学術情報検索における未知語の分類とその処理

1V－10

劉 軼[1] 大野 澄雄[1] 亀田 弘之[2] 藤崎 博也[1]

[1]東京理科大学 [2]東京工科大学

1. はじめに

計算機技術の急速な発展に伴い、厖大な量の情報が生成・蓄積・伝送・公開されるようになったが、その反面、真に必要なものを迅速かつ的確に抽出することのできる情報検索システムはない。このような観点から、我々は「キー概念」[1]を用いた検索方式を提案した[2]。これは、従来のキーワード検索が語の表記上の一致/不一致によって検索を行うのに対して、語の持つ概念による検索を行うことにより、検索もれや不要な検索を低減するものである。

しかしながら、この情報検索システムにおいてもなお、ユーザのシステムへの検索要求の中、あるいは検索対象とするデータベース内に、システムの辞書に登録されていない語が含まれる場合には、それらの概念を適切に推定する必要がある。本報告では、このような状況に鑑み、学術情報検索における未知語の実例を収集して、それらの分類とその処理について検討した結果を述べる。

2. 未知語の定義・分類[3]

我々人間は、語に対する概念を利用する上で、初見(未知)の語であっても、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。

一方、コンピュータでは辞書と、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、システムの辞書に予め載っていない語は、システムにとっては未知となる。以下ではこの「コンピュータにとっての未知語」すなわち未登録単語を単に未知語と呼ぶこととする。

また、未知語には、大きく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。これらはそれぞれ以下のように定義される。

第一種の未知語　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語のこと。これは日本語における表記の多様性に起因する未知語である。第一種の未知語は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等、さらに細分類することができる[4]が、学術情報検索においては、外来語のカタカナ表記の違いにより生じる未知語(例えば、「コンピュータ」と「コンピューター」等)が比較的多く現われる。

第二種の未知語　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のこと。造語された複合語がこの種の未知語となり得る。

第三種の未知語　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。学術情報におけるこの種の未知語としては、人名やカタカナ表記の学術用語などがある。

学術情報検索を行う場合、第一種の未知語に対しては、規則により多様な表記を生成しキーワードに追加することにより対処することができ、また、第三種の未知語に対しては、究極的に表記のみに基づく検索を行うことにより、何らかの情報を引き出すことができる。

一方、日本語においては、必要に応じてさまざまな複合語が日常的に造語され利用される。その多くは第二種の未知語であり、その構成要素それぞれが既知であるため、その意味を推定することは不可能ではない。特に学術情報にはこの種の未知語が多く現われるため、これらの意味・概念を推定することにより検索の柔軟性を高めることが可能である。以下では第二種の未知語について詳細な検討を行う。

3. 学術情報における未知語の収集・分析

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5425件(1998年1月時点)を利用した。各テキストデータには、論文タイトル、著者名、所属、概要、キーワードなどが含まれるが、本研究では日本語で表記された論文タイトル、概要、キーワードのみを利用した。

未知語の抽出には、形態素解析ツール「茶筅」[6]を用いて分析を行い、２つ以上の連続する名詞の列を１つの複合語として取り出した。以下ではこのようにして抽出した未知語のうち、漢字４文字からなる第二種の未知語(約6,500語)に着目し分析を行った結果について述べる。

Classification and Processing of Unknown Words in Academic Information Retrieval

Yi Liu[1], Sumio Ohno[1], Hiroyuki Kameda[2] and Hiroya Fujisaki[1]

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

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

第二種の未知語の構成要素は既知の形態素であり、その形態素間の関係を従来の文法(文文法)を参考にして表層構造の関係として整理した。そのためにまず、構成要素である形態素を、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)の各カテゴリに分類した。付属的要素とは、それ自体では単語としての機能を持たず、他の形態素に付属することにより、正否、肯定、否定、程度、状態、範囲、省略等の意味を表す要素のことをいい、接頭辞、接尾辞もこれに含まれる。また、複合語が３つ以上の構成要素からなる場合には、それらの間の表層構造としての関係は階層的なものとなり得るが、ここではその最上位の構造のみに着目することにした。上述のデータから抽出した未知語に関して構成要素の組合せ(以下、「語構成パターン」と呼ぶ)には以下のものがあった。

a) N ＋ N : 例「信号」＋「帯域」
b) N ＋ V : 例「軌道」＋「計算」
c) N ＋ AFF : 例「時間軸」＋「上」
d) V ＋ N : 例「圧縮」＋「技術」
e) V ＋ V : 例「圧縮」＋「保存」
f) V ＋ ADJ : 例「推論」＋「可能」
g) V ＋ AFF : 例「網設計」＋「上」
h) ADJ ＋ N : 例「動的」＋「経路」
i) ADV ＋ V : 例「単調」＋「増加」
j) AFF ＋ N : 例「一」＋「構成法」
k) AFF ＋ V : 例「不」＋「活性化」
l) AFF ＋ ADJ : 例「非」＋「決定的」

また、上記の語構成パターン以外に最上位の構造が３つ以上の要素からなる例として、「極」＋「零」＋「配置」や「雨」＋「雪」＋「判別」等があった。

4. 未知語処理の方略

第二種の未知語の意味を推定するには、最末尾の語構成要素の意味が大きく関わっていることが多い。つまり、最末尾の語構成要素の意味を、それ以外の語構成要素が限定・修飾する形となっている。ここでは先の分類でも大きな割合で現われた最末尾の語構成要素が動詞的要素の場合を例にとって、その未知の複合語の意味を推定する方略について述べる。

動詞的要素とそれに先行する要素との意味的関係に着目することにより、語全体の適切な意味を推定する。この意味的関係は格文法の意味での深層構造を表すものであり、以下では、動詞的要素のもつ意味および格フレームの情報を「動詞パターン」と呼ぶこととする。また、意味推定には、1) 語構成要素辞書(語構成要素のカテゴリおよび名詞的要素には深層格も記述)、2) 語構成パターン、3) 動詞パターン、の知識を利用する。以下に第二種の未知語の語構成パターンおよび意味推定の手順を示す。

手順１ 未知語候補を入力として受け取る。
手順２ 未知語候補を２つ以上の文字列に分解する。
手順３ 分解された文字列がそれぞれ語構成要素辞書に記載された形態素であるかを調べる。すべての文字列が辞書に記載されている場合、入力は第二種の未知語であり、**手順５**へ。
手順４ **手順２**へ戻り、別の分割の可能性を調べる。別の可能性がない場合、ここでの未知語処理を断念する。
手順５ 検出した語構成要素列が、語構成パターンに合致するか調べ、合致した場合、その語構成パターンを採用し、**手順７**へ。
手順６ 合致しなければ**手順３**に戻り、他の辞書項目について可能性を調べる。
手順７ 最末尾の語構成要素が動詞的要素である場合、その動詞パターンを参照して、未知語の意味を推定する。

例えば、「安全確保」の意味を推定する場合、動詞的要素である「確保」の持つ動詞パターンは、

意味: 確実に保持する
格フレーム: 対象格 → ～を
方法格 → ～な方法で

となる。「安全」には、N、ADJ、ADVの３つのカテゴリが存在するが、ADJは語構成パターンと合致しないため、結局、「安全を確保すること」と「安全に確保すること」の２つの意味が推定される。

5. おわりに

本報では、学術情報検索における未知語の実例を収集・分類し、それを処理するための具体的な方法として、特に動詞的要素を中心とした深層格構造から未知の複合語の意味を推定する手順について述べた。

参考文献

[1] 亀田弘之, 藤崎博也: "テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム," 情報処理学会論文誌, vol. 28, no. 11, pp. 1103–1111 (1987).

[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] 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 軼他: "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).

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

[6] http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html.