学術情報検索における未知語処理

1U－7

阿部 賢司　鈴木 匡芳　飯島 岐勇　片見 憲次　大野 澄雄　藤崎 博也
東京理科大学

1. はじめに

キーワードによる従来の情報検索では、語の表記のみに着目して処理するため、異表記同義・同表記異義の存在が検索性能の低下をもたらす [1]。これを避けるためには、キーワードの概念にまで遡って検索するキー概念検索方式が有効であるが [2]、キーワードがシステムの辞書に登録されていない未知語 [3] の場合には、その概念を適切に推定する必要がある。

このような観点から、我々は、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多数収集した [4]。本稿では、収集した未知語を分類し、その中から特に複合語の未知語に着目して、その概念を語の表層構造および深層構造から推定する方法について検討した結果を述べる。

2. 未知語の収集・分類

学術情報における未知語を収集するため、学術情報センター電子図書館サービス [5] によりテキストデータとして提供される論文概要 5,425 件 (1998 年 1 月時点) に記載されている日本語キーワード (英略語も含む) 延べ 10,006 語の中から未知語を抽出した。その結果、全キーワードの 58.3%(5,829 語) 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 (日本語における表記の多様性に起因するもので、漢字の違いによるもの、送りがなの付け方の違いによるもの、外来語のカタカナ表記の違いによるものなど)、 **(2)** 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もの (造語された複合語)、 **(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 (人名やカタカナ表記の学術用語など)、 の 3 つに大別することができた。本稿では、それらを第 1 種の未知語、第 2 種の未知語、第 3 種の未知語 [3] と呼ぶこととする。 また、これらの未知語以外にも、“語の表記は登録されているが、それに対応する概念が登録されていないもの” (例：“WWW” ⇒ 文書上の概念は “World Wide Web” だが、辞書には “World Weather Watch programme” の概念でしか登録されていない) があった。これは、辞書の信頼性が低いことに起因するものであり、その分野に特化した専門辞書を用いることにより対処し得るが、専門用語や略語の全てを把握するのは事実上不可能であり、この種の語が比較的頻繁に出現する学術情報検索では、これを未知語として取り扱う必要がある。本稿ではこの種の語を第 0 種の未知語と呼ぶこととする。

収集した未知語を上記の分類にしたがって集計した結果が表 1 であり、第 2 種の未知語の割合が圧倒的に多い。このことから、第 2 種の未知語を処理する必要性が最も高いといえる。なお、第 0 種の未知語の異なり単語数の内訳は、英略語が 18 語、カタカナ表記の外来語が 7 語となっている。

表 1 収集した未知語の分類

未知語の種類	異なり単語数 [語]	延べ単語数 [語]
第 0 種	25 (0.5%)	40 (0.7%)
第 1 種	12 (0.2%)	20 (0.3%)
第 2 種	3,951 (81.9%)	4,659 (80.0%)
第 3 種	842 (17.4%)	1,110 (19.0%)

3. 未知語の処理方法の概要

第 0 種の未知語は、文書上の概念と辞書上の概念とが一致するか否かを文脈および共起情報などから推察し、一致しない場合にはシステム管理者に提示して新規登録を依頼することにより処理する。また、第 1 種の未知語は、規則により多様な表記を生成し、辞書上の表記を推定することにより処理する。さらに、第 3 種の未知語は、文脈からおおよその概念を推定し、必要に応じてシステム管理者に新規登録を依頼することにより処理する。

一方、第 2 種の未知語に関しては、日常的に造語されるため数が多く、その全てをシステムに予め登録しておくことは現実的でないこと、また、語を構成する各要素はシステムにとって既知であることから、システムが各要素の概念にもとづいて語全体の概念を自動的に推定することが望ましい。したがって、以下では特に第 2 種の未知語をとりあげ、各構成要素の概念および語の表層構造・深層構造から語全体の概念を推定する方法について検討する。

Processing of unknown words in academic information retrieval
Kenji Abe, Masayoshi Suzuki, Michio Iijima, Kenji Katami, Sumio Ohno, and Hiroya Fujisaki
Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

3.1 第2種の未知語の概念推定

収集した第2種の未知語の構成要素数に着目した場合の分布を表2に示す。分布をみると、要素数が2の場合が最も多く、以降、要素数の増加に伴い減少する傾向を示している。したがって本稿では、出現率が高く、また、処理が比較的簡単であるという2つの理由から、未知語の構成要素数が2の場合の処理について検討することとする。なお、要素数が3以上の場合には、語内要素間の統語構造を階層的に分析することにより処理できるが、これについての検討は機会を改めて報告することとする。

表2 構成要素数に着目した第2種の未知語の分布

要素数	2	3	4	5	6以上
異なり	69.1%	24.2%	5.0%	1.2%	0.5%
延べ	71.0%	22.7%	4.9%	1.0%	0.4%

[概念推定の方略]

第2種の未知語の概念を推定するためには、語の構造を表層、深層の両面から分析する必要がある。本研究では、まず、語構成要素の表層レベルのカテゴリとして、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つを設定し、これらの要素の組合せ(以下、「語構成パターン」と呼ぶ)を分析することにより、語の表層構造を推定する。次に、各要素の深層格(格フレーム)を参照し、一般の複合語では(最末尾の要素が付属的要素でない限り)先行する要素が最末尾の要素の概念を限定・修飾する形となることを参考にして、語の深層構造を推定する。

例えば、「安全確保」の概念を推定する場合、まず、表層構造として "N+V" と "ADV+V" の2つの語構成パターンに適合する。次に、「安全」の深層格が "対象格" あるいは "方法格" となり得ること、また、「確保」の格フレームが "対象格 → ～を" あるいは "方法格 → ～な方法で" であることから、「安全を確保すること」あるいは「安全に確保すること」の2つの概念が推定される。

[表層構造および深層構造の分析例]

上記の方略を具体化するためには、表層構造および深層構造の実体を定量的に把握する必要がある。したがって、本研究では、収集した第2種の未知語のうち語構成要素が2つのものを分析し、語構成パターンの種類およびその出現率を求めた。その結果を表3に示す。また、表3の語構成パターンのうち最も出現率の高かったN+N型の未知語を分析し、深層構造の格フレームの種類およびその出現率を求めた結果を表4に示す。

表3 語構成パターンの種類と出現頻度

語構成パターン	出現率	例
N + N	38.6%	「地域」+「情報」
N + V	24.8%	「画像」+「解析」
V + N	12.1%	「移動」+「能力」
ADJ + N	12.1%	「重要」+「 文 」
N + AFF	5.2%	「高速」+「 化 」
ADV + V	2.8%	「過剰」+「減衰」
V + V	2.5%	「移動」+「受信」
AFF + N	1.0%	「 非 」+「対称」
V + AFF	0.8%	「解析」+「 的 」
AFF + V	0.1%	「 不 」+「検出」

表4 格フレームの種類と出現頻度(N+Nの場合)

格フレーム	出現率
説明格 → ～に関する	22.8%
所有格 → ～のもつ	20.2%
手段格 → ～による	16.3%
目的格 → ～のための	13.8%
主体格 → ～の(そのものの)	12.8%
道具,材料格 → ～を用いた	5.3%
場所格 → ～における	4.7%
存在格 → ～にある(いる)	3.1%
名称格 → ～という	1.0%

4. おわりに

本稿では、学術情報検索における未知語の実例を収集・分類し、特に、第2種の未知語について、その概念を語の表層構造および深層構造から推定する手法について述べた。

参考文献

[1] 劉 軼, 戸井田 和重, 八杉 大輔, 阿部 賢司, 大野 澄雄, 藤崎 博也, 久保村 千明, 亀田 弘之: "学術情報検索における異表記同義・同表記異義の分類・分析および処理," 言語処理学会第4回年次大会発表論文集, pp. 108–111 (1998).

[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," *Proceeding of Eurospeech'97*, vol. 3, pp. 1675-1678 (1997).

[3] 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).

[4] 劉 軼, 大野澄雄, 亀田弘之, 藤崎博也: "学術情報検索における未知語の分類とその処理," 情報処理学会第57回全国大会講演論文集, vol. 3, pp. 219-220 (1998).

[5] http://els.nacsis.ac.jp/nacsis-els-j.html.