed that when a word frequently appears in the MAINICHI database, it also appears in many fields. We used TF_{iM} instead of FF_i and defined the following simple function f_{iNa} as

$$f_{iNa} = \left(\frac{TF_{iNa}}{DF_{iNa}}\right)^{\alpha} \times \frac{1}{TF_{iM} + 0.5}, \quad (3)$$

   where the symbols Na and M represent the NACSIS and MAINICHI databases. We set α and the threshold to 2 and 1 respectively. On this condition, "直交型推論 (orthogonal type reasoning)" and "問題解決機構 (A problem solving system)" in Table 1 are recognized as terms.

4 Experiment and results

4.1 Evaluation and analysis of our system

We extracted term lists from tagged-data and untagged-data. Two kinds of term candidates, Manual-Candidates and Index-Candidates, were prepared by The NTCIR tmrec group, and our lists are close to Manual-Candidates. We believe that the definition of terms for Manual-Candidates is close to ours. We therefore assumed that Manual-Candidates are correct answers and analyzed the difference between them and our term list. Then, we found the following problems.

- Part-of-speech
  All of Manual-Candidates are nouns. On the other hand, our definition has no restriction on the part-of-speech, so our term lists included verbs, adjectives and so on. Most verb candidates of our lists were SAHEN verbs such as "離散化する (discretize)," and their nominalized forms such as "離散化 (discretization)" were mostly included in the Manual-Candidates.
- Noun phrase
  There are several patterns of noun phrases such as "A-no-B," "adjective+noun," and "modifier+noun," and those three patterns of noun phrases, for example, "一階の論理 (first order formula)," "統合的なプラニング (integrated plan-

Table 1: Example of Term Recognition

	Frequency			Morphological information	
Candidates	NACSIS (TF)	NACSIS (DF)	MAINICHI	Tagged	Untagged (Information given by JUMAN)
直交型推論	3	1	0	NN	名詞 (NOUN)
直交	4,430	2,603	3	NS, K	名詞 (NOUN), サ変名詞 (SAHEN)
型	129,388	65,838	18,156	NN, K	接尾辞 (SETSUJI), 名詞性名詞接尾辞 (NOUN-SETSUJI)
推論	7,371	3,251	28	NS, K	名詞 (NOUN), サ変名詞 (SAHEN)
に	2,661,460	331,752	1,179,760	SCA, W	助詞 (JOSHI), 格助詞 (KAKU-JOSHI)
基づく	15,392	12,847	1,671	VKAbs, W	動詞 (VERB)
問題解決機構	35	25	0	NN	名詞 (NOUN)
問題	62,837	39,197	43,085	NN, K	名詞 (NOUN), 普通名詞 (COMMON)
解決	10,195	8,258	5,997	NS, K	名詞 (NOUN), サ変名詞 (SAHEN)
機構	24,303	15,837	3,821	NN, K	名詞 (NOUN), 普通名詞 (COMMON)

ning)" and "説明に基づく類推 (explanation-based analogical reasoning)" were included in the Manual-Candidates. However, our definition excludes them. Those noun phrases could be terms, but if we extracted all noun phrases as term candidates, there would be some risk of extracting unnecessary ones. We need to investigate the behavior of the rightmost noun in a noun phrase and to investigate if there is a regular relationship between the modifier and the modifiee in a noun phrase.

- Compound words

  In our definition, a compound word is a conjunction of nouns, katakana strings, letters, unknown words, a prefix word, and a suffix word. Due to this definition and the feature of our evaluation function, long compound words such as "未知複合語意味推定システム (a system for inferring meaning of unknown words)" and "アブダクティブ論理プログラム処理系 (an abductive logic programming system)" tend to be recognized as terms. These compound words do not appear frequently in any field corpora. Since they could appear only once or twice in a field, it is difficult to judge whether they are terms or not. In this case, humans mostly judge by referring to context. Such a reference mechanism is necessary in order to correctly recognize these long compound words as terms.

4.2 Evaluation function and accuracy

We evaluate the accuracy by using three kinds of indices: recall, precision, and F-measure based on Manual-Candidates. When the candidate in our term list fully matches one of the Manual-Candidates, it is counted as correct answer. As described in Section 4.1, all of the Manual-Candidates are nouns. We therefore extracted only nouns and compound words from our term lists and evaluated them.

The following four kinds of evaluation functions are used in the following additional experiments. Recall—Precision curves for the tagged-data and untagged-data are shown in Figures 1 and 2, respectively. The increase of threshold tends to increase precision and decrease recall.

- Without the information of the other fields.

$$f^1_{iNa} = \frac{TF_{iNa}}{DF_{iNa}} \quad (4)$$

- The ratio of frequencies in two fields.

$$f^2_{iNa} = \frac{TF_{iNa}}{TF_{iM} + 0.5} \quad (5)$$

- Eq. (2)

$$f^3_{iNa} = TF_{iNa} \times log\left(\frac{N_{Na}}{DF_{iNa}}\right) \times log\left(\frac{2}{FF_i}\right) \quad (6)$$

- Eq. (3) ($\alpha = 2$)

$$f^4_{iNa} = \left(\frac{TF_{iNa}}{DF_{iNa}}\right)^2 \times \frac{1}{TF_{iM} + 0.5} \quad (7)$$

We can see from the figures that the accuracy is higher in the ascending order of f^1_{iNa}, f^2_{iNa}, f^3_{iNa}, f^4_{iNa}. This result shows that document frequency in the target field (DF_{iNa}) and term frequency in the other field (TF_{iM}) contribute to the increase in accuracy. However, we cannot conclude that f^4_{iNa} is higher than that calculated from Eq. (1) because we used only two kinds of corpora in these experiments. So in the future we will carry out the experiments by using corpora in many kinds of fields or by dividing the MAINICHI newspaper database according to fields.

In these experiments, α in Eq. (3) was fixed at 2. We carried out additional experiments with the fixed threshold and plotted the relationship between α and F-measure. The threshold was fixed at the value which led to the best F-measure, that is 0.4 for tagged-data and 0.2 for untagged-data. The F-measure is defined as

$$\mathrm{F-measure} = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}.$$

The plot of F-measure vs. α in Figure 3 shows that the F-measure is highest when α is 7 for tagged-data and 5 for untagged-data. For reference, the recall and the precision are listed in Table 2.

5 Conclusion

We have developed and evaluated a system that can perform the term recognition task, one of the subtasks covered by the NTCIR tmrec group. Our system uses different field corpora, and it is based on a model which recognizes a morpheme or a conjunction of morphemes having the following features as terms:

1. They frequently appear in the target field's corpus.
2. They are not common terms in the target field.
3. They less frequently appear in the other fields' corpora.

We analyzed the difference between our term list and Manual-Candidates prepared by the NTCIR tmrec group, and found that it is important to take into account how to deal with parts-of-speech, noun phrases, and compound words in order to recognize terms. Furthermore, we found that our indices, term frequency, and document frequency in the target field's corpus, and term frequency in other fields' corpora, play an important role in recognizing terms from the results of comparative experiments based on Manual-Candidates. However, we could not determine the relationship between the accuracy and the difference of fields because we used only two kinds of corpora, the NACSIS Academic Conference Database and the MAINICHI newspaper database. In our future work, We will verify the importance of the difference between fields by using corpora in many kinds of fields or by dividing the MAINICHI newspaper database according to fields.

Reference

[1] Sadao Kurohashi and Makoto Nagao. *Japanese Morphological Analysis System JUMAN version 3.6.* Department of Informatics, Kyoto University, 1998.

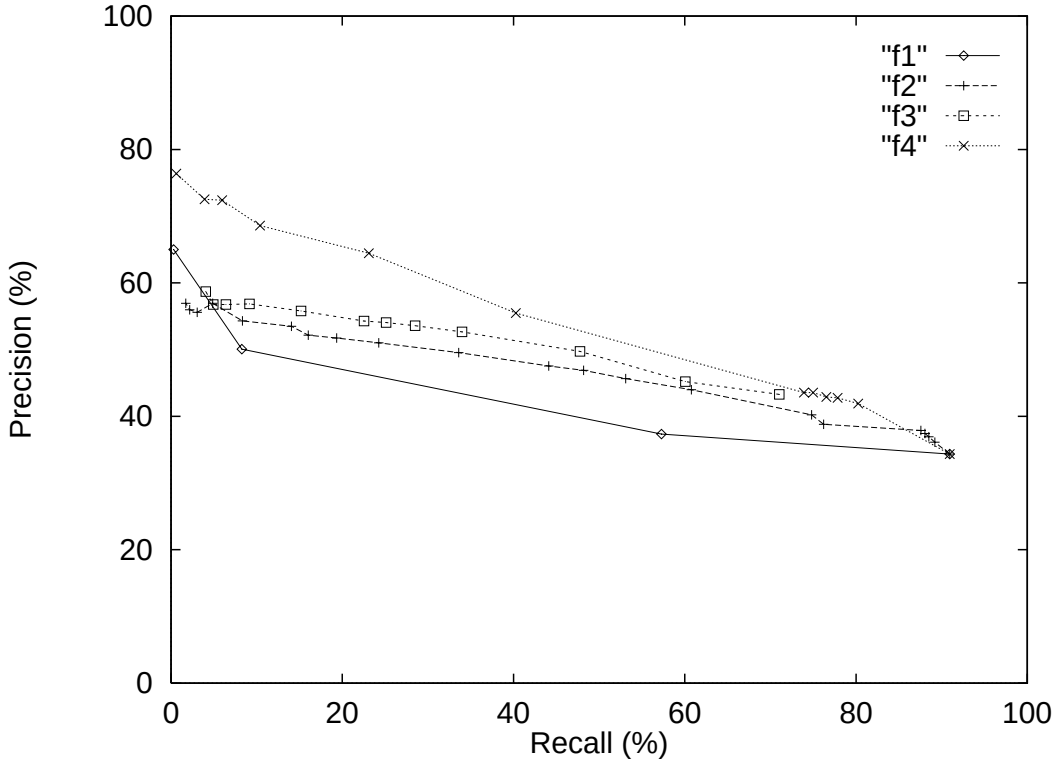

Figure 1: Recall and Precision based on Manual-Candidates (Tagged)

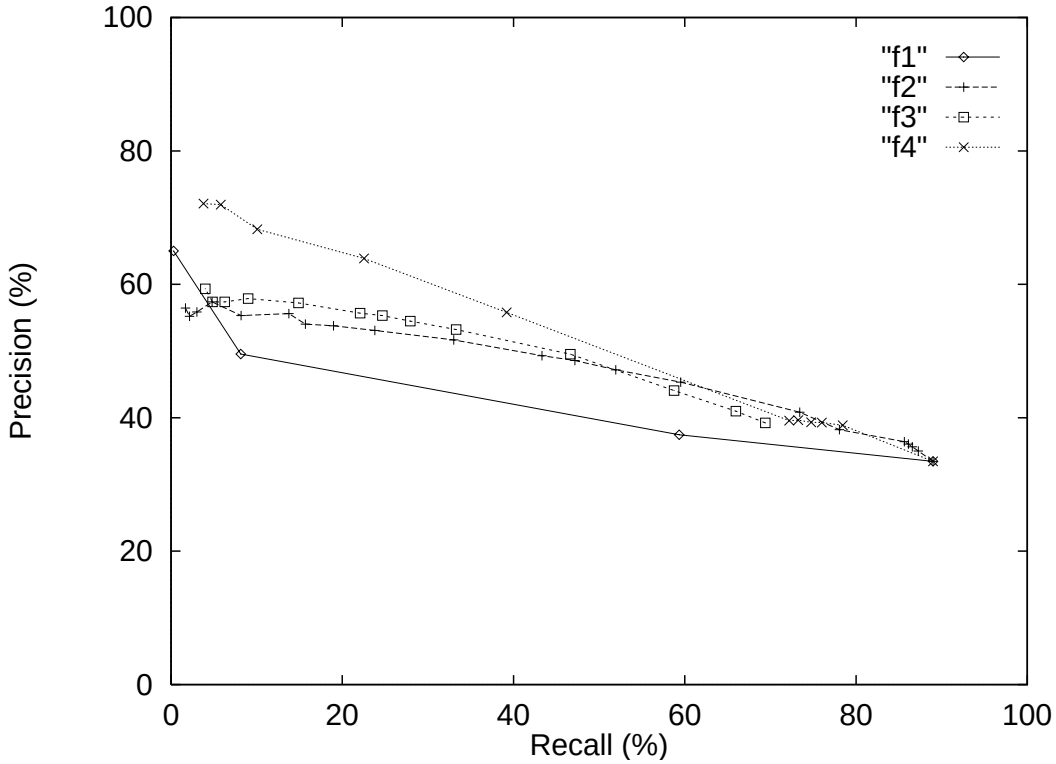

Figure 2: Recall and Precision based on Manual-Candidates (Untagged)

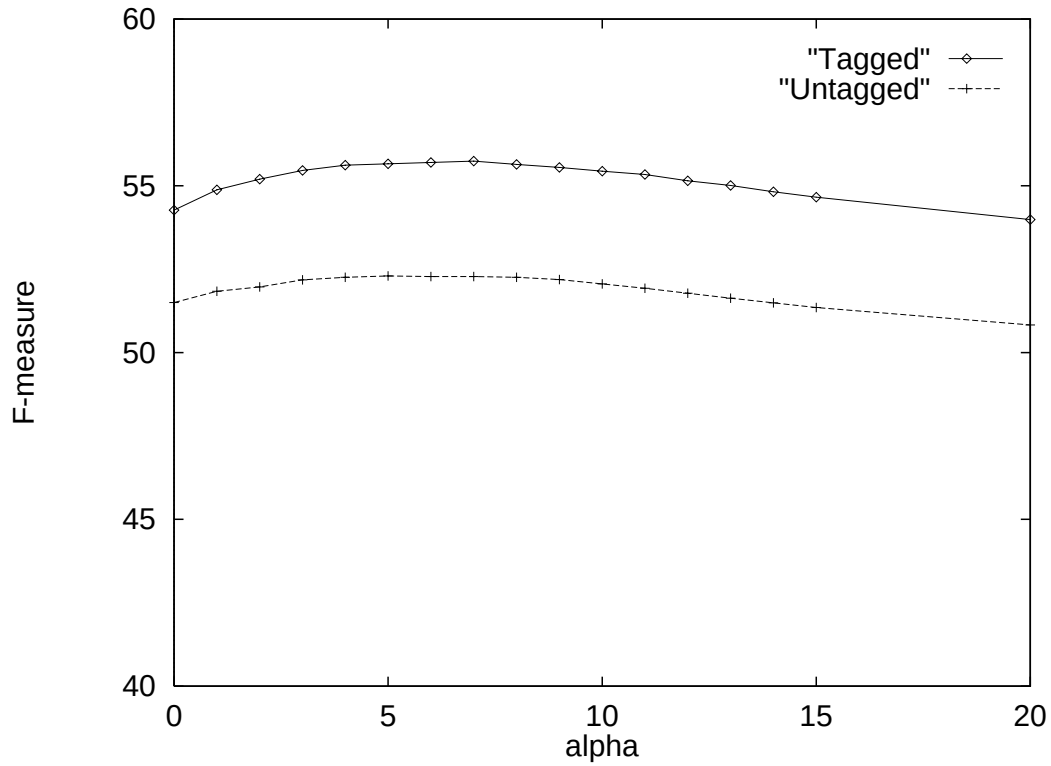

Figure 3: Relationship between α and F-measure

Table 2: Best accuracy by using Eq. (3)

	α	Recall	Precision	F-measure
Tagged	7	82.54% (7,292/8,834)	42.08% (7,292/17,328)	55.74
Untagged	5	81.04% (7,159/8,834)	38.61% (7,159/18,543)	52.30