ワールドワイドウェブを利用した用語説明の自動生成

桜井 裕[†] 佐藤理史[††]

本論文では，与えられた用語に対して，その用語を説明する文章（用語説明）をワールドワイドウェブから収集し，それらを編集してユーザに提示するシステムを提案する．本システムは，(1) 用語説明の収集，(2) 編集，の 2 つのモジュールから構成される．「用語説明の収集」では，まず，サーチエンジンなどを用いて，入力された用語の説明が記述されている可能性が高いウェブページを収集する．次に，収集したウェブページから，用語の説明が記述されている段落を抽出する．最後に，抽出した段落内を解析し，その用語を定義する文（用語定義文）が存在するかどうかを判定し，存在した段落のみを用語説明として出力する．この判定においては，13 種類の用語定義文それぞれに対して設定した文型パターンを用いる．「編集」では，収集した用語説明を語義ごとにグループ化し，それぞれのグループに対して，最適な用語説明と上位語を決定する．最後に，これらをまとめて，結果を語義ごとに出力する．本システムにおいて，用語定義文の判定精度は 87%，グループ化の精度は 81%であり，ほぼ実用レベルに達していると考えることができる．

Automatic Generation of Term Explanation from the World Wide Web

Yu Sakurai[†] and Satoshi Sato[††]

This paper proposes a term explainer that offers us a *virtual dictionary*, which uses the World Wide Web as information source. The system consists of two modules: explanation collector and explanation editor. For a given term, the first module collects related web pages by using search engines, and extracts paragraphs that contain the term explanations. Sentence patterns of thirteen kinds of definition sentences enable automatic detection of definition sentences and automatic extraction of term explanations. The second module classifies the extracted explanations into groups according to the meaning, and determines the best explanation and the best broader term for every group. Finally, the system generates the result in HTML. In an experiment, the system achieved 87% accuracy in detection of definition sentences and 81% accuracy in classification of explanations into groups.

1. はじめに

日々増え続ける情報の中で，我々はしばしば意味の分からない用語に出会う．こうした未知の用語の意味を調べる代表的な方法は，国語辞典や百科事典を引いて調べることであるが，近年，ワールドワイドウェブ（以下，ウェブと略記）の普及にともない，サーチエンジンなどを用いて，用語の意味を「ウェブで調べる」という新しい方法が浮上してきた．

辞書や百科事典は，利用者が用語や概念を調べることを想定して注意深く編集されている．そのため，調べたい用語が収録されているのならば，辞書や辞典で調べるのが最も簡単でかつ効率的な方法である．しかし，新しい用語が生まれ，それが辞書や辞典に収録されるまでには，かなりの年月を要するのが普通である．また，ある限られた分野や領域でのみ使用される専門性の高い用語（専門用語）は，一般的な辞書や辞典に収録されていないことが多い．

これに対して，ウェブは，頻繁に更新されるという特徴を持っているため，最新の用語に関しても多くの説明や解説が存在する．また，ウェブには，非常に専門的な情報も掲載されるため，専門用語に関しても，それに対する解説ページが存在することが期待できる．

このように，「ウェブで調べる」方法は，潜在的には，旧来の「辞書で調べる」方法の欠点を埋める可能性を秘めている．しかし，現在，我々が利用可能なサーチエンジンは，用語の意味を調べるという用途においては，それほど便利なツールとはなっていない．

† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology
†† 京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

通常，調べたい用語を検索質問（クエリ）としてサーチエンジンに入力することになるが，多くの場合，大量の URL が検索結果として得られるため，その中から用語の説明が記述されているページを手作業で見つけることが必要になる．このため，「辞書で調べる」方法と同じような簡便さで，用語の意味を調べることができない．

本研究では，用語を入力すると，その説明文を出力するようなシステムを実現することにより，「用語の意味をウェブで調べる」方法を自動化する．このシステムは，いわば，ウェブを仮想辞書化するシステムであり，ユーザは，このシステムを用いることにより，「辞書で調べる」方法と同じような簡便さで，用語の意味を「ウェブで調べる」ことができる．

2. システムの概要

本システムは，用語を入力として受け付け，その用語を説明する文章を出力する．**図 1** に本システムの入力インタフェースを，**図 2** に入力「ABS」に対する出力を示す．

本システムでは，出力は語義ごとに表示される．この例では，入力「ABS」に対して，「安全システム」の下位語としての「ABS」(「アンチロック・ブレーキ・システム」）と，「証券」の下位語としての「ABS」(「資産担保証券」）の 2 つの語義が表示されている．それぞれの語義に対して，(1) 上位語，(2) 代表的な用語説明とその出典，および，(3) 他の用語説明が掲載されているページへのリンク，が表示される．

本システムの構成を**図 3** に示す．システムは，次の 2 つのモジュールから構成されている．

(1) 用語説明の収集
入力された用語を説明する文章（用語説明）を，ウェブから収集する．

(2) 編集
得られた多数の用語説明を整理・編集して，最終的な出力を作成する．

これらのモジュールについては，それぞれ，4 章および 5 章で説明する．

3. 用語の説明とは何か

3.1 用語説明と用語定義文

システムの詳細を述べる前に，まず，本システムが出力すべきものは何かということを明らかにしておく必要がある．

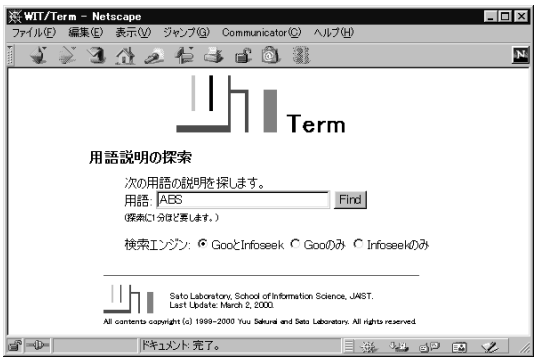

図 1 システムの入力インタフェース
Fig. 1 Input interface of the system.

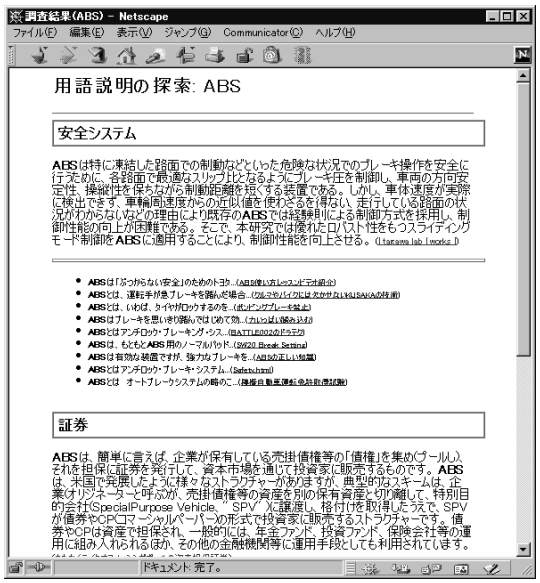

図 2 システムの出力例
Fig. 2 An example of system output.

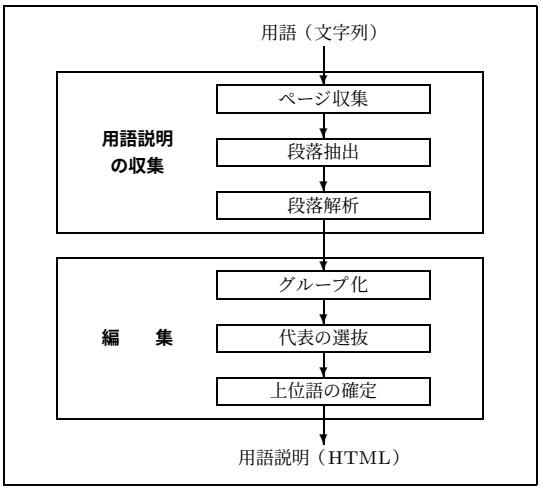

図 3 システム構成
Fig. 3 System configuration.

表 1 用語定義文の分類
Table 1 The classification of term definitions.

	名称	長尾の分類	説明	例
1.	内包的定義		上位語を用いて定義する	
1.a	直接的内包	内包的定義	被定義語がトピック	**JAVA**とは アメリカのサンマイクロシステムズ社が開発したプログラミング言語 です.
1.b	間接的内包	—	被定義語が連体修飾句を構成	「合成洗剤」は 1933 年にドイツで **ABS**という 界面活性剤が開発されたのが最初です.
2.	略記	同義語	略称を示す	**MPEG** とは，Movie Picture Experts Groupの略 である.
3.	特徴説明	特徴による定義		
3.a	性質	性質	性質を定義する	**Java**の特徴は，プラットホームに依存しないことである.
3.b	機能	機能	機能を定義する	**PowerPoint** を使って短時間でプレゼンテーションの資料を作成 することができます.
3.c	目的	目的	目的を定義する	**人工知能**は，知識のメカニズムを解明し，それを計算機上に実現すること を目的とする.
3.d	属性	属性	属性を定義する	**MPEG** は，動画のファイル形式 を定義して います.
4.	例示的定義	外延的定義	具体例を示す	白内障の中で 最も一般的な のが**老人性白内障**です.
5.	他概念との関係	—	他の概念を用いて定義する	**Java** は C++とよく似て いる.
6.	構成的定義	構成的定義	構成要素を示す	**HTML** は，ホームページを作るタグセット で構成されている.
7.	発生的定義	発生的定義		
7.a	歴史的		発生した時間を示す	1995 年 5 月にSunWorld '95 で公式に **Java** が 発表されました.
7.b	現象		発生の経緯を示す	高齢になる につれて 骨から次第にカルシウムが抜けて骨量が減少する.
7.c	原因		発生の原因を示す	**老人性白内障**は老化 が原因で起こり ます.

辞書や辞典では，それぞれの語や項目に対して，それらを説明する文章が書かれている．このような文章のことを，**用語説明**と呼ぶことにしよう．本システムが目指すことは，ウェブを仮想辞書・辞典化することであるから，本システムが出力すべきものは，この用語説明にほかならない．

では，一体，辞書や辞典の用語説明にはどのようなことが記述されるのであろうか．そこで記述される最も基本的なものは，用語（が表す概念）の定義である．本研究では，用語の定義を**表 1** に示す 7 種類（詳細分類では 13 種類）に分類する．この分類は，間接的内包的定義を除いて，長尾による分類[1),2)] にほぼ準拠している．間接的内包的定義とは，文全体がその用語の定義となっているのではなく，「ABS という界面活性剤」のような句から，用語（ABS）の上位語（界面活性剤）が読み取れるものを指す．

なお，以下では，この分類に含まれる文のことを**用語定義文**と呼び，その分類を**定義種別**と呼ぶ．

3.2 用語定義文の判定

「ウェブで調べる」方法を自動化するためには，ある文が用語定義文であるかどうかを機械的に判定する方法が必要となる．本システムでは，先に示した 13 種類の定義種別に対して，どのような文がそれに属するかを文型パターンとして記述し，そのパターンと文とを照合することによって用語定義文の自動判定を実現する．

現在，システムでは，総計で 43 個のパターンを用いている．その一部を**図 4** に示す．これらのパターンは，ウェブ上に存在する用語定義文を実際に調査し，それらを整理することによって作成した．

パターンには，次の 2 種類がある．

- 正のパターン：文がそのパターンにマッチした場合，その文をその定義種別の定義文と判定するパターン．次のような形式で記述する．
  定義種別::パターン
- 負のパターン：文がそのパターンにマッチした場合，その文をその定義種別の定義文ではないと判定するパターン．次のような形式で記述する．
  !定義種別::パターン

図 4 では，3 番目のパターンのみが負のパターンで，それ以外は正のパターンである．

パターンは形態素（<形態素基本形:品詞>）の並びとして記述する．被定義語（対象となっている用語）は `<X:*>` と記述し，任意の形態素は `<*:*>` と記述する．また，?（あってもなくてもよい）や *（任意個の並び），+（1 個以上の並び）などの正規表現を使用することもできる．

たとえば，図 4 の最初のパターンは，直接的内包的

```
直接的内包::<X:*><*:(名詞|未定義語)>*<と:助詞>?<は:助詞><*:*><*:(名詞|未定義語):E>+<だ。:*>
間接的内包::<X:*><*:(名詞|未定義語)>*<と:助詞><(い|言)う:*>+<*:(名詞|未定義語):E>+<*:助詞>
!直接的内包::<X:*><*:(名詞|未定義語)>*<と:助詞>?<は:助詞><*:*><は:助詞>
略記::<の(略|略称|頭文字|略語|訳語):*>
性質::<が(利点|欠点|メリット|デメリット|特徴|特長):*>
機能::<*:動詞><ことが:*>?<(できる|出来る|可能):*>
目的::<することを(目的|目指す):*>
属性::<を(定義する|定める|規定する):*>
例示::<(最も|一番)(多い|一般的だ):*>
関係::<と:助詞><*:形容詞>?<(似る|類似する):*>
構成::<で構成する:*>
歴史::<[\d0-9]+年:*><*:*><(発足|発表|開始|開発|スタート):*>
現象::<(につれる|にしたがう|とともに):*><*:*><する:*>
原因::<(原因|要因):*><(で|のために):*><*:*><(起こる|する|生じる):*>
```

図 4 判定に用いるパターンの例
Fig. 4 Examples of sentence patterns for term definitions.

定義に対する文パターンを定義しており，このパターンは，おおよそ，「X(と) は ... Y だ。」という文とマッチする．なお，直接的内包的定義，および，間接的内包的定義の 2 つの定義種別のパターンに限り，上位語として抽出する部分を「:E」という記号により指定することができる．先の例では，Y（名詞，または未定義語の列）を上位語として抽出する．

4. 用語説明の収集

システムを構成する 2 つのモジュールのうちの 1 つは，用語説明の収集である．本モジュールは，入力された用語を説明する文章（用語説明）を，ウェブから自動収集する．本モジュールは，ページ収集，段落抽出，段落解析の 3 つのステップから構成されている．

4.1 ページ収集

ある用語がシステムに入力されたならば，システムは，まず，次のアルゴリズムに従って，その用語説明が掲載されている可能性があるページ（候補ページ）を収集する．

(1) サーチエンジンを引く
2 つのサーチエンジン☆に対して，それぞれ，次の 4 種類の検索質問（クエリ）を入力する．
「X」，「X とは」，「X という」，「X は」
ここで，X は入力された用語を表す．それぞれのクエリに対して，最大 50URL を収集する．すなわち，最大 400URL 収集する．

(2) ページの取得
収集した URL のそれぞれに対して，ページを取得する．

(3) アンカの抽出
取得したページ中に存在するアンカを抽出する．

(4) リンク先ページの取得
抽出したアンカのアンカテキスト（<A>と</A>で囲まれた文字列）に，入力された用語が含まれていた場合，そのアンカのリンク先ページを取得する．

(5) ステップ (2) と (4) で取得したすべてのページを候補ページとして出力する．

ステップ (1) で，入力された用語以外に，その用語に助詞などを付加したものをクエリとして用いるのは，そのような表現が用語定義文によく用いられるためである[4]．また，ステップ (4) で，アンカテキストに用語が含まれているアンカのリンク先ページを収集するのは，そのようなページにその用語の説明が記述されていることが多いためである．

4.2 段落抽出

段落抽出では，収集したそれぞれのページを調べ，用語説明が書かれた段落を抽出する．実際に抽出するのは，(1) 内包的定義文を含む 1 段落，(2) 用語が見出しとなっている場合の見出しに続く 1 段落，の 2 種類の段落である．いずれの場合も，まず，前処理を行ってページを整形した後，抽出する段落の開始位置と終了位置を決定する．

4.2.1 前処理

まず，以降の処理を簡単にするために，ページの整形を行う．具体的には，以下の規則に従って，句読点の統一，HTML タグの削除，および，改行コードの挿入と削除を行う．なお，以降の処理では，HTML の開始タグ，終了タグは区別せず，同じタグとして扱う．

(1) 「.」「,」は「。」「、」に変換し，句読点を統一する．

☆ Goo（www.goo.ne.jp）と Infoseek Japan（www.infoseek.co.jp）を使用．

(2) 文中に出現する修飾タグ☆1を削除する．
(3) それ以外のタグの前後に改行を挿入し，独立行とする．
(4) 終了記号（。?! ）の後に改行を挿入する．
(5) タグの前後の改行，終了記号の直後の改行，改行の前後の改行を除く，すべての改行を削除する．

ステップ (2) で，文中に存在しうるタグをあらかじめ削除した後，ステップ (3) で，タグを独立行とする．ステップ (4) とステップ (5) により，ページ製作者が設定した改行を無視し，1 文ごとに改行する．

なお，以降の処理は，ここで標準化された行単位で行う．

4.2.2 内包的定義文を含む段落の抽出

用語説明は，内包的定義文で始まることが多い．そこで，内包的定義文を見つけて，その文からその段落の末尾までを抽出する．具体的には，以下の手続きで行う．

(1) 内包的定義文の文型パターンにマッチする行を見つける☆2．
(2) 次の行が空行またはタグであれば，見つかった行のみを抽出する．
(3) それ以外の場合は，その行を段落開始行とする．
(4) 段落開始行から逆方向に行を調べ，最初に見つかった
 以外のタグを段落開始タグとして記憶する．
(5) 段落開始行から順方向に行を調べ，以下のいずれかの条件を満たす行を段落終了行とする．
- 段落開始タグと同じタグが存在する行
-
 タグが存在する行☆3
- そこまでの全文の長さが 200 バイト☆4を超える行

(6) 段落開始行から段落終了行までを抽出する．

4.2.3 見出しで始まる段落の抽出

ウェブ上には，用語集のような形で用語に対する説明を記述しているページも多い．このようなページにおいては，見出しとして用語が示され，その後にその用語の説明が記述されるのが普通である．そこで，以下の手続きを用いて，見出しとそれに続く段落を抽出する．

(1) 用語が含まれている行を見つける．
(2) その行の文字数が，用語+5 バイト以内の場合，見出しと判定する．これにより，「X」，「X とは」，「X は?」などが見出しと判定される．
(3) 見出しの直前にある
 以外のタグを見出しタグとして記憶する．
(4) 見出しの行以降の最初のテキスト行（タグと改行以外の行）を，段落開始行とする．その行の直前のタグを段落開始タグとする．
(5) 段落開始行から順方向に行を調べ，以下のいずれかの条件を満たす行を見つける．これを段落終了行とする．
- 段落開始タグと同じタグが存在する行
- 見出しタグと同じタグが存在する行
-
 タグが存在する行
- そこまでの全文の長さが 200 バイトを超える行

(6) 見出しと段落（段落開始行から段落終了行まで）を対にして抽出する．

4.3 段落解析

段落解析では，抽出された段落内に含まれるそれぞれの文に対して，まず，形態素解析を行い，その結果と 3.2 節で述べた用語定義文の文型パターンとの照合を行って，それぞれの文の定義種別を決定する．内包的定義文と判定された場合は，同時に，その文から上位語を抽出する．なお，見出しで始まる段落の場合は，最初の文のみ，「X（と）は」の省略を許す形で照合する．用語段落解析の例を**図 5** に示す．

現在，上位語として抽出しているのは，「名詞，未定義語の列」，「動詞+こと」，「サ変名詞+すること」のいずれかの構造を持つものである．このうち最初のものは条件が緩いため，図 4 の最初のパターンによって，たとえば「X は ... の略だ。」のような文も内包

#	行	定義種別（上位語）
1	Java はサン・マイクロシステムズ社（Sun）が開発した，プログラム言語です．	直接的内包（プログラム言語）
2	Java は Web ページを作れるので，急速に広まりました．	
3	アプレットでは，動きのあるページを作ることができます．	機能
4	また，Web ページを見ている人がブラウザから要求したことに反応する双方向性が利点です．	性質

図 5 段落解析の実行例
Fig. 5 Examples of paragraph analysis.

☆1 文字を強調したり，他の情報を埋め込むために用いられるタグ．<A>，<B>，<I>，<S>，<U>，<TT>，<CITE>，<FONT>，<SMALL>，<BIG>，<SUB>，<SUP>，<STRONG>，<STRIKE> の 14 種類のタグを指す．
☆2 ここでは，3.2 節で設定した文型パターンをそのまま用いるのではなく，簡略化した文字列パターンを用いる．すなわち，この時点では行を形態素解析することは行わない．
☆3 文と文の間に
 タグが必ず存在する場合は，連続する
 タグを段落の区切りと判定する．
☆4 区切りのない大きすぎる段落を切るための最終手段である．200 バイトという基準は，実際のウェブページを調査して決定した．

的定義文と判定してしまう．これを防ぐためには，文型パターンをより精密に書く方法もあるが，ここでは，上位語としてふさわしくない語のリスト☆を用意し，上位語として抽出された語がそのリストに含まれる場合には，その文を内包的定義文とは見なさないという方法を採用した．

段落解析の最終出力は次のように決定する．

- 段落内に用語定義文が 1 文も見つからなかった場合は，その段落を破棄する．
- 間接的内包的定義文が見つかった場合は，段落ではなく，間接的内包的定義文のみを出力する☆☆．
- それ以外の場合は，段落を出力する．

5. 編　　集

用語説明の収集によって，一般に，複数の用語説明（段落）が得られる．本システムのもう 1 つのモジュールである編集では，まず，同じ内容を表す段落をグループ化し，次に，こうして形成されたそれぞれのグループに対して，代表となる用語説明と上位語を決定する．

5.1 グループ化

入力された用語は，1 つの意味（語義）しか持たない場合もあるが，複数の意味を持つ場合もある．後者の場合，同じ意味を表している説明をグループ化して出力することが強く望まれる．このグループ化を実現するためには，2 つの用語説明が同じ意味内容を表している（同義）かどうかを判定する方法が必要となる．

本システムは，次に示す 3 つの場合に，2 つの用語説明が同義であると判定する．

(1) 上位語が一致する．
それぞれの用語説明から抽出された上位語が一致する場合．上位語の一致は，2 つの単語の後方一致☆☆☆により判定する．

(2) 内包的定義文に内容語の重複がみられる．
それぞれの用語説明に含まれる内包的定義文から抽出した内容語（文字長が 2 以上の名詞，未定義語，動詞，形容詞）リストに，重複が 3 語以上ある場合．

(3) 用語定義文に内容語の重複がみられる．
それぞれの用語説明に含まれる用語定義文から抽出した内容語リストに，重複が全体の 25%以上ある場合．

5.2 代表の選抜と上位語の確定

同一内容を表す用語説明が多数見つかった場合，最終的な出力において，それらをすべて列挙するよりも，最も良さそうな用語説明を代表として表示する方が望ましい．そこで本システムでは，それぞれのグループに対して，次の基準を上から順に適用して用語説明を絞り込み，1 つに絞り込めた時点で，それを代表とする．

(1) 内包的定義文を含むものを選ぶ．
(2) 「X は ... Y である」というタイプの内包的定義文を含むものを選ぶ．
(3) 用語定義文を多く持つものを選ぶ．
(4) 定義の種類を多く持つものを選ぶ．

次に，上位語を確定する．ここでは，代表となった用語説明に基づいて上位語を決定するのではなく，グループ全体を考慮して，次の方法で上位語を決定する．

(1) 核語の決定
　(a) グループ内の用語説明から抽出された上位語を収集する．収集した上位語のリストを**リスト A** とする．
　(b) 収集した上位語がグループ内の用語説明中に何回出現するかを数える．
　(c) 出現数の最も多いものを，そのグループの**核語**とする．
(2) 核語の拡張による詳細化
　(a) リスト A から，核語を文字列として含むものを取り出す．
　(b) それらの中で，用語説明中の出現数の最も多いものを新たな核語とする．
　(c) ステップ a と b を，新たな核語が見つからなくなるまで繰り返す．
　(d) 最後に見つかった核語をグループ全体の上位語として出力する．

たとえば，「Java」の場合，上位語として，「言語」，「プログラミング言語」，「オブジェクト指向プログラミング言語」などが用語説明から抽出される．上記の方法は，まず，上位語の核となる語（この場合は，「言語」）を確定し，それを順次拡張（詳細化）していくことによって，より適切な上位語（この場合は，「オブジェクト指向プログラミング言語」）を決定する．

6. 実　行　例

本システムに「ABS」を入力した場合のシステムの実行過程を以下に示す．

☆ 「もの」「こと」「ため」「わけ」など．
☆☆ 間接的内包的定義文は，「X という Y」のような，用語 X の上位語 Y を間接的に示す部分を持つ文である．この文に続く文において，X に関することが記述されることはほとんどない．
☆☆☆ 2 つの文字列が等しい場合，あるいは，一方が他方の末尾部分文字列となっている場合に，2 つの単語は後方一致すると判定する．

表 2 URL の収集状況
Table 2 Collected URLs.

クエリ	Goo	Infoseek	計（異なり）
ABS	50	50	99
ABS は	50	50	98
ABS とは	26	50	75
ABS という	23	50	73
合計（異なり）	145	53	194

6.1 用語説明の収集

(1) ページ収集：総計で 220 ページを取得した．内訳は，サーチエンジンにより得られたものが 194URL（このうち，183 ページを実際に取得），これらのページのリンクから得られたものが 42URL（このうち，37 ページを実際に取得）である．サーチエンジンによる URL の収集状況を**表 2** に示す．

(2) 段落抽出：24 段落を抽出した．これらはすべて，「内包的定義文に続く 1 段落」であった．

(3) 段落解析：用語定義文を含んでいる段落は，15 段落あり，15 件の用語説明を出力した．このうちの 1 つは，間接的内包的定義文のみから構成される用語説明であった．

6.2 編　　集

(1) グループ化：15 件の用語説明を 4 つのグループにまとめた．

(2) 代表の選抜と上位語の確定：それぞれのグループに対して，代表となる用語説明と上位語を決定した（**表 3**）．

7. 実験と検討

7.1 実　　験

本システムの動作を確認するために，(1) 用語定義文の判定精度，(2) グループ化の精度，(3) 代表の選抜精度，(4) 上位語の確定精度，の 4 つの精度を調べる実験を行った．

実験には**表 4** に示した 95 語の用語を用いた．これらの用語は，次のように選んだ．まず，ネットニュースの記事の中から，普通の辞書には載っていないと思われる用語を 139 語選択した．この 139 語をシステムに入力したところ，120 語（86%）に対して用語説明が出力された．この 120 語に含まれていた多義語 20 語すべてと，残り 100 語からランダムに 60 語を選び，合わせて 80 語の用語集合を作成した．これとは別に，「現代用語の基礎知識 2001」（自由国民社）から，50 ページごとにページ最初の語（たとえば，「中東の石油」のような用語としてふさわしくない語の場合は，次の語）を選択して 24 語を得，このうちシステムで用語説明が得られた 15 語を，先の 80 語に追加した．表 4 で末尾に '*' が付加されているものが，追加した 15 語である．

表 3 「ABS」に対する出力
Table 3 Output for the input "ABS".

上位語：安全システム
代表の用語説明：
ABS は特に凍結した路面での制動などといった危険な状況でのブレーキ操作を安全に行うために，各路面で最適なスリップ比となるようにブレーキ圧を制御し，車両の方向安定性，操縦性を保ちながら制動距離を短くする装置である．しかし，車体速度が実際に検出できず，車輪周速度からの近似値を使わざるをえない，走行している路面の状況が分からないなどの理由により既存の **ABS** では経験則による制御方式を採用し，制御性能の向上が困難である．そこで，本研究では優れたロバスト性を持つスライディングモード制御を **ABS** に適用することにより，制御性能を向上させる．
用語説明数：10

上位語：証券
代表の用語説明：
ABS は，簡単にいえば，企業が保有している売掛債権などの「債権」を集め（プールし），それを担保に証券を発行して，資本市場を通じて投資家に販売するものです．**ABS** は，米国で発展したように様々なストラクチャーがありますが，典型的なスキームは，企業（オリジネーターと呼ぶ）が，売掛債権などの資産を別の保有資産と切り離して，特別目的会社（SpecialPurposeVehicle，"SPV"）に譲渡し，格付けを取得したうえで，SPV が債券や CP（コマーシャルペーパー）の形式で投資家に販売するストラクチャーです．債券や CP は資産で担保され，一般的には，年金ファンド，投資ファンド，保険会社などの運用に組み入れられるほか，その他の金融機関などに運用手段としても利用されています．
用語説明数：3

上位語：熱可塑性樹脂
代表の用語説明：
ABS はアクリロニトリル（A），ブタジエン（B），スチレン（S）の 3 成分の共重合物で優れた耐衝撃性，剛性，耐薬品性，光沢，成形性などを有した熱可塑性樹脂です．
用語説明数：1

上位語：界面活性剤
代表の用語説明：
「合成洗剤」は 1933 年にドイツで **ABS** という界面活性剤が開発されたのが最初です．
用語説明数：1

7.1.1 用語定義文の判定精度

用語説明の収集では，これら 95 語に対して総計で 1,430 件の用語説明が収集された．これらの用語説明において，用語定義文と判定された文は 2,168 文あった．この 2,168 文を，各定義種別ごとに，その種別の定義文としてふさわしい文であるかどうかを手作業で調べた．なお，調査の対象となる文が 100 文以上ある定義種別については，ランダムに選んだ 100 文のみを調べた．

調査結果を**表 5** に示す．実際に調査した 774 文中

表 4 実験に用いた用語（95 語）
Table 4 Terms that were used in experiments.

用語群 A（74 語） ･･･ 語義が 1 つ
ADSL, APEC, AppleTalk, cdmaOne, CERT/CC, D3 端子, DirectX, DOS, DVD, Flash, IMF*, IPv6, ISDN, i モード, Java, JR セントラルタワーズ*, MP3, MPEG, MRI, NMR, NPO, O-157, OpenGL, OSI, PNG, powerpoint, ROM, sendmail, TDMA, VCX, WindowsCE, WML, WTO, XML, Z バッファ, エストロゲン, クローン病, サイレージ, シオニズム*, ストックオプション, ダイオキシン, ダウン症, ツイストパーマ*, データマイニング, てんかん, パーキンソン病, ハイブリッド車, バリアフリー, フラッシュメモリ, ポリゴン, レイヤー, レンダリング, ワールドワイドウェブ, 瑕疵, 家庭用品品質表示法*, 川崎病*, 機械翻訳, 気候変動枠組条約*, 形態素解析, 口蹄疫, 骨粗しょう症, 在職老齢年金*, 債務超過, 酒船石遺跡*, 自然言語処理, 情報検索, 人工知能, 摂食障害, 対外純資産*, 中央競馬*, 日本国憲法*, 脳型コンピュータ*, 白内障, 臨調*
用語群 B（21 語） ･･･ 語義が複数（括弧内は語義数）
ABS(4), ATM(4), B2B(2), FSB*(3), IBS(6), IEC(2), PDC(2), POS(2), RM (4), SWAP(4), TCO(2), TPC(5), VB(2), VIM(3), VOD(2), VR(3), アパッチ (3), エージェント (3), キーボード (2), ハッシュ(3), ブロイラー (2)

*は「現代用語の基礎知識 2001」から採取した語.
無印はネットニュースから採取した語.

表 5 結果：用語定義分類結果
Table 5 Experimental result: classification of term definitions.

定義名称	定義文数	調査文数	判定精度
直接的内包的定義	610	100	89 (89%)
間接的内包的定義	489	100	83 (83%)
略記	86	86	84 (98%)
性質	74	74	58 (78%)
機能	780	100	87 (87%)
目的	394	100	81 (81%)
属性	23	23	22 (96%)
例示的定義	11	11	11 (100%)
他概念との関係	54	54	50 (93%)
構成的定義	12	12	10 (83%)
歴史的	28	28	26 (93%)
現象	3	3	3 (100%)
原因	83	83	72 (87%)
計	2168	774	676 (87%)

676 文が正解であり，判定精度は 87%であった．対象テキストが，不均質で雑音の多いウェブ上のテキストであることを考慮すると，この精度は非常に良い数字であり，実用レベルに達していると考えることができる．

この精度は，情報検索で用いられる 2 つの評価尺度のうちの，precision に相当する．これに対して，もう一方の recall（再現率）は，一般にウェブ検索ではそれほど重要ではない．なぜならば，ウェブ検索において，網羅的に検索するということは，事実上不可能であり，そのようなニーズもほとんどないからである．

表 6 グループ化の結果 (1)
Table 6 Experimental result: grouping #1.

	用語群 A	用語群 B	全体
正解	66 語	11 語	77 語 (81%)
不正解	8 語	10 語	18 語 (19%)
内訳			
(a) 同義が複数グループ	8 語	5 語	13 語 (14%)
(b) グループ内に複数語義	—	3 語	3 語 (3%)
(c) 両方が混在	—	2 語	2 語 (2%)

用語説明の探索という本アプリケーションにおいては，ウェブ上に存在するすべての用語説明が発見できなくても，よいものがいくつか見つかればそれでよい．すなわち，recall よりは，precision の方がより重要である．

7.1.2 グループ化の精度

グループ化の精度を調べるために，まず，表 4 に示す 95 語それぞれに対して，用語説明の収集でいくつの語義が見つかっているかを，手作業で調べた．1 つの語義しか見つからなかった 74 語を用語群 A とし，複数の語義が見つかった 21 語を用語群 B とした．用語群 B の各用語の語義数は，用語の後の括弧内に示している．

グループ化の精度を調べる場合，何を単位（分母）とするかにおいて，いくつかの選択肢がある．ここでは，まず，用語を単位として，グループ化の精度を調べた．この結果を**表 6** に示す．

用語群 A は，語義が 1 つしか存在しないため，1 つのグループにまとまった場合が正解であり，それ以外の場合は，同義の用語説明が 1 つのグループにまとまりきれず，複数のグループを構成してしまったということになる（誤り a）．一方，用語群 B は，見つかった語義の数と同じ数のグループが形成され，それぞれのグループは同義の用語説明のみから構成される場合が正解となる．不正解は，同義の用語説明が 1 つのグループにまとまりきれなかった場合（誤り a），異なる語義の用語説明が 1 つのグループにまとまってしまった場合（誤り b），それら両者が混在する場合（誤り c），の 3 つの場合に分けられる．この評価法は，最も厳しい評価法であるが，95 語中，77 語（81%）が正解であった．誤り a は比較的被害が少ない誤りであるのに対し，誤り b および誤り c は深刻である．この深刻な誤りは 5 語（5%）だけであった．

以上の結果より，全体としては，グループ化は良好であると考えることができるが，用語群 B に含まれる用語の数が 21 語と少ないため，確信を持って判断

表 7 グループ化の結果 (2)
Table 7 Experimental result: grouping #2.

	グループ数
同義の用語説明のみ	64 (93%)
複数の語義の用語説明を含む	5 (7%)

表 8 代表の選抜精度と上位語の確定精度
Table 8 Experimental result: the best explanation and the best broader term.

	適切	不適切
代表	88 (94%)	6 (6%)
上位語	112 (85%)	19 (15%)

することができない．

そこで，今度は，グループを単位にグループ化の精度を調べた．ここでは，用語群 B の 21 語に対して形成された 69 グループを対象に，それぞれのグループが同義の用語説明のみから形成されているかどうかを調べた．その結果を**表 7** に示す．69 グループ中 64 グループ（93%）は，同義の用語説明のみから構成されていた．すなわち，本グループ化の方法は，異なる語義の用語説明を 1 つのグループにまとめてしまう危険性は低いと判断できる．

以上の結果より，グループ化の精度は，ほぼ実用レベルに達していると判断できる．

7.1.3 代表の選抜精度と上位語の確定精度

- 代表の選抜精度：複数の用語説明を持つ 94 のグループにおいて，代表の選抜によって選ばれた用語説明が，そのグループの代表として適切かどうかを人手で調べた．
- 上位語の確定精度：システムが上位語を出力した 131 のグループにおいて，出力された上位語がそのグループの上位語として適切であるかどうかを人手で調べた．

これらの結果を**表 8** に示す．代表の選抜精度は 94%，上位語の確定精度は 85%であった．

7.2 検　　討

用語定義文の判定，グループ化，代表の選抜，上位語の確定，の 4 つの精度は，いずれも良好であり，ほぼ実用レベルに達している．これらの点から，本システムの目標，すなわち，与えられた用語の説明を探し出し，それらを整理して提示するということは，ほぼ達成できたと考えることができる．

これまで述べてきた実験で，システムが設計どおり動作することを確認できたわけであるが，システムの有用性を評価するためには，別の評価実験が必要である．本システムの最終的な目的は，辞書や辞典には掲載されていない新しい用語の意味をウェブで調べることを自動化することであるから，有用性の評価は，理想的には，次の 3 つの観点から行うことが望ましい．

(1) カバレジ：辞書や辞典には掲載されていない新しい用語の説明をどの程度発見できるか．
(2) 出力品質：出力される用語説明の品質はどの程度であるか．
(3) 時間短縮：本システムを用いることによって，調査時間をどの程度短縮できるか．

しかし，これらの観点に沿った評価実験を設計することは，それほど簡単ではない．今後，本システムの有効性をどのようにして評価していくかを考えていく必要がある．

一方，システムの内部においても，各種の改善の余地がある．その中で最も重要なものは，用語定義文の判定精度の向上である．用語定義文の判定は，本システムの根幹部分であり，その精度は，グループ化，代表の選抜，上位語の確定などの他の処理精度に大きく影響する．判定に用いる文型パターンの見直しや，段落抽出における文末判定アルゴリズムの改良などによって，さらなる精度向上を実現する必要がある．

本システムは，用語説明として 1 つの形式段落を抽出する方針をとっているが，これは，その段落全体が，その用語を説明する文章となっていることを仮定している．しかし，そのような仮定は必ずしも成り立つわけではない．たとえば，表 3 の「安全システム」の代表の用語説明では，最初の 1 文は「ABS」の直接的内包的定義を与えているが，次の文は「ABS」の説明とはなっていない．このような場合に対処するために，段落内の文章構造（文間のつながり）を把握し，用語を説明している部分だけを抽出することを実現する必要がある．

8. 関 連 研 究

テキストからの用語説明の抽出に関連する研究に，黒橋らの研究[3]，西野らの研究[4]，木田らの研究[5]，藤井らの研究[6),7)] がある．

黒橋ら[3] は，辞典の用語説明に見られる定義文をパターン化し，同義語文，内包的定義文，外延的定義文を自動的に抽出することを実現した．本研究の定義文抽出の方法論はこの研究に準拠しているが，黒橋らの対象が高品質の辞典のテキストであるのに対し，我々は低品質なウェブ文書を対象としている点が大きく異なる．

西野ら[4] は，一般文書を対象として，用語定義文をパターン化することで，テキストからの用語説明の自動抽出を実現した．彼らが用いたパターンは，「X と

はYである」（直接的内包的定義文）に限定されており，抽出するのは，それにマッチする文のみである．これに対して，我々は，直接的内包的定義以外にさらに12種類の定義文を設定し，定義文を含む1段落を抽出する．

木田ら[5]は，新聞記事から用語集を作成することを目的として，「AはB」というパターンの文の分析を行っている．

藤井ら[6),7)]が提案するウェブからの事典的知識収集法は，目的やシステム構成において，本論文で我々が提案したシステムと多くの共通点がある．しかし，以下の5点で大きく異なる．

(1) 定義文の分類とパターン化：藤井らは，辞典の説明文から定義文のパターンを自動的に生成するが，定義文の種別は導入していない．これに対して我々は，定義文に13種類の種別を設定し，そのそれぞれに対して人間がパターンを作成している．

(2) 用語説明の抽出：藤井らは用語説明として連続する3文を抽出するが，我々は形式段落を自動的に判定し，用語説明段落を抽出する．

(3) 用語説明の解析：藤井らはトライグラムモデルを用いて，抽出した3文の言語らしさを判定し，ノイズ除去を行っている．これに対して我々は，抽出した段落に13種類の定義文が存在するか否かを判定し，その段落がその用語の説明として適切かどうかを判定している．

(4) 用語説明のグループ化：藤井らのクラスタリング手法は，あらかじめクラスタ数を与える必要があるのに対して，我々のグループ化手法は，あらかじめグループ数を与える必要がない．

(5) システム形態：藤井らのシステムはオフラインシステムとして動作するのに対し，我々のシステムは，オンラインシステムとして動作する☆．

本システムと姉妹関係のあるシステムに，ウェブからの住所情報を探索するシステム[8)]，ウェブから人物情報を探索するシステム[9)]がある．本システムを含むこれらのシステムは，いずれも，ある限定された種類の情報（本システムの場合は，用語説明）をウェブから探し出し，その結果を適切に整理して提示する，メタサーチエンジン[10)]と考えることができる．

9. まとめ

本研究では，用語を入力すると，その用語を説明する文章をウェブから探し出し，それらを整理して出力するシステムを実現した．本システムは，次の特徴を持つ．

- 13種類の用語定義文を自動判定する．
- 用語説明として形式段落を抽出する．
- 同義判定によるグループ化，代表の選抜，上位語の確定などの編集処理を行う．

本システムを使用することにより，辞書を引いて調べるのとほぼ同等の簡便さで，ある用語の意味を「ウェブを用いて調べる」ことができる．

☆ 現在の平均的な応答時間は約2分である．

参考文献

1) 長尾 真：知識と推論，岩波書店 (1986).
2) 長尾 真：辞典形式での専門分野の知識の体系的構成法，人工知能学会誌，Vol.7, No.2, pp.320–328 (1992).
3) 黒橋禎夫，長尾 真，佐藤理史，村上雅彦：専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol.7, No.2, pp.336–345 (1992).
4) 西野文人，橋本三奈子，落谷 亮：テキストからの用語とその定義文の抽出，言語処理学会第5回年次大会発表論文集，pp.124–127 (1999).
5) 木田敦子，乾 裕子，落谷 亮，西野文人：新聞記事からの用語集作成のためのテキスト分析，情報処理学会研究報告，1999–NL–134–12, pp. 85–92 (1999).
6) 藤井 敦，石川徹也：用語説明抽出に基づくWeb文書の事典的利用，言語処理学会第6回年次大会発表論文集，pp.296–299 (2000).
7) 藤井 敦，石川徹也：World Wide Webを利用した百科事典的知識の収集法，人工知能学会研究会資料，SIG–KBS–A001–6, pp.31–36 (2000).
8) 佐藤理史：ワールドワイドウェブを利用した住所探索，言語処理学会第6回年次大会，pp.447–450 (2000).
9) 山本あゆみ，佐藤理史：ワールドワイドウェブからの人物情報の自動収集，情報処理学会研究報告，2000–ICS–119, pp.173–180 (2000).
10) Etzioni, O.: Moving Up the Information Food Chain, *AI Magazine*, Vol.18, No.2, pp.11–18 (1997).

(平成12年8月24日受付)
(平成14年2月13日採録)

桜井　　裕（学生会員）

1997 年東京工業大学工学部電気・電子工学科卒業．1999 年北陸先端科学技術大学院大学情報科学研究科修士課程修了．現在，北陸先端科学技術大学院大学情報科学研究科博士後期課程在学中．

佐藤　理史（正会員）

1983 年京都大学工学部電気工学科第二学科卒業．1988 年同大学院博士課程研究指導認定退学．京都大学工学部助手，北陸先端科学技術大学院大学情報科学研究科助教授を経て，2000 年より京都大学大学院情報学研究科助教授．1997 年より 2000 年まで科学技術振興事業団研究員を兼任．京都大学博士（工学）．自然言語処理，機械学習，情報の自動編集等の研究に従事．言語処理学会，日本認知科学会，AAAI，ACL 各会員．著書：『自然言語処理』（共著，岩波書店，1996），『アナロジーによる機械翻訳』（共立出版，1997），『言語情報処理』（共著，岩波書店，1998）等．