高次辞書データベースのための語彙知識自動獲得システム

Automatic Word Knowledge Acquisition System for Advanced Dictionary Database

亀田弘之[1]・藤崎博也[2]
Hiroyuki KAMEDA[1] ・ Hiroya FUJISAKI[2]

1. 〒192 東京都八王子市片倉町1404-1 東京工科大学工学部
2. 〒278 千葉県野田市山崎2641　　東京理科大学基礎工学部

1. Faculty of Engineering, Tokyo Engineering University,
   Katakura 1404-1, Hachioji-City, Tokyo 192, JAPAN
2. Faculty of Industrial Science and Technology,
   Science University of Tokyo,
   Yamazaki 2641, Noda-City, Chiba 278, JAPAN

キーワード：知識獲得，未知語，辞書データベース，自然言語処理
Keywords： knowledge acquisition, unknown word, dictionary database, natural language processing

あらまし： そもそも自然言語の語彙は、人間が世界を表現・記述するために必要に応じて創造するものであり、いわば“開いた集合”である。従って、新しい単語が創造される度毎に、辞書に登録する必要があるが、これを機械により支援するシステムはまだない。本稿では、真に実用的な自然言語処理技術の実現に寄与するとともに、言語学・辞書学等の人文科学の研究にも役立つことを意図して構築中の、単語辞書と統語規則とを用いてべた書き日本語文から未知語を抽出するとともに、新たに得られた単語を新語として単語辞書に自動的に登録することのできる未知語獲得システムについて述べる。また、本システム構築のために補助的に作成したユーティリティの概略についても述べる。

Summary: Vocabulary of a language is in general an "open set," for human creatively to express and describe the world. This fact forces natural language systems to have an ability to provide new words (henceforth: unknown words) with the system dictionary when they are found. But no natural language systems can still support automatic word registration to the system dictionary. In this paper, automatic word knowledge acquisition sysytem, which both aims at realizing apractically useful natural language processing system, and also supports wide areas of studies on the Humanities, e.g. linguistics and lexicology, is presented as well as auxiliary utilities for implementing the system.

1. はじめに

自然言語は、人間相互の意思疎通のための手段であるとともに、人間が自己の内外に生起する様々な事象を認識・記述し、さらにそれらを素材として知識を形成・蓄積し、かつ、思考を巡らせるための媒体である[1]。これゆえに自然言語の語彙は特に、社会や時代の進展・変化にともない、必要に応じて創造され、いわば“開いた集合”となっている。

一方、自然言語を機械により処理する研究は、計算機の黎明期から積極的に手がけられており、現在では機械翻訳システムやワードプロセッサ等が商品化されるに至っているものの、既存の自然言語処理システムでは、辞書・文法はいずれも予め限定されており、新しい表現、すなわち機械にとっての未知の表現に対する処理能力を欠いている[2]。

未知語の取扱いに関する研究は従来から部分的になされており[3-11]、著者らもこの問題を解決し、真に実用的な自然言語処理技術を完成させることを主目的として、未知語処理の研究に早くから従事している[1,3]。本稿ではその内、本格的な自然言語処理用の辞書データベースのための語彙知識自動獲得システム、すなわち、予めシステムに準備された辞書データをもとに、機械が未知語を検出し、その品詞等の語彙知識を推定・獲得することのできる未知語獲得システムについて、その概要と動作例について述べる。

2. 未知語の定義・分類と未知語処理の定義

2-1. 未知語の定義[3]

人間は、言語を媒体として相互の意思疎通を行う場合、状況や背景的知識等を適宜利用して、見かけ上同一の表現であっても多様な意味を表現・伝達することができるとともに、さらには必要に応じて新たな単語や表現を創造する。これに対して、その受信者は、多くの場合何の支障もなく、それらの単語や表現に担われた発話者の意図する意味を円滑に推察し理解することができる。人間がこのようなことをなし得るのは、人間が多義的な表現に対して発信者の意図した意味を推察することができるとともに、新たに創造された初見の表現に対しても、ほとんどの場合その意味を推察することのできる能力を持っているからである。つまり初見（未知）の単語であっても、文字・形態素・造語法・統語情報・談話情報等の言語的知識や、文脈・背景的知識等の言語外知識を用いて、初見であることすら気付くことなく迅速かつ適切にその意味を理解することができる。人間はこのような優れた能力を持ち合わせているために、“人間にとっての未知語”という概念は、必ずしも明確には定義することはできない。

一方、現在の機械は、上述したような高度な処理能力を持っておらず、一般的には、単語辞書と、単語の配列を規定する文法規則（統語規則）とを主たる知識として言語処理を行っているため、多義的な表現や新しい創造的な表現は、十分に処理することができない。特に、システムの単語辞書に予め載っていない単語は、システムにとっては未知となる。本稿ではこのような観点から、「機械にとっての未知語」とは未登録単語のことであるとし、以下の議論ではこの定義によるものとする。

2-2. 未知語の種別

未知語には、大きく分けると３つの種別があり、本研究ではそれらを第一種の未知語・第二種の未知語・第三種の未知語と呼ぶこととする。以下にそれらの定義と実例とを示す。なお、この定義に関する詳しい説明は、参考文献[3,12]を参照されたい。また、以下の未知語の実例（下線の付されたもの）はすべて、広辞苑（第４版・岩波書店）の主見出しとして記載されていないものである。

2-2-1. 第一種の未知語

【第一種の未知語の定義】 単語自体は辞書に登録されているにもかかわらず、表記が辞書のものと異なるために、辞書検索に失敗する単語（異表記同義語）のこと。

この種の未知語は、日本語における表記の多様性によるものであり、例えば、単語“慶ぶ”は、広辞苑（第４版）の主見出しに記載されている表記“喜ぶ”あるいは“悦ぶ”とは一致しないために、この辞書をもとにした処理では、未知語（未登録語）扱いとなる。

第一種の未知語はさらに細かく分類することができる。以下(1)～(5)にその例を示す。

(1) 漢字異表記： 異なる漢字で表記されているために生じる未知語。
例：「喜ぶ」と「慶ぶ」

(2) 送りがな異表記： 送りがなの付け方の違いにより生じる未知語。
例：「行う」と「行なう」

(3) 混ぜ書き異表記： 漢字と平仮名等の混ぜ書きにより生じる未知語。
例：「飛び込む」と「飛びこむ」

(4) 片仮名異表記： 片仮名表記の異いにより生じる未知語。この種類の未知語はさらに３つに分類できる。
・大小文字異表記： 片仮名大小文字の異いにより生じる未知語。
例：「ソフトウェア」と「ソフトウエア」
・長音記号異表記： 長音記号により生じる未知語。

例：「コンピューター」と「コンピュータ」
・外来語異表記：外来語表現の異いにより生じる未知語。
例：「バイオリン」と「ヴァイオリン」
(5) 記号の異表記：　数字表記や単位表記の異いにより生じる未知語。
例：「百」と「１００」、「１個」と「１コ」

第一種の未知語はこのように、表記におけるゆれ・慣用的用法・学術（専門）用語の表記規約等に起因するものの他に、「みんなでガンバロー！」のように特定の単語・表現を強調する等の特殊な用法に起因するものもある[13]。

2-2-2. 第二種の未知語

【第二種の未知語の定義】　単語の各構成要素は辞書に登録されているが、その単語自体は辞書に登録されていない単語（既知語を用いて造語された複合語）のこと。

第二種の未知語の例としては、以下のようなものがある。

例：「情報学」　（情報 ＋ 学）
「数学辞典」（数学 ＋ 辞典）
「再試験」　（再 ＋ 試験）

日本語においては、必要に応じてさまざまな複合単語が日常的に造語され利用されるので、この種の未知語の処理は重要である[14]。また、第二種の未知語を機械処理するためにはさらに、いくつかの分類を行う必要があると考えられる。このような観点から、例えば以下のような下位分類が得られる。

(1) 複合語の単語構成要素中に付属的な形態素があるか否かに着目する分類方法。付属的な形態素とは、その単語構成要素自体は単語としての機能を持たず、他の単語（名詞等）に付属することにより、正否、肯否定、是非、程度、性質、状態、範囲、時間、省略等の意味を表す単語構成要素のことをいう。また、接頭語・接尾語もこれに含まれるものとする。以下に例を示す。

例：（付属的な形態素を含む複合語）
「不採用」、「正社員」、「非常識」、
「弱酸性」、「電子化」、「確実性」、
「例外的」、「政府内」、「各国」、
「処理時」、「～等」
（注：強調文字部分が付属的な形態素）
例：（独立した単語構成要素のみから成り、付属的形態素を持たない複合単語）
「証券取引」、「主旨説明」、「最低気温」

(2) 複合語中の各形態素の意味を合成することにより全体の意味が得られるか否かに着目する分類方法。

例：（意味合成可能単語、すなわち、それぞれの意味を合成することで全体の意味が得られる単語）
「人権侵害」、「転送時間」、「最終決定」
例：（意味合成不可能単語、すなわち、それぞれの意味を合成しても全体の意味が得られない単語）
「湾岸支援」、「企業行動憲章」、
「関税貿易一般協定」

2-2-3. 第三種の未知語

【第三種の未知語の定義】　単語の構成要素として、単語辞書に登録されていないものが含まれるもの。

第三種の未知語も以下のように下位分類することができる。未知語の例とともに示す。

(1) 辞書にない単語構成要素（強調文字で表記）を部分的に含む単語
例：「ＩＰアドレス」
「トラブル・シューティング」、
(2) 単語構成要素すべてが辞書にない単語
例：「ボリス・パンキン」
「インターネット」
(3) その他
例：（省略により第三種の未知語となるもの）
「フ諸島」（フォークランド諸島のこと）
「阪大」（大阪大学のこと）

2-3. 未知語処理の定義

“自然言語処理”という用語は、自然言語の理解と生成との両面を指すのに用いられるのと同様に、“未知語処理”も、未知語の理解と生成とを一般には指すが、本稿では、理解の側面のみに着目し、未知語処理を、未知語の検出、内部構造推定、意味推定の３段階に大別する。

3. 各種未知語の処理方法

機械による自然言語処理の場合には、上述した種々の未知語に対して、一般には様々な処理方法が有り得る[3]。以下では、本研究で試作したシステムにおいて、動作を確認した部分に関連するものに重点を置いて述べる。

3-1. 未知語の検出（各種の未知語に共通）

本研究で作成したシステムでの未知語の検出は、原則的には、入力文の統語解析処理の枠組みにおける下記の処理によって行われる。

①入力文から文字列を切出し単語候補とする。
②単語候補に関して辞書検索する。
③辞書に載っていれば既知語であり、未知語の検出は不成功裏に終了する。

④辞書に載っていなければ未知語候補とみなす。
⑤未知語候補の内部構造を調べ品詞を推定する。
⑥品詞推定に失敗する場合には、処理①へ飛ぶ。
⑦推定した品詞が統語解析に矛盾を生じさせなければ、統語解析は成功裏に終了し、その結果、推定が妥当であるとともに、当該の未知語候補は真の未知語であると確認され、未知語検出は終了する。
⑧推定した品詞が統語解析に矛盾を生じさせるならば、統語規則に基づき、他の品詞の可能性も調べ、その可能性があれば処理⑤へ、なければ入力文から新たな文字列を切出して処理②へ飛ぶ。

なお、上述の内部構造推定処理は、各種の未知語あるいは品詞毎に異なる。

3-2. 第一種の未知語の処理

第一種の未知語は、すべての可能な表記を予め辞書に登録することにより処理するか、未知語が出現する度毎に辞書項目に記載されている形(表記)を推定し処理する方法とがある。これらは、出現頻度や未知語処理の処理量との兼ね合いを考慮して最終的に決定することが必要であるが[15]、本研究では、すべての異表記を予め網羅的に知ることが一般には不可能であること、また、単語辞書が不用意に増大しないようにとの配慮から、辞書項目の表記を推定し、異表記同士の文字列照合を行う方式を採用した。

筆者は既に、図1に示す多重照合方式を提案し、C言語によりそのシステムを作成し、基本的有効性を確認しているが[16]、本研究のシステムには、現在、片仮名照合部分のみインプリメントされている。具体的には、片仮名文字列における部分文字列の書換え機能が実装されており、例えば、「イタリヤ」を「イタリア」と書換え辞書検索することができる。

以下、図1の用語について簡単に説明をするが、詳しくは、参考文献[16]を参照されたい。

- **表記照合**：単語の辞書表記をキーとする検索方法。
- **読み照合**：単語の読み（実際のシステムでは、平仮名表記）をキーとする辞書検索方法。
- **普通照合**：表記照合と読み照合の総称名。
- **大小文字照合**：片仮名表記における片仮名小文字と大文字とを同一視して辞書検索する方法。
- **長音記号照合**：片仮名表記における長音記号の有無を無視して辞書検索する方法。
- **外来語照合**：外来語の表記の場合に生じるゆれとして、上記の大文字小文字と長音記号以外にも、「ヴァ」と「バ」、「ヴィ」と「ビ」等がある。これらの異表記を同一視して辞書検索する方法。
- **片仮名照合**：上記、大小文字照合、長音記号照合、外来語照合の総称名。
- **外国語文字照合**：現代日本語の場合、文章中に英語等の外国文字単語が現れることがある。このような異表記に対応するために、外国文字単語（外国語単語）を、対訳辞書を用いて、日本語の単語に変換し辞書検索を行う方法。
- **送り仮名照合**：送り仮名のゆれによる異表記を同一視して辞書検索する方法。
- **混ぜ書き照合**：漢字と平仮名の混ぜ書きによる異表記を同一視して辞書検索する方法。
- **多重照合**：上記すべての照合方法の総称名。

3-3. 第二種の未知語の処理

第二種の未知語の処理では検出の他に、内部構造の推定と意味推定とを行う。以下では、まず、これらの処理に必要な知識を概観した後に、本研究で採用した第二種の未知語の処理方法とそのインプリメント方法について述べる。

3-3-1. 未知語の意味推定に関連する知識[17]

未知語の意味を推定するためには、種々の知識を必要とする。以下に、そのための知識を列挙する。まず知識は、「言語に関する知識」・「発話行為に関する知識」・「発話内容に関する知識」・「発話者に関する知識」の4つに大きく分類される。

「言語に関する知識」とは、媒体としての言語自信に関する知識のことである。また、言語は、語彙や文法規則が相互に密接に関係し合って構築されているという体系(langue)としての側面を有するとともに、個々の具体的発話として現れた諸例(parole)の側面をも持つ。具体的には、体系としての知識と

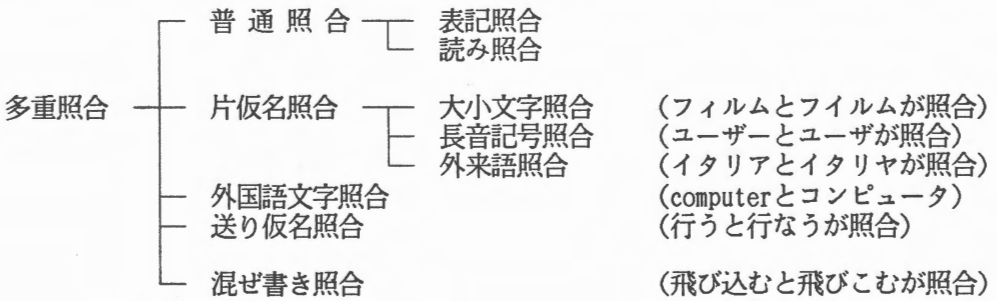

図1. 多重照合の種類

しては、文字セット・文法（統語規則）・各単語の用法等があり、一方、具体的諸例に関する知識としては、用語や文等に関する用例がある。

「発話行為に関する知識」とは、発話の目的・意図に応じて、その目的・意図を達成するための発話計画・方略に関する知識のことである。これはいわゆるコミュニケーション（意思疎通）のノウハウに関するものであり、発話の相手（大人か子供か、専門知識を持っている人か、聴く意志のある人か等）・発話の形態（直接対面、電話、手紙等）などに応じて、伝達したい内容をどの様に整理し、どの様な順番で、どの様な用語・表現を用いれば、より効率的・正確な意思疎通を行ない得るのか、に関する知識のことである。

「発話内容に関する知識」とは、発話により述べられる事実に関する知識のことであり、専門的知識（百科事典的知識）・背景的知識（個別的出来事に関する知識）・常識等がある。

「発話者に関する知識」は、具体的な発話状況において、発話の際の主体者（発話者と聴き手）自体に関する知識のことであり、上記の発話に関する知識を実際に運用する際に利用されるべきものである。

上述の内容をまとめたものを下図に示す。

①言語に関する知識：
- 体系としての言語に関する知識
  （文字・文法・用法等）
- 事例（用例）としての言語の知識
  （用語・例文等）

②発話行為に関する知識：
- コミュニケーション行為のノウハウに関する知識

③発話内容に関する知識：
- 情報伝達媒体としての言語により伝達される情報・知識自体の内容に関する知識
  （対象に関する知識・専門的知識・世界に関する知識・背景的知識等）

④発話状況に関する知識：
- 具体的発話者の特性 等

図2. 未知語の意味推定に関連する知識

未知語の処理を行うためには、このように様々な知識を必要とするが、筆者らは、現在のところ、まず上記①の言語に関する知識のみに着目する処理方法に重点をおき、規則に基づく処理方法と、用例からの類推に基づく処理方法とを取り上げて研究を行っている[18-25]。本研究で作成したシステムでは、そのうち規則に基づく処理が現在組込まれている。

3-3-2. 規則に基づく処理方法[18-20, 23-25]

第二種の未知語の多くはいわゆる複合語であること、また、筆者の行った語彙調査では[3]、未知語の多くは単語構成要素が2個の第二種未知語であることから、本研究では、まず単語構成要素が2個の未知複合語に処理対象を限定してシステムを作成した。このシステムでは、単語構成要素間の関係を規則化することにより、未知語の内部構造と意味とを推定する。

(1) 語内文法

従来の文法（文文法）を参考にして、単語構成要素とそれらの相互関係を分析し、表層レベルと深層レベルの2層の形式で整理した。

①表層レベル：　このレベルでの単語カテゴリとして、名詞的要素、動詞的要素、形容詞的要素、副詞的要素の4つを設定し、このうち、名詞的要素と動詞的要素を中心的に分析、整理した。最終的にインプリメントした規則は以下の通りであり、プログラミング言語prologにより記述されている。

```
表層構造関係(名詞的要素,　動詞的要素).
表層構造関係(動詞的要素,　動詞的要素).
表層構造関係(形容詞的要素, 名詞的要素).
表層構造関係(副詞的要素,　動詞的要素).
表層構造関係(名詞的要素,　名詞的要素).
```

図3. 表層構造に関する規則の例

②深層レベル：　動詞的要素とその他の要素との意味的関係に着目し、まず後者に対して、深層格と意味カテゴリとを設定した。また、動詞的要素に対しては、意味カテゴリと意味パタンとを設定した。なお、意味カテゴリとは、単語構成要素の担う意味の分類カテゴリのことであり、意味パタンとは、動詞的要素に関わる（格文法の意味での）意味構造のことである。以下に、深層格と動詞的要素の意味カテゴリを列挙する。

- 深層格：動作格、対象格、源泉格、目的格、経験者格、受け手格、関係格、受益格、道具格、方法格、役割格、比較格、程度格、場所格、時間格、期間格、原因格、結果格、手段格、目的格、条件格、内容格、範囲格（23個）
- 動詞的要素の意味カテゴリ：存在、属性、占有、関係、知覚状態、感情状態、自然現象、物理的遷移、占有遷移、属性移動、身体動作、生産、社会動作、精神遷移、知覚動作、感情動作、思考動作(17個)

これらの整理された知識をもとに、次頁の図4と図5とに示すような規則をインプリメントした。

```
深層構造関係(表記(Element1), 品詞(Cat1),
      意味(Meaning1), 表記(Element2),
      品詞(Cat2), 意味(Meaning2),
      深層構造(Deep_str), 総合的意味(M))
 :-  Cat1 = 名詞的要素, Cat2 = 動詞的要素,
      動詞パタン(動詞的概念(Element2),
        動作主格(Element1)),
      Deep_str = 動詞パタン(動詞的概念
        (Element2), 動作主格(Element1)),
      M = [Meaning1, が, Meaning2, する].
```

図4. 深層構造に関する規則の例

```
動詞パタン(動詞的概念(旅行), 動作主格(Agnt))
        :- 人間(Agnt).
動詞パタン(動詞的概念(旅行), 場所格(Plc))
        :- 場所(Plc);建造物(Plc).
動詞パタン(動詞的概念(旅行), 時間格(Time))
        :- 時間(Time).
動詞パタン(動詞的概念(旅行),道具格(Instr))
        :- 交通手段(Instr).
```

図5. 動詞的要素の意味パタンの例

(2) 単語構成要素辞書

単語構成要素は多くの場合、単独で単語となり得るので、単語辞書で兼用することも可能であるが、単語と単語構成要素とを明確に区別して取り扱うために、便宜上、単語構成要素は単語構成要素辞書に分離記述した。なお、単語構成要素は現在のところ、６９個登録されている。

```
登録要素(表記(鳥), 品詞(名詞的要素),
        意味(カラス)).
登録要素(表記(鳥), 品詞(形容詞的要素),
        意味(カラスのような)).
登録要素(表記(見学), 品詞(動詞的要素),
        意味(実地に見て知識をひろくする
        (見学する) こと)).
登録要素(表記(見学), 品詞(名詞的要素),
        意味(見学)).
登録要素(表記(方式), 品詞(名詞的要素),
        意味(やり方)).
登録要素(表記(３０日), 品詞(名詞的要素),
        意味(３０日)).
登録要素(表記(小), 品詞(形容詞的要素),
        意味(小さな)).
```

図6. 登録した単語構成要素の例

3-4. 第三種の未知語

入力文から切り出される文字列は、先にも述べたように、まず、単語辞書に登録されているか調べられ、登録されていれば既知語として処理され、もし、登録されていなければ、未知語候補として処理される。この際、未知語候補は、まず、第二種のものと仮定されて処理が続行され、処理に矛盾が生じた場合には、第一種の未知語として処理される。さらに矛盾が生じた場合には、第三種として処理する。なお、この仮定にも矛盾が生じた場合には、切り出された文字列は単語ではないと判断する。

4. 未知語獲得システム

4-1. システムの概要

上述した未知語処理方法を統合し、未知語を獲得するシステムを、DEC製 ノート型パーソナルコンピュータDigital HiNote CT475（主記憶20MB、ハードディスク350MB）上に、Arity/Prolog（Version 5.1, ライフボート社製）を用いてインプリメントし、以下の処理モジュールと知識ベースからなる。

A．処理モジュール
- 主処理モジュール部
- テキスト入力モジュール部
- 統語解析モジュール部
- 第一種未知語処理モジュール部
- 第二種未知語処理モジュール部
- 第三種未知語処理モジュール部
- 単語獲得モジュール部
- 補助関数モジュール部

（これら全体で約２，４００行）

B．知識ベース
- 単語構成要素データベース
- 単語辞書データベース
- 造語規則データベース
- 統語規則データベース

4-2. 処理の流れの概要

本システムの処理の流れの概要を次頁の図７に示す。処理制御の流れは、prologシステムに依存しており、トップダウン的に処理する。以下に図７に準じてアルゴリズムの概略を述べる。

①テキスト読込み： 漢字仮名交じりべた書きの日本語文を、文字列としてキーボードから読込む。

②文リスト構築： 読込まれた文字列を文字毎に分解しリスト構造の形式に変換する。変換結果を以下、文リストと呼ぶ。

③文法による先頭単語品詞の推定： 文法（統語規則）を参照しながら、文リストの先頭に位置する単語の品詞を、トップダウンに推定する。

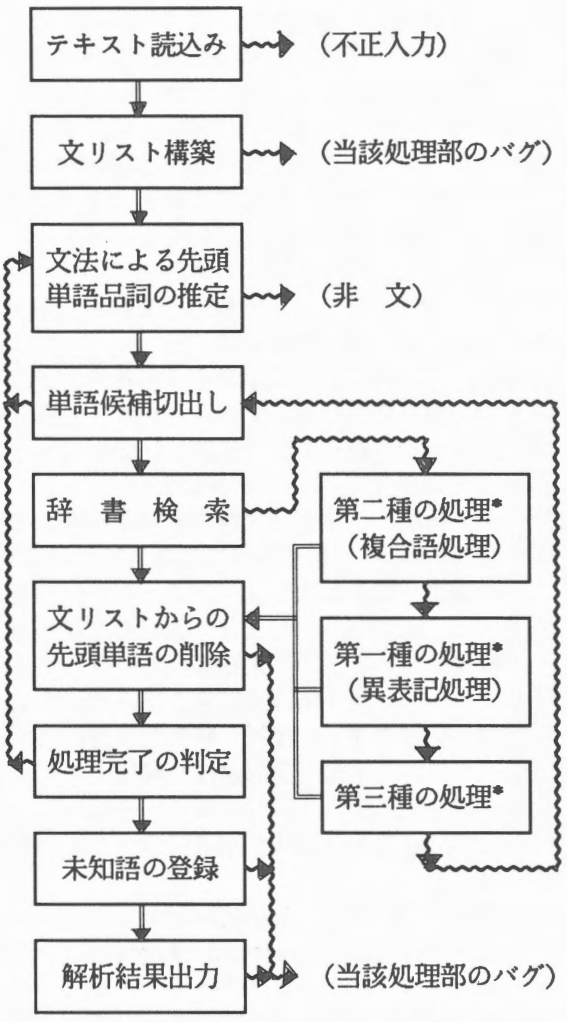

図7. 未知語獲得システムの処理の流れの概要

④**単語候補切出し：**　文リストの先頭側から単語候補として、部分リストを切出す。

⑤**辞書検索：**　単語候補としての部分リストが、辞書に登録されているか検索する。検索に成功すれば、これは未知語ではなく、処理は⑨に移る。検索に失敗する場合は、これを未知語候補とみなし、処理は次の⑥へ移る。

⑥**第二種の処理：**　未知語候補が、第二種の未知語かどうか調べる。第二種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑦へ移る。

⑦**第一種の処理：**　未知語候補が、第一種の未知語かどうか調べる。第一種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ⑧へ移る。

⑧**第三種の処理：**　未知語候補が、第三種の未知語かどうか調べる。第三種の未知語とみなし得る場合には、処理は⑨へ、そうでなければ④へ移る。

なお、上記⑥～⑧が未知語処理の中核部分である。

⑨**文リストからの先頭単語の削除：**　文リストの先頭に位置する単語を削除し、残りのリストを新たな文リストとする。

⑩**処理完了の判定：**　文リストが空リストか調べる。空リストならば、統語解析の処理は完了しているので⑪へ、そうでなければ③へ移る。

⑪**未知語の登録：**　統語解析の際に、未知語が検出されていれば、推定品詞等の情報も統合して、新たな辞書項目として辞書に登録する。

⑫**解析結果出力：**　統語解析結果をディスプレイ上に表示する。

4-3. 動作例

例えば、“イタリヤ”（第一種の未知語）、“見学旅行”（第二種の未知語）、“シェーンな”（第三種の未知語）を含む文として、「シェーンなイタリヤの見学旅行に行った」を入力すると出力として、

『文(主部(名詞句(名詞句no(連体詞(プチな，第三種未知語)，名詞句(名詞(第一種未知語(名詞(イタリア)))、助詞(の)))，名詞句(未知複合語(見学旅行)、助詞(に)))，述部(動詞句(動詞(行く)))』

が表示される。この例では、統語規則の知識とともに、“シェーンな”が連体詞の語尾「な」を持っていること等から連体詞と推定され、“イタリヤ”が“イタリア”の異表記単語として照合され、さらに、“見学旅行”は“見学”と“旅行”が複合語の構成要素になり得るとの知識および造語規則から、第二種の未知語と推定されている。

5. その他のユーティリティ

上述した未知語獲得システムを実際に辞書データベースのために稼働させるためには、単語辞書と統語規則とを充実させる必要がある。本研究では、そのために、以下のようなユーティリティを作成した。

5-1. 単語辞書作成のためのユーティリティ

CD-ROMに格納された広辞苑（第4版）のデータを素材として、prologプログラム形式の単語辞書を作成するユーティリティを作成した。記述言語は、文字列操作言語jgawkである。なお、処理の一部は、テキストエディタVZ(Version 1.6)の文字列検索・置換機能を用いており、現在単語辞書には約6万単語が登録されている。

5-2. 統語規則作成のためのユーティリティ

光学的文字読取り装置(OCR, Optical Character Reader)により電子化した単文(「スペイン語基本文2000」, 大学書林)を、テキストエディタVZを利用して、手作業により単位切りおよび品詞情報のタグ付けを行い、そのタグ付きテキストから品詞列情報およびprolog形式の統語規則を自動抽出することのできるユーティリティを作成した。記述に使用した言語は、jgawkである。現在、これらを利用して統語規則作成のための基礎的資料を蓄積・分析中である。

6. おわりに

本稿では、辞書データベースのための未知語獲得システムの概要とその試作結果および、単語辞書・統語規則作成用ユーティリティについて述べた。

なお、本研究の一部は、文部省科学研究費補助金試験研究(B)(1)(課題番号:07558274, 研究代表者藤崎博也)により行われた。

<<参考文献>>

[1] 藤崎:“言語的思考過程の定式化”,林大(編), 講座 現代の言語2「言語と思考の発達」第2章, 三省堂(1984).

[2] 藤崎・亀田:“知識獲得研究の展望”, 電子情報通信学会第二種研究会「言語獲得と概念形成」, LA90-1, pp.1-8(1990).

[3] 藤崎・亀田 他:“人間の言語処理過程のモデルに基づく自然言語理解システムの構築”, 昭和63年度科研費特定研究「言語情報処理の高度化のための基礎的研究」第6班研究発表資料集(1989).

[4] 吉村・武内・津田・首藤:“未登録語を含む日本語文の形態素解析”, 情報処理学会論文誌, Vol.30, No.3(1989).

[5] 永瀬:“形態素タイプを用いた日本語構文解析前処理”, 情報処理学会第41回全国大会(1990).

[6] 大沢・藤崎:“未知語を含む文の形態素解析システム”, 情報処理学会第42回全国大会(1991).

[7] 植田・小松・横尾・宮崎:“部分複合語による複合名詞構造解析”, 情報処理学会第43回全国大会(1991).

[8] 石川・伊藤・牧野:“文節オートマトンを用いた未知語処理法”,電子情報通信学会第二種研究会「言語獲得と概念形成」,LA92-17,pp.1-8(1993).

[9] 山田・山村・佐川・大西・杉江:“英文における未登録語の意味推定の検討”,情報処理学会研究報告, Vol.93,No.1, 93-NL-93, pp.63-70(1993).

[10]神岡・安西:“仮説生成機構を用いた未知語を含む文の解析”, 人工知能学会誌, Vol.3, pp.627-638(1988).

[11]P. Norvig:“Paradigms of Artificial Intelligence Programming”,Morgan Kaufmann(1992).

[12]亀田・藤崎・森田・倉島:“未知語の分類とその処理に関する考察”, 情報処理学会第36回全国大会講演論文集, 5T-5, pp.1195-1196(1988).

[13]富田隆行・眞田和子:“表記”, 教師用日本語教育ハンドブック2 改定版,国際交流基金(1988).

[14]荻野綱男:“名詞辞書に含まれるべき見出しの範囲 ── 特に複合名詞の扱いをめぐって ──”, ソフトウェア文書のための日本語処理の研究-8 ── IPAL補完文法 ──, 情報処理振興事業協会, 61技-072, pp.207-221(1987).

[15]亀田:“未知語処理機能をもつ自然言語処理システムにおける語処理の効率について”, 電子情報通信学会第二種研究会「言語・知識の運用と獲得」研究発表資料, LK92-4(1992).

[16]亀田・藤崎:“日本語文章の形態素解析における未知語の獲得”,電子情報通信学会第二種研究会「言語獲得・概念形成」,LA90-15,pp.1-8(1991).

[17]亀田:“用例からの類推にもとづく知識の獲得と一般化について ── 未知複合語の獲得を中心にして ──”, 電子情報通信学会第二種研究会「言語・知識の運用と獲得」研究発表資料(1993).

[18]亀田:“言語知識の獲得過程を解明するための心理実験と未知複合語解析システムの試作”, 平成3年度科研費重点領域「知識科学」成果報告書(1992).

[19]KAMEDA:“A Processing Method of Class-2 Unknown Japanese Compound Words of Two Components with Use of In-Word Grammar and Its Prototype System Implementation”, IEEE, TENCON'92, pp.710-714(1992).

[20]亀田:“未知語の意味推定過程解明のための実験と未知複合語の意味推定システム基本処理部の試作”, 平成4年度科研費重点研究「知識科学」成果報告書(1993).

[21]波多野・小嶋:“未知語の意味の推定・獲得の過程”, 平成4年度科研費重点領域「知識科学」シンポジウム論文集, pp.45(1993).

[22]亀田・波多野・小嶋:“用例からの類推に基づく未知語意味推定システム”, 人工知能学会第8回全国大 会23-7,pp.653-656(1994).

[23]亀田・桜井:“語内文法に基づく未知複合語意味推定システムの作成と評価”,人工知能学会第8回全国大会, 23-8, pp.657-660(1994).

[24]亀田・桜井:“べた書き日本語文からの未知語獲得システムの作成”, 電子情報通信学会「思考と言語」技報TL94-11, pp.17-24(1994).

[25]亀田・桜井:“統語解析処理にもとづく未知語獲得システムの試作”, 電子情報通信学会総合大会講演論文集「基礎・境界」, pp.474-478(1995).