語内構造に着目した未知複合語の概念推定の一方法

1N－4

藤崎 博也[1] 阿部 賢司[1] 鈴木 匡芳[1] 亀田 弘之[2] 白井 克彦[3]

[1] 東京理科大学 [2] 東京工科大学 [3] 早稲田大学

1. はじめに

自然言語処理において、システムの辞書に登録されていない語、すなわち未知語 [1] の存在は、処理精度の低下をもたらす大きな要因となる。また、我々は、キーワードの概念にまで遡って検索する知的情報検索システムを提案しているが [2]、概念検索を行なう上でも、未知語の存在は検索精度を低下させるという点で大きな問題となる。したがって、テキスト中の未知語を自動的に抽出し、その概念を推定することの必要性は極めて高い。

このような観点から、我々は、特に、学術情報検索における未知語の実態を定量的に把握することを目的とし、学術論文のキーワードから未知語の実例を多数収集し、分析した。本報では、収集した未知語のうち、特に、未知複合語に着目し、その概念を語内構造から推定する方法を述べる。

2. 未知語の収集・分類

学術情報検索における未知語の実態を定量的に把握するため、学術情報センターより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] に収録されている学術論文の日本語キーワード (英略語も含む) の中に含まれる未知語の実例を収集した。その結果、キーワードの約 6 割が未知語であることを確認した。また、収集した未知語を分析した結果、(1) 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、辞書照合に失敗するもの、(2) 語構成要素は辞書に登録されているが、語全体としては登録されていないもの、(3) 語の構成要素として、辞書に登録されていないものが含まれるもの、の 3 つに大別することができた。本報では、それぞれの未知語を第 1 種、第 2 種、第 3 種の未知語 [1] と呼ぶこととする。また、それらの出現頻度を調べた結果、全体の約 8 割が第 2 種の未知語であることを確認した。

本報では、特に第 2 種の未知語 (未知複合語) に着目して、その概念を語内構造から推定する方法について検討した結果を述べる。

3. 未知複合語の処理方法の概要

未知複合語の概念を推定するための一方法として、語内構造からその概念を推定する方法を提案する。この方法では、未知複合語の各語構成要素の係り受けを表層・深層の両面から分析し、語全体の概念を推定する。

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し、これらの要素の組合せ (語構成パターン) を分析することにより、語の表層構造を推定する。

次に、深層レベルでは、従来の文法 (文文法) における格文法の考え方を参考にして、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、原因格の 29 種類の深層格を設定し、それらの格構造を分析することによって、語の深層構造を推定する。なお、格文法といった場合、一般に動詞的要素に着目した場合の意味表現のことを指すが、ここでは、動詞的要素以外のものにも拡張して用いる。

また、本研究では、各語構成要素に対して、要素の概念・表層カテゴリ・深層格の情報を付加した語構成要素辞書を人手によって作成し、語内構造の解析に用いた。なお、解析に用いた語構成要素辞書は、一般単語辞書とは区別して取扱うものとする。

A method for inferring the concepts of unknown compound words based on analysis of intra-word structures

Hiroya Fujisaki[1], Kenji Abe[1], Masayoshi Suzuki[1], Hiroyuki Kameda[2], and Katsuhiko Shirai[3]

[1]Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

[2]Tokyo University of Technology, 1404-1 Katakura, Hachioji, 192-8580

[3]Waseda University, 3-4-1 Okubo, Shinjuku, 169-8555

4. Shift-Reduce パーザを用いた語内構造の解析

本研究では、Shift-Reduce パーザを用いて、未知複合語の語内構造を解析する。

以下に、未知複合語：「設計支援環境」の解析例を示す (図 1)。解析手順としては、まず、2 つの語構成要素がスタックに蓄積するまで、入力バッファから、先頭の要素を順に読み込む。次に、「設計」と「支援」がスタックに読み込まれた段階 (ステップ 3) において、2 つの要素の係り受けを表層・深層の両面から分析する。その結果、表層レベルでは "N+V" の語構成パターンに適合し、深層レベルでは、格フレーム "対象格, ～を … すること" に適合する。そのことより、「設計」と「支援」は、接続可能だと判断され、動作規則 "設計, 支援→支援" が適用される。このことは、「設計支援」の概念を「支援」という概念に置き換えることを意味する。なお、最末尾の要素が付属的要素の場合には、先行する要素と接続した形で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と判断された場合には、入力バッファの先頭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、未知複合語の最末尾の要素になるまで、処理を繰り返す。このことより、「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、各要素の係り受けを構文木で表すと、図 2 のようになる。

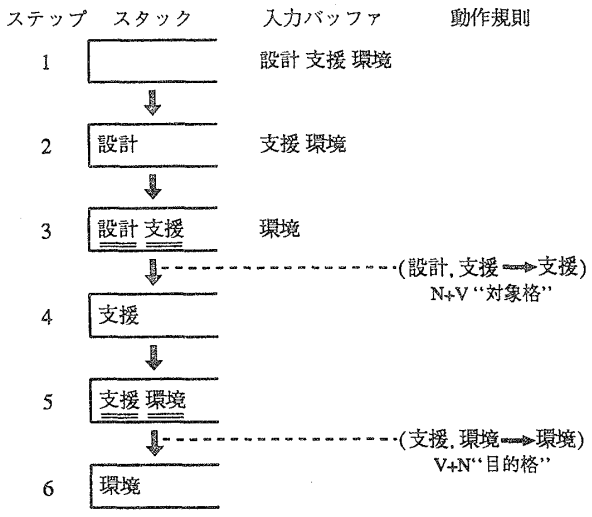

図 1. 語内構造の解析例 (「設計支援環境」の場合)

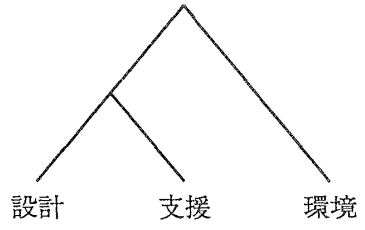

図 2. 「設計支援環境」の場合の構文木

上記の方法を用いて、400 例の未知複合語の概念を推定した結果、1 つの未知複合語あたり平均 2.10 個の概念が解として出力された。推定した概念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といえる。

5. 語構成要素が登録されていない場合の処理

前節では、すべての要素が語構成要素辞書に登録されている場合の処理方法について説明したが、すべての要素が必ずしも語構成要素辞書に登録されているとは限らない。したがって、着目する未知複合語に語構成要素辞書に登録されていない要素が含まれる場合には、その要素の表層カテゴリおよび深層格を推定する必要がある。

本研究では、このような場合、語構成要素辞書に登録されている要素の共起情報を利用して、未知の要素の表層カテゴリおよび深層格を推定した。この方法を用いて、語構成要素辞書に登録されていない要素を含む未知複合語の概念を推定した結果、約 90% が概念推定に成功した。

6. おわりに

本稿では、学術情報検索における未知語の実態を定量的に把握することを目的とし、学術論文のキーワードから、未知語の実例を収集し、分析した。また、収集した未知語のうち、特に、未知複合語に着目し、その概念を語内構造から推定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] H. Fujisaki, H. Kameda, S. Ohno, T. Ito, K. Tajima and K. Abe : “An intelligent system for information retrieval over the Internet through spoken dialogue,” *PROCEEDINGS of EUROSPEECH'97*, vol. 3, pp. 1675-1678 (1997).

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