NTCIR Experiments at Matsushita: TMREC Task

FUKUSHIGE, Yoshio
fuku@trl.mei.co.jp

NOGUCHI, Naohiko
noguchi@trl.mei.co.jp

Multimedia Systems Research Laboratory, Matsushita Electric Industrial Co., Ltd.
4-5-15, Higashi-Shinagawa, Shinagawa-ku, Tokyo 140-8632 JAPAN
+81-3-5460-2744

Extended Abstract

In this report, we describe the results of our experiments in the NTCIR TMREC task group. We participated in all the three tasks, namely the term recognition task, keyword extraction task, and the role analysis task.

Section 1 of the main body of this paper describes the outline of our experiments, section 2, our experiments for the term recognition task, section 3, the keyword extraction task, section 4, the role analysis task.

Because the main body of this paper is written in Japanese language, we describe, in this extended abstract, the summary of this paper for non-Japanese readers.

1. Term Recognition Task

In the term recognition task, participants are required to extract from the given articles the terms that are considered to be typical to the artificial intelligence domain.

1.1 Two Approaches

We took two different approaches to the task. The first is based on an idea that the domain-typical terms are the ones that can discriminate the documents in that domain from a larger document set. The second is based on another idea that the domain-typical terms are the ones that are difficult to understand and thus which are to be collected in the glossary for that domain.

1.2 First Approach

As described above, our first approach to the term recognition task is to extract the terms that can discriminate the documents of the domain from a larger document set. Therefore, we tried to find the terms using the statistical features of the in-the-domain documents against the whole document set. We use the NTCIR J-Collection corpus as the whole document set.

We did four experiments using four different tools in each of which the following scores are employed as the measure respectively, namely (a) χ^2 value, (b) score based on the probabilistic model, (c) tf・idf value, and (d) the product of (a) and (b).

The results of these experiments showed that while the terms with moderate frequencies also appear in the top part of the term list ordered with the value of (a), terms with low frequency appear mainly there with the tools (b) and (c). In case of the tool (c), most top terms are of the frequency 1 or 2, which we did not think is a good result.

The tool (a) refers to the divergence of the term occurrence from the mean (or expected) value, but does not care the number of the documents in which the term occurs. In contrast, tools (b) and (c) refer to the number of the documents in which a term occurs, but don't refer at all to the divergence.

We thought it important to keep balance between the two, so adopted the product of (a) and (b) as the measure in the formal run.

We learned through the examination of the formal run result, that:

- Some sort of normalization or tuning by parameters is essential in evaluating the term weights.
- We need some theoretical support for the threshold, into which we did not investigate so far this time.

In addition, in counting the term frequencies, we used the maximal term frequencies that can be obtained with the use of MEISTER's index module. Analysis of the difference between the result we got this time and the one that is to be obtained by using plain n-gram frequencies may be of some interest.

1.3 Second Approach

Our second approach to the term recognition task is based on an idea that what we should extract as terms typical to a domain is the terms that cannot be easily understood without the knowledge in that domain. This is the standard with which terms are extracted in making a glossary for that domain.

Terms that are to be in the glossary are of two types, i.e. (a) terms that also appear in a general dictionary with different or special meaning or usage in that domain, and (b) new terms which do not appear in such dictionaries. New (i.e. not-in-the-dictionary) terms are divided in three categories: simple words (i.e. non-compounds), compounds, and word-like phrases.

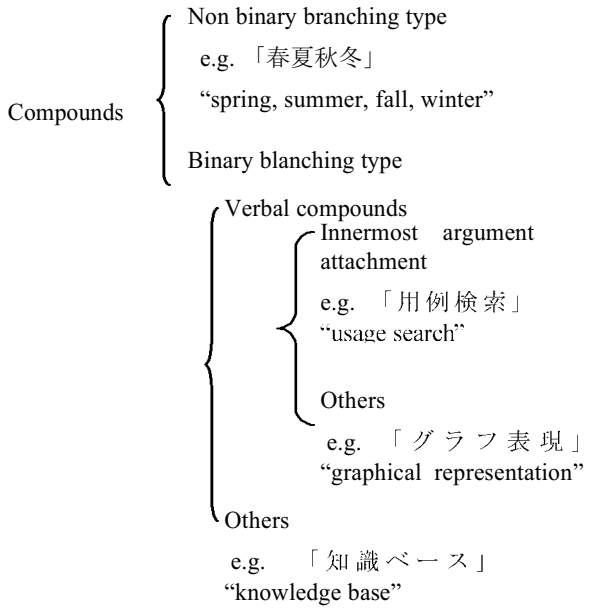
This time, we mainly concentrated in distinguishing compounds, especially compound nouns.

It is known that meanings of some compounds are just systematical syntheses of those of the morphemes that form the compounds and thus easily understood, while meanings of other compounds are not. With our idea on the domain-typical terms above, our target should be the latter.

As in the English language, in the Japanese language also, compounds are morpho-syntactically of two types: ones with conjunctive structure and ones with binary branching structure. The latter are again of two types: verbal compounds and others. A verbal compound is the one that has as its head (usually the right-hand morpheme) a noun with the verbal origin. Verbal compounds are again categorized into two types in regard to the semantic nature of the left-hand element to its head. Those whose left-hand element is semantically the innermost argument of the head are very productive and easily understood, while those of the rest type are not so freely introduced and have specialized meaning to some extent. With our intention to extract terms with a domain typical meaning, we should extract the verbal compounds with the former type of inner structure.

On compounds of binary blanching type other than verbal compounds, we did not study in detail, but adopted some simple criteia. Some nouns are qualified as synthetic suffixes, and compounds that have such nouns as their right-most element are labelled as synthetic.

We applied another set of simple criteria to the simple words.



In the analyses of terms' inner structures, we consulted to the data from the tagged corpus. We refer, for example, to the collocation data to see if the left-hand noun of a verbal compound occurs as a direct object of the right-hand verb, and if so, the compound is labeled as synthetic.

For the result of comparison between our list of candidate terms and the lists provided by the staff, see Table 2.2-1 in the main body.

Considerable portion of the divergence came from the limited size of the corpus data, while further improvement of our criteria on the correspondence between the compound's inner structure and syntactic behaviors of the element words or morphemes is needed.

As to the question whether our assumption that domain typical terms are the ones that are difficult to understand from general knowledge is appropriate or not, more divergence than our anticipation was observed. Many compounds with the synthetic structure are listed as terms typical to the domain, if one of the elemental words or morphemes is typical to the domain.

However, from the point of making of glossary, it is quite questionable to put such words into the glossary only because they contain words or elements typical to the domain, which leads to the overflow.

Combining the two methods we adopted in our experiments may give us an acceptable standard.

2. Keyword Extraction Task

For the keyword extraction task, we adopted the same first approach we have taken in the term recognition task. At most 10 keywords per one document are extracted employing (c) tfidf value as the measure.

From the evaluation by NTCIR TMREC group, we observed that the precision and the recall are relatively higher in the evaluation against author keywords than in the evaluation against AI-dictionary keywords.

It might be due to the fact that we chose our first candidates from the author keywords in the J-Collection documents which includes the target documents.

3. Role Analysis Task

The aim of the role analysis task is to extract triplets that describe the contents of an academic document. Triplets consists of three terms which represent (1) the theme (what authors treated), (2) the method (the means they took), and (3) the operation (procedure or action applied to the theme), respectively.

Our stance to this task is to extract such triplets that could be used effectively in an article search. With that stance, we did not restrict the expression of each role to the terms extracted in the Term Recognition Task.

Our procedure is as follows: (1) Extract sentences that have typical expression (mainly certain types of verbs) that express the operational acts, (2) Extract collocation data between words (mainly those between a predicate and its argument), (3) Extract terms (words) which bear the target roles (theme and method), (4) Abridge, enlarge, or paraphrase the terms, (5) Score and choose the five best triplets.

As to the result of comparison between our result and the answer given by the staff, see Table 4.4-1 in the main body.

The study of the result tells that the top four of the most frequent types of inappropriate triplets are (1) description of the behavior of the target, (2) description of the method, (3) description of the background of the research, and (4) error in the collocation analysis.

Apart from the errors, the most difficult question we faced in this task was how far in detail we should describe the terms, or how briefly the terms should be described, in other words.

Studies on evaluation of term's information as a role (or argument) of a certain statement in a document, or study on triplet's information may yield an acceptable answer to that question.

Keywords

term, keyword, role, tfidf, domain, chi-square value, compound, synthetic, verbal compound, predicate, argument, binary branching, argument structure, inner most argument, theta-criterion, term expansion, term abridgement,

1. 概要

我々は、TMREC グループの 3 課題に参加した。本稿では、２節にて、用語抽出タスクに対する取り組み、３節にて、キーワード抽出タスクに対する取り組み、4 節にて役割分析タスクに対する取り組みについて述べる。最後に 5 節にて、まとめを行う。

2. 用語抽出タスク

筆者らは、「専門用語とは何か」という問題に対して、異なる二つの考えに基づく抽出実験を行った。

第一の実験は、「専門用語とは、専門分野の文書を一般の文書から区別する際の特徴となる語である」という考えに基づくタグなしコーパスからの抽出実験である。

第二の実験は、「専門用語とは、一般の知識から、意味が理解困難な語である」という考えに基づくタグ付きコーパスからの抽出実験である。

2.1 抽出実験１：タグなしコーパスからの用語抽出

2.1.1 実験手法

タグなしコーパスからの用語抽出タスクでは、「専門用語とは、専門分野の文書を一般の文書から区別する際の特徴となる語である」という考え方に基づき、単語出現の統計情報を用いて抽出実験を行った。

タグなしコーパスでは、単語分割すらなされていないため、用語認定の以前に、単語認定という作業が必要である。単語認定という作業は、形態素解析処理（やそれが用いる辞書）に関連し、未知語認定や複合語の認定などの問題を内含する。今回は、この問題に対して、文書の単語分割（形態素解析）は直接的には行わず、一旦用語候補を抽出した後に、それらの出現頻度の統計を取得した後に用語の選択を行う、という手法を採用した。具体的な実験手法は、以下の通りである。

（１） 用語候補の抽出：与えられた文書から、用語候補となる文字列をなるべく多く抽出する。ここでは、元辞書としてＥＤＲ辞書を利用し、(a)我々の保有する固有名詞辞書の語、(b)比較文書全体の日本語キーワードフィールドに記載された語、(c)本文フィールドおよびタイトルフィールドの文章から、*MEISTER* の単語抽出機能[1]を利用して抽出した文字列を集積して、用語候補の全体とした。*MEISTER* の単語抽出機能は、字種情報を用いて単語を抽出するので、漢字やカタカナ連続からなる長単位の複合語を数多く抽出することができる。

（２） 出現頻度の取得：（１）で得られた各用語候補に対して、タグなしコーパス（人工知能分野の文書）および比較文書全体（J コレクション）での出現頻度統計を取得する。ここでは、*MEISTER* の索引化モジュールを利用し、対象文書の極大単語索引[2]を構成して、出現頻度を取得した。（従って、各用語候補の、極大単語としての頻度が計測される）

（３） 用語の選択：（２）の出現頻度を用いて、用語候補に重み付けを行い、重みの重いものから順に用語を選択する。ここでは、用語候補 t_I の人工知能分野における重み w_{ij} を、以下の 4 種類の式により実験的に計算した。ここで、人工知能分野の文書数、のべ語数をそれぞれ df_j、wf_j とし、比較文書全体での文書数、のべ語数をそれぞれ df、wf とする。また、t_i の人工知能分野での出現頻度、出現文書数それぞれ f_{ij}、df_{ij} とし、比較文書全体での出現頻度、出現文書数それぞれ f_i、df_i とする。

(a) χ^2値 [3][4]：

$$w_{ij} = \frac{(f_{ij} - m_{ij})^2}{m_{ij}}$$

ただし、$m_{ij} = f_i \times (\frac{wf_i}{wf})$

(b) 確率モデル [5][6]：

$$w_{ij} = \frac{df_{ij} \times (df - df_j - df_i + df_{ij})}{(df_j - df_{ij}) \times (df_i - df_{ij})}$$

(c) *tfidf* ：

$$w_{ij} = \frac{f_{ij}}{f_i} \times (\log(\frac{df}{df_i}) + 1)$$

(d) (a) × (b)

用語候補抽出（１）により、用語候補の総計は約 90 万語となった。増強された辞書を用いて頻度統計の取得（２）を行った。その概略は、以下の通りである。

表 2.1.1 用語抽出の概略

	文書数	のべ語数	異なり語数
人工知能分野	1870	298283	18105
比較文書全体	339477	60518844	589432

上記人工知能分野に出現する用語候補 18105 語に関して、（３）の各基準で重み付けを行い、上位 3200 語を正解とした。（語数は、異なり用語候補数の１／５程度を目安として決定し、重みの閾値といった基準は用いなかった）

2.1.2 実験結果および考察

重みの評価式の性質から、(b)(c)は低頻度語が重要視される傾向にあり、(a)は中頻度語も上位に出現する傾向がある。特に、(c)の場合は、上位語はほとんど頻度１か２の語となるため、回答としては採用しなかった。また、(a)は単語の文書集合における出現頻度の偏りのみを見ており、出現文書数については全く考慮していない。一方、(b)は単語の出現文書数のみを考慮しており、文書あるいは文書集合における出現頻度の偏りは考慮していない。我々は、これらのバランスをとることは重要であろうと考え、今回は、 (a)(b)を組み合せた式(d)を、正式回答として採用した。

事務局により判定された評価結果を以下に示す。

表 2.1.2 事務局による正解との一致数

全候補数	正解との一致率	
	対マニュアル	対索引
3200	1537	162

また、事務局による分析では、対マニュアル正解の場合、我々の結果は長単位の語を抽出する傾向があるとのことである。これは、（１）の用語候補としての文字列の抽出の際に、字種による長単位語の切り出しを数多く行ったこと、（２）の頻度統計取得の際に、極大単語の頻度を採用したことによると思われる。

今回我々の採用したアプローチには、以下のような問題点があると認識している。

- 用語候補の抽出は、本来は n-gram 頻度統計などに基づき、網羅的に行う必要がある。ただ、頻度統計だけに頼ると、低頻度のものが漏れてしまう危険性もある。
- 用語候補の頻度統計には、今回は、極大単語頻度を採用した。これは *MEISTER* の索引化モジュールを利用したことが直接的な要因であるが、これは、例えば文字列頻度（n-gram 頻度）を用いた場合とは様相が多少異なってくると思われる。これも今後の検討課題である。（我々は、極大単語頻度は擬似的な単語頻度を表現していると考えているが、それを積極的に利用して単語分割を行おうという試み [7]もある。）
- 重みの計算式においては、何らかの正規化や、パラメータの導入によるチューンアップが必要である。今回はその吟味が不十分であった。特に、(d)においては、単純な積ではなく、何らかの正規化が必要であると考えている。
- また、用語の最終的な選択基準については、今回は、異なり語の１／５程度を目安としたが、積極的な意味付けがあるわけではない。重みの計算式から自明な閾値が定まることが理想的だが、その吟味も今後の課題である。

2.2 抽出実験２

第２の実験においては、「専門用語とは、一般の知識から、意味が理解困難な語である」という考えに基づきタグ付きコーパスからの抽出を行った。

2.2.1 抽出ターゲット

専門用語抽出の目的の一つに、用語辞書(glossary)作りが考えられる。

専門用語を glossary に載せるべき語と考えるなら、対象となるのは、特殊な主題や分野に関する、難しい語、通常の意味とは異なる意味を持つ語（句）である。

glossary に登録すべき語句を分類すると、まず、(1)一般の辞書に既に登録されている語句で、当該の専門分野において、意味・用法が異なるもの、(2) 一般の辞書に登録されていない新規語句、と大別される。

さらに、新規語句は、構造的特徴から、 (2a)単純語、(2b)複合語、(2c)句、と分けられる。

このうち、今回の実験では、対象を、上記(2)の、一般辞書に登録されていない新規語句に絞り、中でも、複合名詞に対する専門用語性の判定に力を注いだ。

複合語に関しては、既に多くの研究がなされており、分類に関しても、様々な視点からの分類が行われている [8],[9]が、ここでは、複合語を構成する語の意味を機械的に組み合わせたものが、全体の意味となっているものと、そうでないものとがあることに着目する。今回の用語抽出の目的は、glossary に載せるべき、意味理解が困難な語の抽出と設定したので、そうした観点からは、複合語のうち抽出すべきなのは、後者となる。以下では、前者を構成的な複合語と呼び、後者を非構成的な複合語と呼ぶ。

構造的に見ると、複合語は、並列構造を持つものと、そうでないものとに分かれ、並列構造を持たないものは、基本的に二分木枝分かれ構造を持つ。

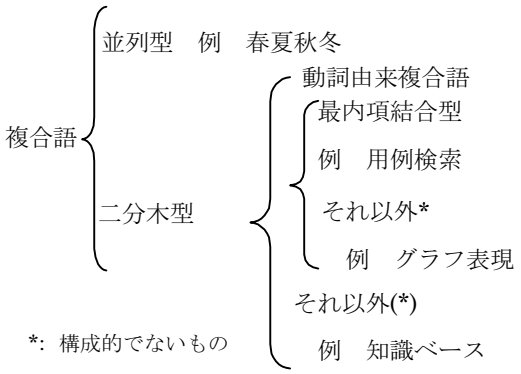


図 2.2.1 複合語の分類

このうち、並列構造を持つものに関しては、要素の意味を並列的に合成したものが全体の意味となるので、構成的であり、理解は容易である。

二分木構造を持つ複合語は、動詞を主要部として右側に持つ動詞由来複合語と、それ以外に分けられる。

このうち、動詞由来複合語は、動詞の持つ項構造における最内項を左に結合したものと、それ以外の項や副次的要素を左に結合したものとに分けられるが、前者が、非常に生産的であり、その意味も、構成語の意味から構成的に決定されるのに対して、後者は、少なからず語彙化を受けており、生産性が低く、意味も特殊化されていることが指摘されている。今回の目的からは、後者が抽出対象となる。

動詞由来複合語以外の二分木構造を持つ複合語に関しては、今回は詳しい分析・分類の対象としなかったが、基本的には、意味解釈の自由度が高く、逆にいえば、実際に使われる際の意味は特殊化されているといえるので、非構成的とした。ただし、一部の名詞は、非常に生産的に動詞の右側に結合して複合語を形成し、その意味も比較的構成的に決定されることが観察できる。今回は、そのような名詞を構成的接辞とし、それらによる複合語は、構成的とし、専門用語としては抽出しないことにした。

単純語に関しては、別の粗い基準から抽出を行った。

2.2.2 概要

上記の基本方針のもと、タグ付きコーパスから規則ベースの抽出実験を行った。ただし、外部基準辞書として(株)日本電子化辞書研究所(EDR)の日本語単語辞書 1.5 版[10]を利用し、同辞書に一般語としてのみ登録されている語は、抽出対象としなかった。

抽出は、以下のステップにより行った。

ステップ１　普通名詞及びサ変名詞の抽出

タグ付きコーパスから、普通名詞およびサ変名詞と品詞付けされた語を、候補として抽出する。ただし、他の語の部分語としてしか現れないものは、抽出しない。

ステップ２　一般語の除外

候補のうち、EDR の日本語単語辞書に一般語としてしか現れないものを除く。

EDR 日本語単語辞書には、一般語と、情報処理分野の専門語が登録されている。今回の対象専門分野が、人工知能分野であるということを考慮し、情報処理分野の専門語としてのエントリのあるものは、残す。

ステップ３　複合語と単純語への分類

タグつきコーパス中の複合語マーカーを元に、候補を複合語と単純語に分ける。

ステップ４　複合語候補の二分木分割による分類

複合語と分類された候補を、以下の基準で、二分木分割できるものと、それ以外とに分ける

- コーパス中で二つの形態素に分割されている語は、二分木分割可能とする。
- 3 つ以上の形態素への分割しかない複合語に関しては、コーパス中に現れる二つの語あるいは形態素に分割できるなら、二分木分割可能とする。
- どちらの基準も満たさないものは、二分木分割不可能とする。

ステップ５　各候補集合からの専門用語抽出

二分木分割可能な複合語候補、二分木分割不可能な複合語候補、単純語の候補のそれぞれに対して、後述の基準にしたがって専門語候補を抽出する。

各候補集合からの専門用語抽出に関しては、以下で詳しく説明する。なお、以降の例において、*印がついているものは、意図に反した結果の例である。

なお、ここで参照するために、タグ付きコーパスから、品詞パタンマッチによる単語共起データを取得する。

取得される共起データは、以下に示すような 3 つ組みである。このうち、関係ラベルは、共起関係の種類を表すラベルであり、述語－項関係およびにおいては、関係ラベルは、項に後接する助詞等を宛てた。判定、連体修飾、連用修飾においては、それらの関係をそのまま用いた。

拡張する　－　を　－　類推の理論

理論　－　の　－　類推

(主要部要素)　(関係ラベル)　(非主要部要素)

図　2.2-2 「類推の理論を拡張する」からの共起データ抽出例

ステップ６　抽出結果のマージ

各候補集合から得られた候補の和集合を、最終的な抽出結果とする。

以下に、分類された各候補集合からの抽出方法を述べる。

2.2.2.1 二分木型複合語の集合からの抽出

上記により二分木分割可能と判定された複合語の集合から、特定の品詞を構成語として持つものを除き、残ったものに、構成語の品詞組別の判定基準を適用し、構成的とされなかったものをこの分類からの候補とする。

以下に各手順について述べる。

1.　構成語の品詞による絞込み

接辞、数詞、形式名詞などを含むものを候補から除く。

除かれる例

構文構造同士, 各クラス, 二通り, *視覚心理学, *サブダイアローグ, *反復法, *3 つ組法

2.　「普通名詞＋サ変名詞」からの構成的複合語取得

まず、コーパスから抽出した共起データを調べ、複合語の構成要素のうち左側の普通名詞が、右側のサ変名詞の「を」格になっている例があれば、構成的とする。

構成的とされた例

工程設計, 用例検索, 仮説検証　*図形表現,

次に、複合語全体が「を」格をとる例が共起データとして得られていれば、非構成的とする。これは、一般に「を」格の要素は動詞の最内項であるので、それが複合語の外にあるのなら、θ 基準により、動詞に結合しているのは最内項ではありえないからである。

非構成的な例（ただし、今回は該当例なし）

特異値分解、指数変換

次に、右側のサ変名詞の「を」格となっている語 w_a が、複合語全体 w_b に関して「w_a の w_b」という形で、コーパス中に出現していれば、複合語を非構成的とする。これも、θ 基準に基づく。

非構成的とされた例(括弧内は、「w_a の」の例)

（部品の）空間配置、（式の）グラフ表現、*（文の）整合性解析、*（データの）依存関係分析

右側のサ変名詞が「を」格を取らない場合も、非構成的とする。

非構成的とされた例

正規分布交叉、単位行動

以上で非構成的とされなかったものは、構成的とする。

3. 「サ変名詞＋普通名詞」からの構成的複合語の取得

左側のサ変名詞の「を」格となっている語 w_a が、複合語全体 w_b に関して「w_a の w_b」という形で、コーパス中に出現していれば、右側を、構成的接辞とし、構成的接辞を右側に持つ複合語は、構成的とすし、そうでないものは、非構成的とする。

これは、構成的接辞は、述語に付加して項構造を継承する、形式名詞あるいは派生接辞的な性質を持つものであり、全体の意味も理解容易であるという観察に基づく。

構成的接辞とされた例

方法、過程、手法、環境、問題、モデル、機構、システム、技術、*アルゴリズム、*原理、*関数

構成的複合語とされた例

解決手法、プランニング過程、マージ手法、実現環境、 *制約問題、*遺伝アルゴリズム、*協調モデル

非構成的とされた例

航行めがね、隣接アブデューサ、交叉オペレータ、解析木、自律ロボット、導出節

4. 「サ変名詞＋サ変名詞」からの構成的複合語取得

この場合は、まず「普通名詞＋サ変名詞」とみなして構成的か調べ、次に「サ変名詞＋普通名詞」として構成的か調べ、どちらかの基準で構成的とされたものは構成的とし、そうでないものは非構成的とする。

構成的とされた例

助言生成、散逸防止、 *反復学習、*直交分解; 学習機能、設計実験、*交叉設計、*分散処理

5. 「普通名詞＋普通名詞」からの理解可能語取得

この場合は、右側の普通名詞が、３で求めた構成的接辞のリスト中にあれば、構成的とし、そうでなければ非構成的とする。

構成的とされた例

知識ベースシュミレーションシステム、知識獲得方式、*ヒューリスティック関数

非構成的とされた例

自然言語文、Ｃ言語、ＤＳ理論、*文献資料、*誤り訂正回数、*助詞選択法

6. １の残りからの構成的複合語の除外

最後に、1 で残った候補に対して、2〜5 で構成的と判定されたものを除き、残ったものを、二分木分割可能な複合語集合からの最終候補とする。

2.2.2.2 非二分木型複合語の集合からの抽出

この分類には、本来的に二分木構造を持たない複合語と、コーパスの規模の問題により、二分木構造を明らかにできなかった複合語が混在している。また、複数単語からなる英語の人名やシステム名も含まれる。

これらに対しては、以下の基準によりペナルティを与え、あらかじめ設定した基準値よりペナルティが高いものを候補から外し、残りを最終候補とする。

- 接辞、数詞、形式名詞、区切り記号などを含めば、大きいペナルティを与える。
- 固有名詞を含めば、ペナルティを減らす。
- 「普通名詞＋サ変名詞」あるいは「サ変名詞＋普通名詞」で終われば、小さいペナルティを与える。

(一番右の語と、それより左の部分からなる二分木構造を持つ語とみなしての粗い基準)

専門用語でないと判定された例

Virtural Tour, 質問生成ウィンドウ, ２実体間、*ポスト大量生産パラダイム、*rough 集合理論

専門用語と判定された例

非線型擬似ブール代数、型つき素性構造、受付応対会話、ソリッド合成処理

2.2.2.3 単純語候補集合からの抽出

単純語と判定された語については、(1) コーパス中に普通名詞以外として現れる場合、および (2) 普通名詞として複合名詞の要素となっている例がコーパス中にない場合にペナルティを与え、ペナルティがあらかじめ設定した値を越えるものを捨て、残りを候補とした。

専門用語とされなかった例

プログラミング、競合、GEMACS, *バックトラック

専門用語とされた語

ベクトル、ノード、*段落、*場所

2.2.3 実験結果および考察

残念ながら、提出した試験結果においては、EDR 辞書を使ったチェックに誤りがあり、本来一般語として扱うべき語が、候補に残っていた。また、普通名詞＋普通名詞に対する処理にもバグがあった。そのため、提出した結果は、本来目指した結果と若干異なるものとなった。

次表に、公式結果及び、上記バグ修正後の結果について、事務局による正解との比較結果を示す。

表 2.2-1　事務局による正解との一致数

	全候補数	「正解」との一致数	
		対マニュアル	対索引
公式結果	6,913	3,368	281
バグ修正後	6,444	3,243	258

構成的複合語と非構成的複合語の判別という視点から今回の結果を見ると、誤りの多くは、コーパス中に直接的な証拠が得られなかったことが原因と思われた。

ただし、「普通名詞＋サ変名詞」型の分析において、左側要素が右側要素に対する最内項以外の項であったり、付加詞であったりする例を、コーパス中に積極的に求めていなかったため、本来非構成的とされるべきものがデフォルトで構成的とされてしまった例が多く見られた。

また、形容詞を構成素として持つ複合語の構成性に関しても検討すべきであった。

さらに、二分木分割できなかった複合語、および「サ変名詞＋サ変名詞」「普通名詞＋普通名詞」のパタンを持つ複合語に関しては、項構造の同型性を調べるなど、並列構造の可能性を調べるべきであった。

なお、今回の判定においては、一般語として辞書に登録されている語については、分類の対象としなかったが、それらに対する専門用語性の判定のためには、分野における選択制限に関する振舞いの違いを調べることが必要になる。

一方、「専門用語とは、一般の知識からは理解困難な語である」という我々の前提から、与えられた正解を眺めた場合、予想以上に違いが認められた。

特に、複合語が構成的に生成される場合でも、構成語自体が特徴的な場合は、専門用語としての判定を受けているように思われる。ただし、glossary 構築という観点からは、規模の爆発を防ぐために、構成的な語はなるべく採用しないのが望ましい。

逆に、理解困難な語が、かならずしも「当該分野の」専門語とは限らない、という問題もある。

対象の語彙性や分野依存性も考慮した、何らかの判定基準が必要である。

2.3 二つの実験結果からの考察

我々は専門用語に対する二つの捉え方に基づき、抽出実験を行った。

第一の、テキスト群の分別力に基づく定義による手法では、特に低頻度語の扱いが今後の課題といえる。

第二の、理解困難性に基づく定義による手法では、分野依存性の評価が、今後の課題といえる。

今後は、両手法を組み合わせによる、これらの課題の解決に取り組みたい。

3. キーワード抽出タスク

3.1 実験手法

キーワード抽出タスクに対しては、用語抽出タスクの１番目のアプローチ、即ち、「各文書を他の文書から区別する際の特徴となる語」を抽出する、というアプローチで実験を行った。2.1.1 に述べた方法で用語候補の抽出、出現頻度の取得を行い、用語の選択においては、（c）の式（tfidf）により重み付けを行って、各文書当たり上位最大 10 語をその文書のキーワードとした。

3.2 実験結果および考察

事務局による評価の結果を以下に示す。

表 3.2-1 事務局による評価

	著者 KW	AI 辞典	人手
Recall(%)	23.3	18.9	27.6
Precision(%)	10.7	7.2	39.9

事務局による他のサイトとの比較では、著者キーワードとの一致率は高かったが、AI 辞典や人手で付与されたキーワードとの一致率は低かった。これは、次の２つの理由によるものと思われる。

- 初期の用語候補（頻度統計を取得する語）として、比較文書全体の日本語キーワードフィールドから抽出を行ったため、抄録本文にこれらが現れている限り、キーワードとして抽出されやすかった。
- 初期の用語候補としては、抄録本文から字種情報を用いて切り出した比較的長単位の複合語も多く含まれていた。また、それら用語候補に対して極大単語頻度にて頻度統計を取得したために、長単位語が選択されやすい傾向があったと思われる。これが、AI 辞典との一致率が低かった原因かも知れない。

4. 役割分析 タスク

4.1 タスクの説明

「役割分析タスク」は、人工知能分野の学術論文の表題及び要約から、当該論文の論述内容として、「どのような対象に対して」「どのような手段により」「どのような操作を行ったか」を示す用語を 3 つ組として抽出するタスクである。

4.2 アプローチ（方針）

筆者らは、検索におけるキーとして利用しやすいような、論述役割および、その組を抽出する、という方針で、上記タスクに取り組むことにした。その観点から、抽出する 3 つ組の要素は、用語抽出タスクにおいて抽出した用語に限定しなかった。

さて、要約中の文は、おおまかに(1) 研究開発の背景を述べる文、(2) 研究開発行為について述べる文、(3) 研究開発対象の動作、性質などについて述べる文、(4) 論文自体について述べる文、の４種類に分けられる。

このうち、「対象に対して変化を与える」ということを含意する狭義の「操作」を述べる文といえるのは、(1)のみであるが、今回は「～の概要を述べる」のような報告行為や、「～であることを示す」のような証明行為など(4)に分類されるものも、広く「操作」に含め、抽出した。また、操作の「手段」も、広く「操作」に含めた。

なお、形態素パタンによる共起関係抽出を行うために、タグ付きコーパスを利用した。

4.3 抽出手順

4.3.1 概要

上記方針に基づき、1. 論述役割抽出対象文の抽出、 2. 共起関係抽出、3. 論述役割抽出対象述語の抽出、4. 「対象」「手段」の取得、5. いいかえ、拡張、6. 優先度付けと絞込み、という手順で論述役割組の抽出を行った。以下で各手順について説明する。

4.3.2 論述役割抽出対象文の抽出

まず、背景説明文、および対象の性質を述べる文を除くために、記事本文から、特定の文末表現を持つ文、および特定の語を含む文のみを、論述役割抽出対象文として、抽出する。

文末表現に対する基準としては、操作を表す動詞のリストを、人手で作成した。同リストには、「示す」「論じる」など、狭義には操作を表さない語も含めた。

4.3.3 共起関係抽出

抽出された論述役割抽出対象文から、用語抽出の第 2 の実験で用いた共起関係抽出ツールを使って、品詞パタンマッチにより、共起関係を抽出する。

抽出する共起関係は、述語・項関係、「ＡはＢである」のような判定、「名詞句」＋「の」＋「名詞句」、名詞句による連体修飾、関係節による名詞の修飾、動詞句に対する連用修飾など。

4.3.4 論述役割抽出対象述語の抽出

抽出された共起関係を参照し、論述役割抽出対象文中の、主文の動詞および、主文を連用修飾している動詞を抽出し、論述役割抽出の対象述語とする。

4.3.5 「対象」「手段」の取得

論述役割抽出対象述語について、共起関係抽出結果を元に、「対象」「手段」について、当該の述語と、それぞれ｛を、と｝、｛により, から, に基づく, ため, 連用修飾｝の最も先頭にある関係ラベルを伴って共起している語（句）を取得する。

4.3.6 いいかえ、拡張

キーワード化を意識して、求めた各要素の言い換えを行う。

- 名詞句化

人手で作成した、述語言い換えリストを参照し、用言や複文を名詞句に変換する。その際、必要に応じて、助詞なども変換する。

変換例

構造を推定する　→　構造推定

拡張できること　→　拡張可能性

- 項の拡張

項の内容をなるべく限定されたものにするために、あらかじめ設定した一定の回数まで、修飾句、関係節を再帰的に項に付加し、項の形を整える。

4.3.7 優先度付けと絞込み

単純に各要素の文字列長の和の大きいものを優先する。ただし「対象」が空のものは除く。

4.4 結果の評価

事務局作成の正解に基づく評価結果は、以下の通り。

表 4.4-1 事務局による正解との比較

抽出３つ組数	評価結果		
	○	△	×
347	163(47.0%)	92(26.5%)	92(26.5%)

ただし、評価対象記事数は 90 記事。評価結果の「○」は、主要論述に対応するもの、「△」は、主要論述に部分対応または、派生論述に対応するもの、「×」は、それ以外を意味する。

×評価となった 3 つ組について、原因の分析を行ったところ、上位は、1.操作対象の動作に関する記述を抽出したもの(20%)、2. 手段に関する記述を抽出したもの(16%)、3. 研究の背景に関する記述を抽出したもの(14%)、4. 係り受け間違い(10%)、であった。

4.5 課題と考察

今回のシステムにおいては、係り受け解析に関しては、形態素パタンマッチによる簡単な解析しか行わなかったため、共起関係の抽出精度は、必ずしも満足の行くものではなく、結果の精度を下げる原因となっていた。今後の課題として、係り受け解析の精度向上が挙げられる。

また、論述役割組の要素として項や述語をどこまで限定的に記述するかについて、利用目的に応じた、各要素の語彙性や情報量を考慮した基準の設定方法をも検討課題である([11])。

また、論述役割組に関する重み付け（重要度評価）も、今後の課題である。抽出した３つ組の持つ「情報量」を、抽出元のテキストや、テキスト集合から定義（計算）することも、興味深い課題であると考える。

5. まとめ

筆者らは、今回、用語抽出タスク、キーワード抽出タスク、論述役割分析タスクに参加した。

専門用語抽出タスクにおいては、テキスト群の分別力という視点と、意味の特殊性（非構成的性）という二つの視点からアプローチを行った。

論述役割分析タスクの将来課題に関する議論において、論述要素の語彙性、情報量や、3 つ組に対する情報量の与え方が重要課題であると論じたが、これらは、専門用語抽出タスクでの課題と、密接に関連する。今後は、それらの課題に取り組みたい。

参考文献

[1] Ito, H., Sato, M. and Noguchi, N. NTCIR Experiments at Matsushita: Ad-hoc and CLIR Task. In this volume.

[2] Noguchi, N., Kanno, Y., Kurachi, K., and Inaba, M. New Indices for Japanese Text: A New Word-based Index of Non-segmented Text for Fast Full-text-search Systems. Transactions of IPS Japan, Vol.39, No.4 (1998), 1098-1107.

[3] 長尾,水谷,池田. 日本語文献における重要後の自動抽出. 情報処理学会論文誌, Vol.17, No.2 (1976).

[4] 渡辺,村田,竹内,長尾. χ^2法を用いた重要漢字の自動抽出と文書の自動分類. 情報処理学会研究報告, FI39-4 (1995).

[5] 海野. 出現頻度情報に基づく単語重み付けの原理. Library and Information Science, No.26 (1988), 67-88.

[6] 谷口. 情報検索モデル-その数量的アプローチ-. 「図書館情報学における数学的方法」, 日外アソシエーツ (1994), 80-122.

[7] 永田昌明. 単語頻度の再推定による自己組織化単語分割. 情報処理学会研究報告, NL121-2 (1997).

[8] 景山太郎. 文法と語形成, ひつじ書房(1993).

[9] 大石強. 形態論, 現代の英語学シリーズ, 開拓社(1988).

[10] ＥＤＲ日本語単語辞書 http://www.iijnet.or.jp/edr/J_index.html

[11] 岡, 小山, 上田. 句表現要約の句合成手法. NL 129-15(1999).