キー概念に基づく情報検索システム方式の高度化(2)

3V－5

―キーワードの同表記異義の処理―

藤崎博也　大野澄雄　阿部賢司　片見憲次　飯島岐勇　鈴木匡芳

東京理科大学

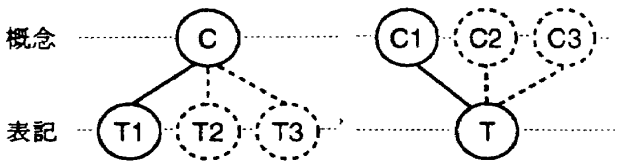
1. はじめに

キーワードによる従来の情報検索では、表記のみに着目して処理するため、異表記同義の存在による検索洩れや、同表記異義の存在による不要な検索が避けられない。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が [1] ～ [4]、異表記同義への対処の方法は既に提案した [5]。本報では、キーワードの同表記異義の例を収集・分析・分類し、その処理の方法ついて検討した結果を述べる。

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。

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検索洩れが生じる (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不要な検索が生じる (図 1(b))。



(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. 同表記異義の例の収集・分析・分類

キー概念に基づく検索方式を具体化するため、学術情報センター (http://els.nacsis.ac.jp/nacsis-els-j.html) に公開されているデータベース 5425 件 (1998 年 1 月現在、以下データと記す) における同表記異義の例を収集し、以下のように分析・分類した。

3.1 収集

データ中の全論文のキーワードと、全論文のタイトルに使われている名詞を収集した。そのうち、概念が複数存在するものを同表記異義とし、その判断は EDR の概念辞書によって行った。

キーワードの中で同表記異義を持つものは 168 語、出現回数は延べ 860 回、タイトルの中の名詞では 1364 語、出現回数は延べ 18445 回存在した。

収集した同表記異義の概念を比較した結果、下記の例のように、概念は微妙に異なっているが、実際の検索においては問題にならないような同表記異義は省いた。

例. 電子メール
- 電気通信的手段で送られる郵便
- 郵便で送るような情報を電気通信的手段で送るサービス

3.2 分析・分類

検索の際に問題となる同表記異義は、表記上で 4 種類に分類できる。以下にそれぞれの例を示す。

(1) 略語・記号

例. SAR
- supervisor analysis router
- storage address register

(2) 英単語

例. permutation
- 順列
- 置換

(3) 外来語

例. グラフ
- 関連する 2 つ又は 2 つ以上のものの数量や，関数関係をあらわした“図形”
- 写真を中心にした“雑誌”

An advanced information retrieval system based on key concepts (2)
— Processing of polysemy in keywords —
Hiroya Fujisaki, Sumio Ohno, Kenji Abe, Kenji Katami, Michio Iijima and Masayoshi Suzuki
Science University of Tokyo, 2641 Yamazaki, Noda, 278-8501

(4) 漢字・かな

例．米

・植物の“コメ”

・“アメリカ”という国

・長さの国際単位系基本単位“メートル”

表1にキーワード中の同表記異義の語数と、タイトル中の同表記異義の語数を示す。また、括弧内の数字は出現回数を表す。

表1 同表記異義の分類結果

	キーワード中の同表記異義の語数 (回数)	タイトル中の同表記異義の語数 (回数)
(1) 略語・記号	21 (79)	38 (1662)
(2) 英単語	2 (6)	4 (14)
(3) 外来語	23 (71)	203 (2210)
(4) 漢字・かな	9 (22)	523 (7775)
合計	55 (178)	588 (11651)

4. 同表記異義の簡単な処理方法とその効果

4.1 処理方法

同表記異義の処理方法として、以下のように検討した。

(1) データ内の非省略形の言葉の利用

略語に関してデータ内に非省略形の表記がある場合には、それを利用して特定できる。

例．… the specific absorption rate (SAR) …

(2) 英語表記の利用

日本語では同表記異義を持つ語でも、英語が併記されており、英語で異表記であれば特定できる。

例．米→ rice

(3) 前後の単語を繋げる

キーワードでは切られて短い単語になったことが原因で、同表記異義になっている場合がある。切られる前の長い単語に戻して、どの意味で使用されていたかを調べることにより特定できる。

例．グラフ→無向グラフ

4.2 処理結果

前節の方法で収集した、同表記異義を持つキーワードについて、上記の方法により特定可能となるものを表2に示す。ここで特定可能率は、キーワード中の同表記異義の出現回数に対する、特定可能な同表記異義の出現回数の割合である。

表2 処理結果

	キーワード中の同表記異義の出現回数	特定可能な同表記異義の出現回数	特定可能率 [%]
(1) 略語・記号	79	48	60.8
(2) 英単語	6	3	50.0
(3) 外来語	71	29	40.8
(4) 漢字・かな	22	8	36.4
合計	178	88	49.4

簡単な処理を行うことにより、表2から、問題となる同表記異義をもつキーワードの約半分が特定可能である。なお、上記の方法で特定不可能なものに関しては、語の共起に関する情報の利用が考えられ、現在検討中である。

5. おわりに

本報では、キー概念検索方式を具体化することを目的とし、同表記異義の実例を収集・分析・分類し、それを処理するための方法を検討した。

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] 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).

[3] 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).

[4] 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).

[5] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 劉軼, 戸井田和重, 八杉大輔 : “キー概念に基づく情報検索システム方式の高度化 (1) －キーワードの異表記同義の処理－,” 情報処理学会第56回全国大会講演論文集, vol.3,pp. 128-129 (1998).