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

劉 軼[1] 戸井田 和重[1] 八杉 大輔[1] 阿部 賢司[1] 大野 澄雄[1] 藤崎 博也[1]
[1] 東京理科大学

久保村 千明[2] 亀田 弘之[2]
[2] 東京工科大学

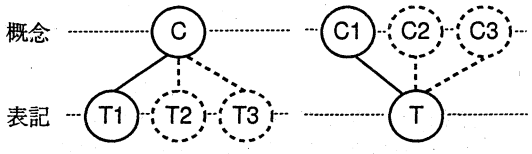
1. はじめに

計算機技術の急速な発展に伴い、膨大な量の情報が伝送・蓄積・公開されるようになり、その中から、真に必要なものを迅速かつ的確に抽出するできるような情報検索システムが必要となっている。しかし、キーワードによる従来の情報検索では、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を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、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、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上で重要なキー概念検索をとりあげ、異表記同義・同表記異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、本稿では、文字言語の場合についてのみ論ずる。表記の異なる複数の語が概念のレベルで1つに縮退する場合を異表記同義の現象と呼び、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こととする。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。



(a) 異表記同義　(b) 同表記異義

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

従来のキーワード検索では、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。以下では学術情報検索における異表記同義及び同表記異義の実例を収集、分析、分類し、それらについての処理方法を検討した結果について述べる。

3. 異表記同義の分析・分類

異表記同義の場合は、(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つに大別され、それらは、以下のようにさらに細分化される。

(1) 表記の多様性による異表記同義

(1.a) 送りがなの多様性によるもの
　例. 引数 / 引き数

(1.b) 漢字仮名混じりによるもの
　例. 勾配 / こう配

(1.c) 漢字表記の多様性によるもの
　例. 函数 / 関数

(1.d) 旧字体と現在表記字体によるもの
　例. 壱 / 一

Classification, Analysis, and Processing of Synonymity and Polysemy in Academic Information Retrieval

Yi Liu [1], Kazushige Toida [1], Daisuke Yasugi[1], Kenji Abe[1], Sumio Ohno[1], Hiroya Fujisaki[1], Chiaki Kubomura[2] and Hiroyuki Kameda[2]
[1]Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510
[2]Tokyo Engineering University, 1404-1 Katakura, Hachioji, 192-8580

(1.e) 音読みと訓読みによるもの

例. 置換 ／ 置き換え

(1.f) 外来語の表記の多様性によるもの

例. アーキテクチャ ／ アーキテクチャー

(1.g) 数字や単位に関する表記の違いによるもの

例1. 5 ／ 五 ／ V　例2. 一個 ／ 1コ

(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 ／ 書籍

(2.b) 接尾語的な変化によるもの

例. 運転手 ／ 運転者

(2.c) 複数の言語にまたがるもの

例. 情報検索 ／ information retrieval

(2.d) 略語使用の有無によるもの

例. information retrieval ／ IR

(2.e) 分野によって表現が異なるもの

例. 電場 ／ 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 ／ 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 ／ 触媒

(3.c) 商品名が普通名詞と同等に扱われるもの

例. サウンドスペクトログラフ ／ ソナグラフ

4. 異表記同義の処理方法

異表記同義の現象は、表記-概念対応辞書を利用することにより対処することができるが、一般に、すべての表記が辞書に登録されているわけではないため、未知語の処理を避けて通ることはできない。

特に、先の (1.f) の外来語の表記は多様性に富んでおり、また、個別的な対応が必要であるとともに、学術情報文献において外来語の出現頻度が高いという特徴がある。

4.1 外来語の異表記

上記の知見から、本研究では外来語表記の処理を重要視し、下記のような文字列置換処理を設定した。

4.1.1 通常型文字と小型文字による異表記

通常型文字ア, イ, ウ, エ, オ, カ, ケ, ツ, ヤ, ユ, ヨ, ワ, の１２文字を小型文字ァ,ィ,ゥ,ェ,ォ,カ,ケ,ッ,ャ,ュ,ョ,ヮ, に置換し、小型文字は通常型文字に置換。

例.「フィルム」と「フイルム」

4.1.2 長音記号の有無による異表記

(1) 長音記号がある場合

1) 長音記号を削除。

例.「フィルター」と「フィルタ」

2) 長音記号を直前の文字の母音に置換。

例.「カール」と「カアル」

3) 長音記号の直前の文字がエ列の場合は、長音記号をイに置換。

例.「スケール」と「スケイル」

4) 長音記号の直前の文字がオ列の場合は、長音記号をウに置換。

例.「ボール」と「ボウル」

5) 長音記号をルに置換。

例.「フォーム」と「フォルム」

(2) 長音記号がない場合

1) 長音記号を挿入。

2) 文字ア, イ, ウ, エ, オ, があり、かつ、その直前の文字の母音と一致する場合、母音を長音記号に置換。

3) 通常型文字イがあり、その直前の文字がエ列の場合は、イを長音記号に置換。

4) 通常型文字ウがあり、その直前の文字がオ列の場合は、ウを長音記号に置換。

5) ルを長音記号に置換。

4.1.3 その他の異表記

特定の文字列を他の文字列に双方向に置換。なお、網羅的な調査によりこの種類の置換規則は現在、41個にまとめられている。

例1.「イタリ（ア）」と「イタリ（ヤ）」

例2.「(ヴァ）イオリン」と「(バ）イオリン」

例3.「テレ（フォ）ン」と「テレ（ホ）ン」

4.2 辞書的な概念の一致による異表記同義の処理

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。

例えば、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は以下のようになり、(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) で見られるように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。

4.3 事実上の概念が一致する場合の異表記同義の処理

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、時や場所が重要な場合もあり、抽出情報を概念レベルで把握する必要がある。

5. 同表記異義の分析・分類

同表記異義は同じ語から派生したが、使用される分野によって意味が異なるもの、と異なる語が表記のレベルで偶然一致するもの、の 2 つに大別され、それらは、以下のようにさらに細分化される。

(1) 同じ語から派生したが、使用される分野によって意味が異なるもの
　(1.a) 漢字表記のもの。　例. 交流
　社会的現象 ／ 電気現象
　(1.b) カタカナ表記のもの。　例. プログラム
　計画 ／ コンピュータの用語
(2) 異なる語が表記のレベルで偶然一致するもの
　(2.a) 略語によるもの。　例. PC
　パーソナルコンピュータ ／ プロピレンカーボネート
　(2.b) 表記が偶然に一致するもの。　例. 工夫
　工場で働く人 (こうふ) ／ 考えついたよい方法 (くふう)

6. 同表記異義の処理方法

ユーザから呈示されたキーワードに同表記異義が存在する場合には、ユーザとの対話に基づき、その概念を特定することができる。しかし、検索対象の論文に含まれるキーワードの概念は、その著者以外の人には特定することができないため、概念レベルでのマッチングは行えない。したがって、論文中のキーワードの概念がユーザの意図と適合するか否かを判定する手法、言い換えれば、ユーザが呈示したキーワードを含む論文がユーザの要求と適合するか否かを判定する手法が必要となる。

本報では、以下の 2 つの手法を検討する。

(1) キーワードの概念と共起する概念の有無を手がかりにする方法。
(2) 論文の分野や学会誌名などを手がかりにする方法。

(1) の方法では、着目するキーワードの近傍に、ユーザが意図する概念と共起する傾向の強い概念を持つキーワードが存在する場合には、検索結果はユーザの要求と適合するものとし、逆に、ユーザが意図しない概念と共起する傾向の強い概念を持つキーワードが存在する場合には、ユーザの要求と適合しないものとする。この方法では、全てのキーワードに対する共起概念の情報をシステムに予め与えておく必要がある。

一方、(2) の方法は、論文の分野や学会誌名などを手がかりとしてキーワードの概念を推定し、検索結果

とユーザの要求との適合度を評価するもので、その手続きは(1)の方法よりもはるかに簡単である。例えば、「交流」をキーワードとして検索した場合、社会的現象としての「交流」、および、電気現象としての「交流」に関する情報が抽出されるが、実際の検索結果(抽出件数29件)において、それぞれの概念で検索されたものの件数を分野および学会誌名に着目して分類すると表1に示す結果が得られた。

表1 「交流」をキーワードとして検索した結果

分野名	学会名	概念	
		社会的現象	電気現象
農学	日本家政学会	22	0
人文科学	日本独文学会	1	0
工学	情報科学技術協会	1	0
	電子情報通信学会	0	5

この表では、分野および学会誌によって「交流」の二つの概念が完全に区別されており、同表記異義を処理するための手がかりとして利用できることを示している。しかし、分野および学会誌名でキーワードの概念を特定できない場合も多いため、(1)の方法と併用する必要がある。

7. おわりに

本報では、キー概念検索方式を具体化するために、異表記同義と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具体的な方法について述べた。

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] 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, 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).

[5] 藤崎博也, 大野澄雄, 伊東卓哉, 阿部賢司, 佐久間聖仁, 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).

[6] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 伊東卓哉, 佐久間聖仁 : "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).

[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).