既知形態素からなる未知複合語の概念推定とその辞書登録

鈴木 匡芳[1] 中村 宏[1] 阿部 賢司[1] 藤崎 博也[1] 亀田 弘之[2]

[1] 東京理科大学 [2] 東京工科大学

1. はじめに

人間社会においては、日々新たな事実が発見され、また新たな出来事が発生するため、それらを述べる学術論文や新聞記事では、従来にはない表現や語句が作られ、比較的高い頻度で用いられる。従来の自然言語処理システムでは、辞書と、単語の配列を規定する文法規則とを主な知識として処理を行なうため、システムの辞書への未登録語 (以下、未知語 [1] と呼ぶ) の存在は、処理精度の低下を招く。従って入力文章中の未知語を検出し、その概念を推定して辞書に登録することは、処理精度の向上のために極めて重要である。

我々はすでに限定的ではあるが、未知語を収集・分類し、未知語の概念推定の方法について提案した [4][5]。本稿では、学術論文のタイトル、概要、キーワードから、未知語の実例を多数収集し、分析した結果について述べる。さらに、収集した未知語のうち、特に出現頻度の高い、未知複合語に関して、語内構造と語構成要素の概念から、語全体の概念を推定し、辞書に登録する方法について述べる。

2. 未知語の収集・分類

学術論文における未知語を処理するためには、まず、大量のテキストデータについて未知語の実態を定量的に把握する必要がある。本研究では、学術情報センターから提供されている情報検索システム評価用テストコレクション 1[3] の学術論文を利用した。各論文のテキストデータには、論文タイトル・著者名・所属・概要・キーワードなどが記載されている。

一方、システムの辞書としては、日本電子化辞書研究所において作成された、EDR 電子化辞書の日本語単語辞書 (登録語数：395,014)、および、専門用語辞書 [情報処理](登録語数：196,921)[2] を用いた。

ここではまず、上記テキストデータの一部、論文100件に掲載されている日本語名詞列 2,501 語 (延べ 4,456 語) を人手により取り出し、システムの辞書に登録されていない語を未知語として抽出した結果、異なりで 55.5%(1,387 語)、延べで 42.0%(1,866 語) が未知語であった。

さらにこれらの未知語を、筆者らの定義 [4] によって、4 種類に分類した結果は表 1に示す通りである。なお、参考までに第 1 ～第 4 種の未知語の定義を以下に示す。

第 1 種の未知語： 語の表記が辞書に登録されているが、それに対応する概念が辞書に登録されていない語。

第 2 種の未知語： 表記が辞書に記載されているものと異なるために、辞書照合に失敗する語。日本語における表記の多様性に起因するもので、漢字の異なり、送りがなの付け方の異なり、外来語のカタカナ表記の異なりによる。

第 3 種の未知語： 語の各構成要素は辞書に登録されているが、その語自体は辞書に登録されていない語。日本語では、造語の自由度が高いため、新造語の出現頻度が高い。

第 4 種の未知語： 上記の第 1 ～第 3 種の未知語以外の語。辞書に登録されていない人名やカタカナ表記の学術用語などが多い。

表 1は、第 3 種の未知語の割合が極めて高いことを示している。

表 1 未知語の分類結果

未知語の種類	異なり単語数 [語]	延べ単語数 [語]
第 1 種	50 (3.6%)	75 (4.0%)
第 2 種	15 (1.1%)	20 (1.1%)
第 3 種	1,145 (82.5%)	1,520 (81.4%)
第 4 種	170 (12.8%)	251 (13.5%)

3. 未知語処理システム

未知語処理システムの構想を図 1に示す。以下では、図 1に則して未知語処理のアルゴリズムの概略について述べる。

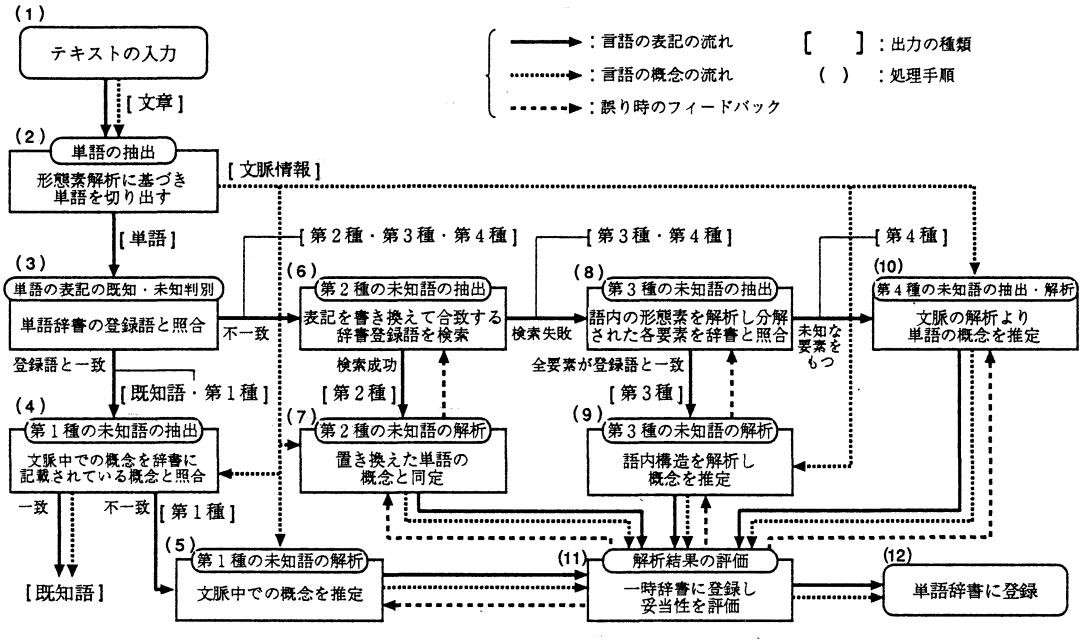

図 1. 未知語処理システムの概要

(1) テキストの入力： 漢字かな混じりの日本語文を読み込み、(2) へ移る。

(2) 単語の抽出： 入力文に対して、形態素解析を行ない、名詞、付属語 (接頭語や接尾語)、形容詞語幹、形容動詞語幹、動詞語幹がそれぞれ連接する場合には、それらを接続して、1つの単語とし、(3) へ移る。

(3) 単語の表記の既知・未知判別： 単語をシステムの辞書に記載されている表記と照合する。照合できた場合には (4) へ、そうでない場合には (6) へ移る。

(4) 第1種の未知語の抽出： (3) において、辞書と照合できた語の中には、文脈中での概念と辞書に記載されている概念が一致する既知語と、一致しない第1種の未知語が混在するため、文脈情報を利用してこれらを分離する。第1種の未知語の場合は (5) へ、既知語の場合は処理を終了する。

(5) 第1種の未知語の解析： (4) において、抽出された第1種の未知語の概念を文脈情報によって推定し、(11) へ移る。

(6) 第2種の未知語の抽出： (3) において、辞書と照合できなかった語は、置き換え規則に従って表記を置換し、辞書に記載されている表記と照合を行なう。照合できたものは (7) へ、そうでない場合には (8) へ移る。

(7) 第2種の未知語の解析： (6) において、置き換えた単語の概念と抽出された第2種の未知語の概念を同定し、(11) へ移る。

(8) 第3種の未知語の抽出： (6) において、第2種の未知語として抽出されなかった語に対して、形態素解析を行ない、分かち書きができた語を第3種の未知語として抽出し、(9) へ移る。そうでない場合には (10) へ移る。

(9) 第3種の未知語の解析： (8) において、第3種の未知語として抽出された語について、語内構造を解析して、その概念を推定し、(11) へ移る。

(10) 第4種の未知語の抽出・解析： (8) において、第3種の未知語として、抽出されなかった語を第4種の未知語として抽出し、文脈情報よりその語の概念を推定し、(11) へ移る。

(11) 解析結果の評価： (5)、(7)、(9)、(10) において、概念が推定された語を一時辞書に登録し、辞書の使用

状況などから妥当性を評価する。推定された概念が妥当と判断されたものについては、(12) へ移る。そうでない場合には、誤った推定結果であると判断し、(5)、(7)、(9)、(10) へ戻る。

(12) 単語辞書への登録：(11) において、未知語の概念が妥当であると判断されたものを、単語辞書に登録し、処理を終了する。

ここで、(4)、(5)、(10)、(11) に関しては、文脈情報を用いた処理の完全な自動化が困難なため、現在のところは人手で行なった。

4. 未知複合語の処理の概要

第 3 種の未知語の語内構造を表層・深層の両面から解析し、その概念を推定する方法は以下の通りである。

4.1 語構成要素辞書の作成

語内構造にもとづいて、未知複合語の概念を推定するには、語構成要素の表層的な役割、深層的な役割を記述した語構成要素辞書が必要である。この辞書は、システムの辞書として扱われる単語辞書には記述されていない、それぞれの要素が取り得る、表層的な役割、深層的な役割を記述したものである。

まず、表層レベルでは、名詞的要素 (N)、動詞的要素 (V)、形容詞的要素 (ADJ)、副詞的要素 (ADV)、付属的要素 (AFF) の 5 つのカテゴリを設定する。これらの要素 2 つから構成される組み合わせを語構成パターンと呼ぶ。与えられた未知複合語に含まれる 2 つの要素が文法上許されるパターンのどれに属するかを分析し、語の表層構造を決定する。

次に深層レベルでは、格文法の考え方を参考にする。一般的に格文法は、動詞に対する前後の他の語の意味的な役割を問題として取り扱うが、語内構造では、動詞的要素以外のものにも拡張して用いる。

ここでは、主体格、対象格、目的格、手段・方法格、材料・道具格、場所格、存在格、時間・期間格、条件格、仕様格、状況格、状態格、事象格、源泉格、方向格、程度格、所有格、動作格、使役格、受け手格、受益格、名称格、結果格、可能格、役割格、関係格、経験者格、範囲格、原因格、対等接続格の 30 種類の深層格 (格フレーム) を設定し、2 要素間の格関係を分析して、語の深層構造を決定する。

ここでは、語構成要素の表層的な役割、深層的な役割を記述した登録要素数 1,217 個の語構成要素辞書を人手により作成し、これを語内構造解析に使用した。

4.2 語内構造解析

語内構造の解析には、Shift-Reduce パーザ [5] を用いた。その動作を「設計支援環境」の解析を例として説明する。(図 2参照)

まず、2 つの語構成要素がスタックに蓄積されるまで、入力バッファの先頭の要素から順次読み込む。「設計」と「支援」がスタックに読み込まれた段階 (ステップ 3) で、2 つの要素の係り受けを表層構造、深層構造の両面から分析する。その結果、表層構造では、語構成パターンは “名詞的要素 (N) ＋動詞的要素 (V)” と判断され、深層構造では、格フレームは “対象格, ～を･･･すること” と判断される。従って、「設計」と「支援」は、接続可能であると判断され、動作

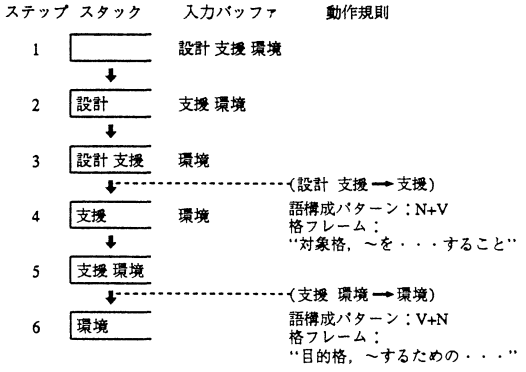

図 2. 語内構造の解析例 (「設計支援環境」の場合)

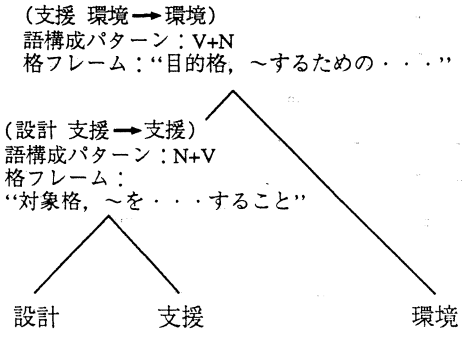

図 3. 「設計支援環境」の語内構造

規則 "A,B → B" が適用される。これは、「設計支援」の概念を「支援」という上位概念に置き換えることを意味する。なお、最末尾の語構成要素が付属的要素 (AFF) の場合には、先行する要素と接続した形で語構成要素辞書に登録することによって、処理を継続する。また、表層構造、深層構造を分析した結果、2つの要素が接続不可能であると判断された場合には、入力バッファの先頭の要素をスタックに追加する。以下、同様の操作を行ない、スタックの要素が、最末尾の要素になるまで、処理を繰り返す。これによって、「設計支援環境」の概念は、「設計することを支援するための環境」と判断される。図3は、このようにして解析された語内構造を示すものである。

1,145 個の未知複合語に対して、語内構造を解析した結果、1,279 通りの語内構造の解析結果が得られた。これを人間が判断した結果、1,159 通り (全体の 90.6%) は、語内構造が正しかった。また、全体の 8.6% にあたる 98 語は、複数の語内構造を持つことが確認された。この曖昧性を軽減するためには、語外の文脈を対象とした構文解析、意味解析が必要であるので、その処理をさらに検討中である。

4.3 辞書登録

概念推定に成功した未知複合語を辞書登録する際には、辞書に登録されている既知語と関連付ける必要がある。本研究では、その一方法として、未知複合語に対して行なった語内構造解析を、辞書の既知語に対しても同様に行ない、既知語の表層構造と深層構造に関する情報を辞書に付加した。これにより、既知語と未知語とを、語内構造に着目して、関連付けることが容易になる。

この関連付けの一例として、「若齢者」の場合を具体的に説明する。「若年者」はシステムの辞書に登録されているが、「若齢者」は、登録されていないものとする。「若年者」を構成する「若年」と「者」、「若齢者」を構成する「若齢」と「者」は、「者」は共通で、「若年」と「若齢」は同じ概念の要素である。また表層構造では、語構成パターンは "形容詞的要素 (ADJ) + 名詞的要素 (N)"、深層構造では、格フレームは "～な..." の語内構造をとるため、「若齢者」は「若年者」と関連付けることができる。

5. おわりに

本稿では、既知形態素からなる未知複合語を語内構造、および、各語構成要素の概念から、概念を推定し、辞書に登録する方法について検討した結果を述べた。

参考文献

[1] 亀田 弘之 : "日本語文章理解における未知語とその処理," 知識科学の最前線シンポジウム論文集別添資料, pp. 1-11 (1993).

[2] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (第 2 版) (1995).

[3] http://www.nacsis.ac.jp/nacsis.index.html

[4] 阿部 賢司, 鈴木 匡芳, 大野 澄雄, 亀田 弘之, 藤崎 博也: "対話による支援を考慮した未知語の概念推定," 言語処理学会第 5 回年次大会発表論文集, pp. 381–384 (1999).

[5] 藤崎 博也, 阿部 賢司, 鈴木 匡芳, 亀田 弘之, 白井 克彦: "語内構造に着目した未知複合語の概念推定の一方法," 情報処理学会第 59 回全国大会講演論文集, vol. 2, pp. 325-326 (1999).