対話による支援を考慮した未知語の概念推定

阿部 賢司[1] 鈴木 匡芳[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

我々人間は、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概念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うため、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登録されていない語は、システムにとって未知の語(以降、これを単に未知語[1]と呼ぶ)となり、処理精度を低下させる大きな要因となる。そのため、未知語を含む大量のテキストを処理する自然言語処理システムおいては、未知語を自動的に処理する機能を賦与する必要性が極めて高い。

このことは、自然言語処理技術の導入が不可欠な情報検索システムにおいても同様である。我々は先に、ユーザの負担をできるだけ軽減し、また、ユーザの意図に即した検索を高精度かつ効率的に行なうことを目的として、音声対話を主としたマルチモーダルなマン・マシン・インターフェース、および、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をとりいれた新しい情報検索システムを提案したが[2]、この情報検索システムにおいても、ユーザの検索要求の中、あるいは検索対象とするデータベース内に未知語が存在する場合には、その概念を適切に推定する必要がある[3]。本稿は、このような観点から、情報検索(特に学術情報検索)における未知語の概念を、対話による支援を考慮して自動的に推定する手法について検討するものである。

以下、第2節では対話による支援を考慮した未知語処理システムの概要を説明する。次に、第3節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類した結果を述べる。さらに、第4節では特に複合語の未知語に着目し、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. 未知語処理システムの概要

我々が提案した情報検索システムでは、ユーザとシステムとの対話にもとづいてユーザの検索要求を明確にする[2]。したがって、ユーザの検索要求の中に未知語が存在する場合にも、究極的には対話による質問応答によりその概念を特定することができる(ただし、ユーザの負担を軽減するため、可能な限りシステムが概念を自動的に推定することが望ましい)。一方、検索対象のデータベース中に未知語が存在する場合には、その全てを対話によって処理するのは現実的ではないため、基本的には、システムが辞書的知識や語内外の統語構造に関する知識にもとづいてその概念を自動的に推定する必要がある。しかし、既存の知識のみでは処理できない場合や、言語表現のもつ多様性のため複数の結果が得られる場合には、必要に応じてシステムの処理過程(処理結果)をユーザに提示し、ユーザの判断を処理結果に反映することも重要となる。このような観点から、我々は、ユーザとシステムとの対話による支援を考慮した未知語処理システムを提案する。システムの概略を図1に示す。

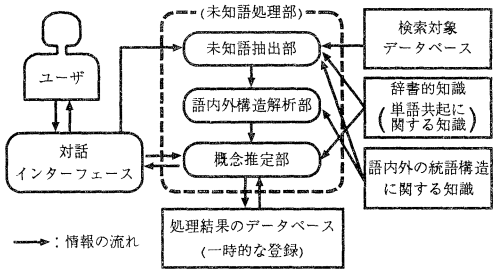

図1. 未知語処理システムの概略図

このシステムは、主として対話インターフェースと未知語処理部から構成される。未知語処理部は未知語抽出部[1]・語内外構造解析部・概念推定部の3つの部分により構成され、対話を介して入力されるユーザの検索要求、および、検索対象のデータベースの内容を、辞書的知識(単語共起に関する知識を含む)・語内外の統語構造に関する知識・過去の処理データにもとづいて処理する。ここで、既存の知識では処理できない場合や複数の処理結果が得られた場合には、必要に応じて処理過程や処理結果を対話インターフェースを通してユーザに提示し、ユーザの判断を加味する。図2は、処理結果をユーザに提示する際の例を示したものであり、処理結果の第一候補が予め選択されている。もし、ユーザがその結果に異議がない場合には、選択されている項目がそのまま処理結果として登録される。また、ユーザが未知語の同義語を入力することにより、未知語と辞書上の同義語との関連付けができるようになっている。処理結果は、

[1] 未知語を処理するためには未知語の抽出処理が必須となるため、このシステムでは、未知語の有無にかかわらず、すべての入力を未知語抽出部で処理する。本研究では、この未知語抽出部も未知語処理部の一部とする。

一時的な記憶領域に登録され、その結果を複数回採用しても矛盾が生じない、あるいは、ユーザに提示したとき異議を唱えられないなど、その信頼性が十分認められた場合には正規の単語辞書に登録される。

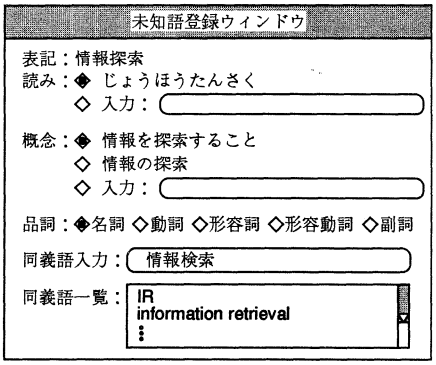

図 2. 概念推定結果の提示例

なお、このシステムでは、ユーザの部分をシステム管理者に置き換えて利用することもできる。

3. 未知語の収集・分類

提案した未知語処理システムを具体化するためには、未知語の実態を定量的に把握する必要がある。このような観点から、我々は、学術情報センター電子図書館サービス [4] 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) を利用して未知語の実例を収集した。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、本研究ではキーワードのみを利用し、記載されている日本語キーワード (英略語も含む)7,091 語 (延べ 10,006 語) の中からシステムの辞書[2] に登録されていない語を未知語として抽出した。その結果、全キーワードのうち、異なりで 68.1%(4,830 語)、延べで 58.3%(5,829 語) が未知語となった。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 語であった。

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%)

ここで、各種の未知語の処理方法について検討すると、第 0 種の未知語に関しては、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ユーザ (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ことができる。また、第 1 種の未知語に関しては、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ことができる。さらに、第 3 種の未知語に関しては、文脈からおおよその概念を推定し、必要に応じてユーザ (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ことができる。一般に、第 1 種と第 3 種の未知語に関しては、機械による概念推定が比較的困難であるため対話による支援を多く必要とするが、出現率が低いこと、また、新たに造語されることが比較的少ないことから、対話によるユーザへ負担も比較的小さいといえる。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、その全てを対話による支援に依存させる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く、その必要性は他と比べるととりわけ高い。したがって、次節では特に第 2 種の未知語をとりあげ、その概念を各構成要素の概念および語の表層構造・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] 本研究では、システムの辞書として EDR 電子化辞書 [5] 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) を用いた。

4. 第2種の未知語の処理方法

システムの未知語処理部では、未知語の抽出、語内外の統語構造解析、概念推定の順に処理が進められる (第2節)。以下に、第2種の未知語を処理する場合の具体的な方略を説明する。

4.1 未知語の抽出

前節で収集した未知語の実例は、単語として掲載されているキーワードから抽出したものであるが、実際には、ユーザの検索要求や論文の概要といったべた書きの文章から抽出する必要がある。本研究では、以下の手順でべた書きの文章から未知語を抽出する。なお、抽出に関しては、第2種以外の未知語に関しても同様に処理することができる。

(手順1) 入力文に対して形態素解析[3]を行なう。この際、数字列、アルファベット列は1単位 (名詞) とみなす。また、形態素解析に失敗する場合には、解析が中断するポイントに未知語が存在すると考え、そのポイントから n 文字を未知語 (名詞) とみなして解析を継続する。ここで、n の値は初め1とし、解析の継続が可能となるまで値を1ずつ増やす。

(手順2) 形態素解析の結果において、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形態素 (1単語) とする。

(手順3) 区切られた単語のうち、システムの辞書に登録されていないものを未知語として抽出する。

ここで、第3節で述べた学術論文のテキストデータから任意にとりだした日本語概要50件を入力とし、上記の手順にしたがって処理した結果、99.3%の精度で未知語の抽出に成功した。なお、抽出に失敗した例の全てが、読点の不足により過接続したものであったが (例：一方2次元FIRディジタルフィルタ ⇐ 本来は “一方” と “2次元” との間に読点 “、” が入る)、接続が不自然な部分は、次の処理 (語内の統語解析) により検出することができる。

4.2 語内の統語解析

語内の統語構造は表層・深層の両面から解析する。まず、表層レベルでは、語構成要素の表層カテゴリとして、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。つぎに、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各要素に深層格 (格フレーム) を設定し、それらの格構造を分析することにより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、本研究では、動詞的要素以外のものにも拡張して用いる。

ここで、第二種の未知語の語内構造の実態を定量的に把握するため、第3節で収集した第2種の未知語のうち、語構成要素数が2つのものを分析して、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2に示す。また、表2の語構成パターンのうち、出現率の高かった上位3つのパターン (N+N型, N+V型 V+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それぞれ表3~5に示す。本研究では、これらの結果にもとづいて、各語構成要素に対し、要素の概念・表層構造のカテゴリ・深層格 (格フレーム) の情報を付加した語構成要素辞書を作成し、それを参照することにより、隣接する要素間の語構成パターンおよび格構造を分析する。なお、現段階では、適用範囲を限定した小規模な語構成要素辞書を人手により作成しているが、一般単語辞書から自動的に拡張する方法、および、獲得機能を用いて小規模な辞書を自動的に拡張する方法についても現在検討している。

表2 語構成パターンの種類と出現率

語構成パターン			出現率	例
N	+	N	36.5%	「地域」+「情報」
N	+	V	25.0%	「画像」+「解析」
V	+	N	15.2%	「運動」+「能力」
ADJ	+	N	9.7%	「重要」+「 文 」
ADV	+	V	4.4%	「過剰」+「減衰」
N	+	AFF	3.8%	「高速」+「 化 」
V	+	V	3.4%	「移動」+「受信」
V	+	AFF	1.5%	「解析」+「 的 」
AFF	+	N	0.3%	「 非 」+「対称」
AFF	+	V	0.2%	「 不 」+「検出」

表3 格フレームの種類と出現率 (N+Nの場合)

格フレーム			出現率
主体格	→	の～	34.3%
対象格	→	に関する～	21.3%
仕様格	→	という仕様における～	11.3%
手段格	→	という手段による～	9.4%
所有格	→	のもつ～	4.4%
目的格	→	のための ～	3.1%
条件格	→	という条件における～	3.0%
名称格	→	という名称の～	2.9%
存在格	→	に存在する～	2.6%
材料格	→	という材料を用いた～	2.0%
道具格	→	という道具を用いた～	1.9%
場所格	→	という場所における～	1.5%
状況格	→	という状況における～	1.5%
事象格	→	という事象における～	0.8%

[3] 本研究では最長一致法を用いて形態素解析を行った。

表 4 格フレームの種類と出現率 (N+V の場合)

格フレーム	出現率
対象格 → を～すること	66.4%
手段格 → という手段で～すること	11.4%
動作主格 → が～すること	6.7%
仕様格 → という仕様において～すること	4.1%
道具格 → という道具を用いて～すること	2.7%
場所格 → という場所で～すること	2.6%
条件格 → という条件で～すること	1.8%
状況格 → の場合に～すること	1.7%
材料格 → という材料を用いて～すること	0.9%
目的格 → という目的で～すること	0.5%
方向格 → へ～すること	0.5%
事象格 → という事象において～すること	0.3%
時間格 → の時に～すること	0.2%
位置格 → に～すること	0.2%

表 5 格フレームの種類と出現率 (V+N の場合)

格フレーム	出現率
目的格 → するための～	41.8%
状態格 → した～	35.2%
動作格 → する～	23.0%

上記の結果は、語構成要素が 2 つの場合の構造を示しているが、要素数が 3 以上の場合には、語内要素間の統語構造をこれらの結果にもとづいて階層的に分析することにより処理する。例として、“音声自動認識”(要素数 3) の分析手順を以下に示す。

(手順 1) 語を要素単位に区切る (音声/自動/認識)。
(手順 2) 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、“音声/自動” と “認識” との関係を表層・深層の両面から分析する。この際、上記の理由から、“音声/自動” の概念は最末尾の要素の “自動” が支配していると仮定し、実際には、“自動” と “認識” との関係を分析する。その結果、表層レベルで語構成パターン “ADV+V” と適合し、深層レベルで格フレーム “状況修飾格 → (的に) ～すること” に当てはまる。
(手順 3) つぎに、“音声/自動” に着目し、“音声” と “自動” との関係を分析する。その結果、表層レベルで “N+ADV” という語構成パターンは存在しないため、“音声” と “自動” は構造的に接続しないことが判明する。(もし、“音声” と “自動” が接続することが可能な場合には、“((音声/自動)/認識)” という解が出力されて処理が終了する。)
(手順 4) 最後に、“音声” と “認識” との関係を分析する。その結果、表層レベルで語構成パターン “N+V” と適合し、深層レベルで格フレーム “対象格 → を～すること” に当てはまる。以上の結果から “(音声/(自動/認識))” という解が生成される。

ここで、第 3 節で収集した第 2 種の未知語の構成要素数に着目した分布は表 6 のようになっており、要素数が 2 および 3 の場合の処理を検討することにより全体の約 9 割が処理できることを示している。

表 6 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%

なお、未知語の概念推定のためには、語外の統語構造解析も重要となるが、これに関しては、従来の一般的な統語解析手法を利用することができるため、具体的な処理方法に関しては省略する。

4.3 概念推定

第 2 種の未知語の概念は、語内の統語構造の解析結果から推定することができる。例えば、先に示した “音声自動認識” の場合には、語内の深層構造の解析結果から、“音声を自動的に認識すること” という推定結果が得られる。しかし、語内の深層構造の解析結果のみから概念を推定するのでは一般に不十分であるため、本研究では以下の結果も加味する。

(1) 語外の統語構造解析により得られる、文脈における未知語の意味的役割や品詞の推定結果。
(2) 過去に処理した未知語のうち、語内構造が類似するものの処理結果。
(3) 対話処理によって提示されるユーザ (システム管理者) の判定結果。

なお、ユーザの負担を最小限に抑えるため、(3) の結果を得るための対話処理は、既存の知識では処理できない場合や複数の処理結果が得られた場合など、その処理が必要な場合にのみ行う。

5. おわりに

本稿では、対話による支援を考慮した未知語処理システムを提案した。また、処理の必要性が最も高い第 2 種の未知語に着目し、その具体的な処理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,” *Proceeding of Eurospeech'97*, vol. 3, pp. 1675-1678 (1997).

[3] 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).

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

[5] 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).