キー概念に基づく情報検索方式の高度化(1)

5Y－4

―キーワードの異表記同義の処理―

藤崎 博也[1] 大野 澄雄[1] 亀田 弘之[2] 阿部 賢司[1] 劉 軼[1] 戸井田 和重[1] 八杉 大輔[1]

[1] 東京理科大学 [2] 東京工科大学

1. はじめに

通常の情報検索においては、多くの場合、ユーザは検索すべき対象を当初から明確に意識しているわけではなく、また、十分な知識を持ち合わせているとは限らない。人間が介入するデータベースの検索においては、検索の専門家(サーチャー)とユーザとの音声による対話を通じて検索要求を明確化させ、迅速かつ的確な検索を行なうことを助けている。したがって、機械による情報検索システムにおいても、ユーザとの音声対話によりその検索要求を明確にすることが望ましいと考えられる。また、キーワードによる従来の情報検索では、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。これを避けるには、キー概念を用いることが有効であるが[1]、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[2],[3]の場合には、その概念推定が必要となる。また、検索効率を向上させるための様々な知識はシステムが自動的に獲得する必要があり、さらに、様々な処理を自律的かつ協調的に遂行するためには、複数のエージェントを導入する必要がある。このような見地から、我々は、音声対話・キー概念検索・未知語処理・知識獲得・エージェント技術を組み合わせた、新しい情報検索システムを提案した[4]～[7]。

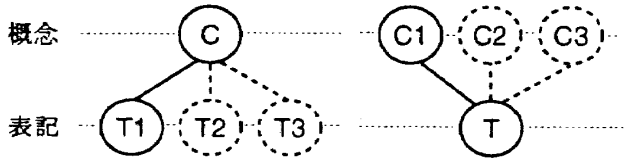
本報では、このシステムを具体化するうえで重要な要素の1つとなるキー概念検索をとりあげ、異表記同義を具体的に処理する方法について検討する。

2. 情報検索における異表記同義・同表記異義

本報では、1つの語は、1つの表記と1つの概念から構成されるものと考える。ここで語の表記とは、文字言語の場合には文字を、音声言語の場合には音声を意味するものとする。ただし、ここでは、文字言語の場合について議論する。このとき、表記の異なる複数の語が概念のレベルで1つに縮退する場合があり、これを異表記同義の現象と呼ぶ。また、概念の異なる複数の語が表記のレベルで1つに縮退する場合を同表記異義の現象と呼ぶ。異表記同義・同表記異義が存在する場合の表記と概念との関係を図1に示す。

従来のキーワード検索では、語の表記のみに着目するため、キーワードに異表記同義が存在する場合には、ユーザが呈示した表記T1では、概念Cに関わるもののうち、表記T2, T3のキーワードで表されるものは抽出できない(図1(a))。また、キーワードに同表記異義が存在する場合には、ユーザが呈示した表記Tを手がかりとして概念C1に関するものを求めようとすると、概念C2, C3に関する不要なものまで抽出してしまう(図1(b))。すなわち、異表記同義の存在は検索洩れをもたらし、同表記異義の存在は不要な検索をもたらす。

これらを避けるには、キー概念のレベルにまで遡った検索が必要である。



(a) 異表記同義　検索洩れをもたらす

(b) 同表記異義　不要な検索をもたらす

図1. 異表記同義・同表記異義が存在する場合の表記と概念との関係

3. 異表記同義の分析・分類

キー概念に基づく検索方式を具体化するため、学術情報検索における異表記同義の実例を収集・分析し、その結果に基づいてそれらを分類すると、(1)表記の多様性によるもの、(2)辞書的な概念の一致によるもの、(3)事実上の概念が一致するもの、の3つに大別することができ、それらは、以下のように、さら

An advanced information retrieval system based on key concepts (1)
— Processing of synonymity in keywords —
Hiroya Fujisaki[1], Sumio Ohno[1], Hiroyuki Kameda[2], Kenji Abe[1], Yi Liu[1], Kazushige Toida[1], and Daisuke Yasugi[1]
[1]Science University of Tokyo, 2641 Yamazaki, Noda, 278
[2]Tokyo Engineering University, 1404-1 Katakura, Hachioji, 192

に細分化することができる。

(1) 表記の多様性による異表記同義

(1.a) 日本語の送りがなの多様性よるもの

例. 引数 と 引き数

(1.b) 外来語の表記の多様性によるもの

例. アーキテクチャ と アーキテクチャー

(2) 辞書的な概念の一致による異表記同義

(2.a) 異なる単語が概念レベルで一致するもの

例. 本 と 書籍

(2.b) 複数の言語にまたがるもの

例. 情報検索 と information retrieval

(2.c) 略語使用の有無によるもの

例. information retrieval と IR

(2.d) 分野によって表現が異なるもの

例. 電場 と 電界

(3) 事実上の概念が一致する場合の異表記同義

(3.a) 恒常的に概念が一致するもの

例. 補間 と 内挿

(3.b) 時や場所を限定したとき概念が一致するもの

例. 白金 と 触媒

4. 異表記同義の処理方法

(1) の、表記の多様性による異表記同義に関しては、表記の多様性に関する知識をシステムに与えることにより処理できる。

(2) の、辞書的な概念の一致による異表記同義に関しては、あらかじめ用意した表記-概念対応辞書に基づいてキー概念が共通する語をキーワードとして用いることにより、検索洩れを回避することができる。

具体的な例として、学術情報センター電子図書館サービスを利用した情報検索において、互いに異表記同義の関係にある (a)「情報検索」, (b)「information retrieval」, (c)「IR」をキーワードとして検索した結果を示すと以下のようになり、(a) のみで検索するよりも、(b), (c) を追加して検索する方が検索洩れが軽減される。

(a)「情報検索」で検索した結果

抽出件数：13 件

適合件数 ： 13 件

(b)「information retrieval」で検索した結果

抽出件数：19 件 ((a) と重複するもの：7 件)

適合件数 ： 19 件

(c)「IR」で検索した結果

抽出件数：6 件 ((a) と重複するもの：1 件)

適合件数 ： 1 件

一方、(c) で見られるような不要な検索が行われる場合もある。これは、同表記異義によってもたらされたものであり、「IR : infrared (赤外線)」に関する情報を抽出している。

(3) の、事実上の概念が一致する場合の異表記同義に関しては、辞書的な概念が一致しないため、表記-概念対応辞書に基づく方法を用いるためには、表記と事実上の概念との対応を示した知識をシステムに与える必要がある。また、(3.b) のように、時や場所の状況が重要になる場合もあり、抽出する情報の内容を概念レベルで把握する必要がある。

本報では、キー概念検索方式を具体化するために、異表記同義の実例を収集・分析・分類し、それを処理するための具体的な方法を検討した。なお、同表記異義に対する具体的な処理方法は、現在検討中である。

参考文献

[1] 藤崎博也, 亀田弘之, 河井恒 : "新聞記事情報の階層構造に基づく記事分類・検索システム," 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[2] 亀田弘之, 藤崎博也, 森田敏生, 倉島顕尚 : "未知語の分類とその処理に関する考察," 情報処理学会第 36 回全国大会講演論文集, 5T-5, pp. 1195-1196 (1988).

[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] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 伊東卓哉, 佐久間聖二 : "キー概念の抽出と未知語の処理に基づく情報検索方式の高度化," 情報処理学会第 54 回全国大会講演論文集, vol. 3, pp. 23-24 (1997).

[5] 藤崎博也, 亀田弘之, 田島 研, 大野澄雄 : "対話による高度情報検索システムの構築," 言語処理学会第 3 回年次大会発表論文集, pp. 261-264 (1997).

[6] 藤崎博也, 大野澄雄, 伊東卓哉, 阿部賢司, 佐久間聖二, 亀田弘之 : "知的エージェントを用いるインターネット上の情報検索システム," 電子情報通信学会総合大会講演論文集, p. 186 (1997).

[7] 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).