キー概念検索のための自然言語処理とその評価

阿部 賢司　武田 和也　堀越 修平　藤崎 博也

東京理科大学

1. はじめに

従来のキーワード検索では語の表記のみに着目して検索するため、異表記同義・同表記異義の存在により検索精度が低下する。これを回避するためには、キーワードの概念（キー概念 [1]）のレベルにまで遡って検索することが有効であるが、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を推定することも必要である。

筆者らは既に、キーワード検索方式に代わるキー概念検索方式を提唱し、さらに、それを実現するために、異表記同義・同表記異義・未知語の処理方法について検討したが [4-8]、本稿では、これらの処理を統合して学術情報検索に適用し、その有効性を実験的に検証した結果について述べる。

2. 表記-概念対応辞書と検索対象データベース

キー概念検索を行うためには、キーワードの表記と概念とを対応付けるための辞書が必要となる。この研究では、日本電子化辞書研究所から出版されている EDR 電子化辞書 1.5 版 [9] 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）にもとづいて、表記から概念を参照するための表記概念対応辞書（登録表記数：415223）、および、概念から表記を参照するための概念表記対応辞書（登録概念数：278525）を作成した。この際、語の表記および概念としては、EDR 電子化辞書における “単語見出し”、“概念識別子” をそれぞれ用いた。表記概念対応辞書、および、概念表記対応辞書の登録形式を以下に示す。なお、本稿では、これらの辞書を総称して “表記-概念対応辞書” とよぶこととする。

＜表記概念対応辞書の登録形式＞
単語見出し i：概念識別子 1, ⋯, 概念識別子 m
(例. AD：0e63bf, 0e63c0, 0e63c1, 0f9803)

＜概念表記対応辞書の登録形式＞
概念識別子 j：単語見出し 1, ⋯, 単語見出し n
(例. 2f167c：IR, information retrieval, 情報検索)

また、キー概念検索の有効性を実験的に検証するためには、検索対象となるデータベースが必要となる。この研究では、検索対象を特に学術情報と定め、NACSIS から研究用に公開された情報検索システム評価用テストコレクション 1 (NTCIR-1) [10] に収録されている 339,483 件の学術論文に関する情報（論文タイトル、著者名、学会名、概要、キーワードなど）を検索対象として用いた。

以下、本稿では、表記-概念対応辞書にもとづいて抽出した異表記同義・同表記異義・未知語の処理方法、および、これらの方法を学術情報検索に適用し、その有効性を検証した結果について述べる。

3. 異表記同義の処理方法 [6]

ユーザが提示したキーワードと異表記同義の関係にあるキーワードはユーザの検索意図と合致する場合が多いため、それら全てを検索式に追加しないと検索もれが生じる。したがって、表記-概念対応辞書にもとづいて着目するキーワードに異表記同義の現象が存在するか否かを判断し、存在する場合には、異表記同義の関係にある全てのキーワードを検索式に追加することにより検索もれを回避する。例えば、ユーザが「情報検索」というキーワードを提示した場合、表記-概念対応辞書上では、「information retrieval」と「IR」が異表記同義の関係にあるキーワードと判断される。この場合、「情報検索」という検索キーワード（検索式）を、「情報検索 *or* information retrieval *or* IR」の様に拡張することにより、検索もれを回避する。

異表記同義の現象を定量的に把握するため、前節で述べた NTCIR-1 に収録されている 339,483 件の論文情報の中から、1,000 件分の情報を任意に抽出し、その中の “キーワード” の項目に記載されているキーワード 3,906 語を調査した。その結果、1,198 語が辞書に登録されていない未知語であり、残りの 2,708 語のうち、830 語に異表記同義の現象が存在した。また、これらの語を分析した結果、1 つの概念あたり平均 3.3 個の表記が対応することを確認した。さらに、それらの異表記同義を、異表記となった要因に着目して分類した結果、(1) 表記の多様性によるもの、(2) 略語によるもの、(3) 上記 (1)、(2) 以外で、辞書上の概念が一致したことによるもの、の 3 つに大別することができ、それらは以下の様に細分類することができた。各種異表記同義の出現割合を表 1に示す。

(1) 表記の多様性によるもの
  - (a) 漢字仮名混じり表記：例. 捻りモーメント / ねじりモーメント
  - (b) 片仮名表記：例. コンピュータ / コンピューター
  - (c) 数字表記：例. 3 次元モデル / 三次元モデル
  - (d) アルファベット表記：例. fiber / fibre

(2) 略語によるもの

(e) 片仮名表記： 例. ミリメートル波 / ミリ波

(f) アルファベット表記: 例. ESB / Electron Spin Resonance

(3) 辞書上の概念が一致したことによるもの

(g) 同言語間： 例. 同定 / 識別

(h) 多言語間： 例. 誤り率 / error rate

表 1 各種異表記同義の出現割合

異表記同義の種類	出現割合 [%]
(1) 表記の多様性によるもの	(小計：24.8)
(a) 漢字仮名混じり表記	3.8
(b) 片仮名表記	20.0
(c) 数字表記	0.1
(d) アルファベット表記	0.9
(2) 略語によるもの	(小計：1.4)
(e) 片仮名表記	0.4
(f) アルファベット表記	1.0
(3) 辞書上の概念が一致したことによるもの	(小計：73.8)
(g) 同言語間	43.2
(h) 多言語間	30.6

4. 同表記異義の処理方法 [7]

ユーザが提示したキーワードに複数の概念が対応する場合、一般には、ユーザの検索意図と合致する概念はその中の 1 つだけであり、それ以外の概念でキーワードを用いている文書は、ユーザにとって不要なものである。したがって、同表記異義の関係にある語を、それらの共起情報にもとづいて概念ごとに分離し、さらに、その中からユーザの意図と合致するものを特定することによって不要な検索 (誤検索) を回避する。具体的には、(1) 着目したキーワード K を含む論文の検索空間における位置を、共起情報にもとづいてベクトルで表し、(2) そのベクトルにもとづいて論文間の距離を求め、それを論文間の類似度を表す指標とし、(3) 階層的クラスター分析の手法を用いて、距離が近い論文同士から階層的にリンクさせ、(4) ある距離を閾値としたときのクラスタリングの結果を参照し、ユーザの検索要求との合致度が最も高いクラスターのみを検索することにより、誤検索を回避する。

同表記異義の現象を定量的に把握するため、前節で述べた 3,906 語のキーワードを調査した。その結果、既知キーワード 2,708 語のうち、221 語に同表記異義の現象が存在した。また、これらの語を分析した結果、1 つの表記あたり平均 4.0 個の概念が対応することを確認した。ここで、これらの異表記同義を表記に着目してに分類した結果を以下に示す。また、各種同表記異義の出現割合を表 2に示す。

(1) 漢字表記： 例. 分散
概念 1： variance (of a variable)
概念 2： dispersion (of light, matter)

(2) 平仮名表記： 例. ひずみ
概念 1： deformation of a material
概念 2： distortion of a waveform

(3) 片仮名表記： 例. チャネル
概念 1： communication channel
概念 2： a band of radio waves

(4) アルファベット表記 (英略語)： 例. IR
概念 1： information retrieval
概念 2： infrared

表 2 各種同表記異義の出現割合

同表記異義の種類	出現割合 [%]
(1) 漢字表記	56.6
(2) 平仮名表記	0.4
(3) 片仮名表記	39.8
(4) アルファベット表記 (英略語)	3.2

5. 未知語の処理方法 [8]

キーワードが表記-概念対応辞書に登録されていない、すなわち、未知語の場合には、その概念を推定し、新たに辞書に登録する必要がある。未知語は以下の 4 つ、すなわち、(1) 第 1 種の未知語：表記は辞書に登録されているが、それに対応する概念が登録されていない語、(2) 第 2 種の未知語：概念は辞書に登録されているが、それに対応する表記が登録されていない語、(3) 第 3 種の未知語：各語構成要素は表記・概念ともに辞書に登録されているが、その語全体では登録されていない語 (4) 第 4 種の未知語：語構成要素の一部、あるいは語全体が、表記・概念ともに辞書に登録されていない語、に大別することができるが [2, 8]、この研究では特に、日常的に造語されるため出現確率が高く、処理の必要性が最も高い第 3 種の未知語の処理について検討した。

第 3 種の未知語は、各語構成要素は表記・概念ともに辞書に登録されているため、これらの情報にもとづいて語全体の概念を推定する。具体的には、語の表層構造を表すための要素として名詞的要素、動詞的要素、形容詞的要素、副詞的要素、付属的要素の 5 つのカテゴリを、また、語の深層構造を表すための要素として 30 種類の深層格 (主体格、対象格、目的格、場所格、手段・方法格、存在格、条件格、仕様格、状況格、時間・期間格、状態格、事象格、源泉格、方

向格、材料・道具格、程度格、所有格、動作格、名称格、受け手格、結果格、可能格、使役格、受益格、経験者格、役割格、関係格、範囲格、原因格、対等接続格) を設定し、各語構成要素間の表層レベルおよび深層レベルにおける係り受けを Shift-Reduce パーザを用いて解析することにより、語全体の概念構造を推定する。また、同じ概念構造を持つ既知語が存在する場合には、それと異表記同義の関係にある語として辞書に登録し、存在しない場合には、新しい概念識別子を付加して辞書に登録する。

第 4 節、第 5 節で述べたキーワード 3,906 語のうちの未知語 1,198 語の種類毎の例を以下に示す。また、それらの出現割合を表 3に示す。

(1) 第 1 種の未知語：例. ホール
- 文書上の概念 (辞書に無かった概念)：“正孔”
- 辞書上の概念：“ゴルフでボールを入れるグリーン上の穴”、“大広間”

(2) 第 2 種の未知語：例. ウィルス
(辞書上の表記は「ウイルス」)

(3) 第 3 種の未知語：例. 人間情報工学
(「人間」、「情報工学」は辞書に登録されている)

(4) 第 4 種の未知語：例. ナスダック
(語全体が辞書に登録されていない)

表 3 各種未知語の出現割合

未知語の種類	出現割合 [%]
(1) 第 1 種の未知語	0.5
(2) 第 2 種の未知語	0.2
(3) 第 3 種の未知語	81.9
(4) 第 4 種の未知語	17.4

6. 提案手法の学術情報検索への適用

6.1 異表記同義・同表記異義・未知語の処理を採り入れた情報検索

提案手法を適用した情報検索の流れは図 1の様になり、オンラインで処理する部分とオフラインで処理する部分とに大別される。

オンラインでは、以下の手順で処理を行う。

(1) キーワード抽出：テキスト形式で与えられるユーザの検索要求に対して形態素解析を行い、名詞列をキーワードとして抽出する。この際、予め作成した重要度の低い語のリスト (不要語リスト) を参照し、このリストに登録されていない名詞列のみを抽出する。なお、形態素解析には、日本語形態素解析システム “茶筌 (version 2.0) [11]” を用いた。

(2) 異表記同義の検出：表記-概念対応辞書を参照し、抽出したキーワードと同じ概念を持つ語が存在するか否かを判定する。

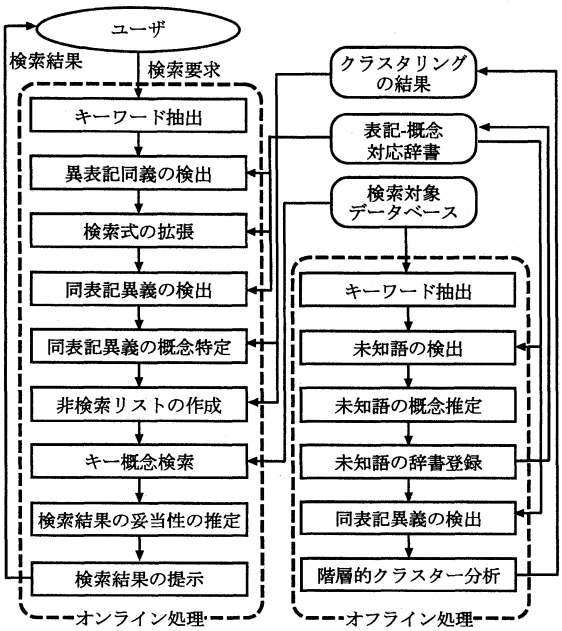

図 1. 提案手法を適用した情報検索

(3) 検索式の拡張：異表記同義が検出された場合には、異表記同義の関係にある全てのキーワードを検索式に追加する。

(4) 同表記異義の検出：表記-概念対応辞書を参照し、検索式の各キーワードの表記に複数の概念が対応するか否かを判定する。

(5) 同表記異義の概念特定：同表記異義が検出された場合には、オフライン処理において予め求めた、同表記異義のキーワードを含む文書のクラスタリングの結果を参照し、検索式との合致度が最も高いクラスターを特定する。

(6) 非検索リストの作成：手順 (5) の結果にもとづいて、不要な文書のリスト (非検索リスト) を作成する。

(7) キー概念検索：検索式および非検索リストにもとづいてキー概念検索を行う。

(8) 検索結果の妥当性の推定：キーワードの出現回数、出現位置、検索式の条件との合致度に着目して検索結果の妥当性 (ユーザの検索意図との合致度) を推定する [12]。

(9) 検索結果の提示：手順 (8) の結果にもとづいて、ユーザの意図との合致度が高い文書からユーザに提示する。

また、オフラインでは、以下の手順で処理を行う。

(a) キーワード抽出：検索対象データベースの各文書に対して形態素解析を行い、キーワードを抽出する (オンライン処理の手順 (1) と同じ)。

(b) 未知語の検出：抽出したキーワードが表記-概念対応辞書に登録されているか否かを判定する。

(c) 未知語の概念推定：未知語が検出された場合には、第5節の方法にしたがって未知語の概念を推定する。

(d) 未知語の辞書登録：手順(c)の結果にもとづいて未知語を表記-概念対応辞書に登録する。

(e) 同表記異義の検出：表記-概念対応辞書にもとづいて、抽出したキーワードから同表記異義を検出する(オンライン処理の手順(4)と同じ)。

(f) 階層的クラスター分析：第4節の方法にしたがって、同表記異義のキーワードを含む文書をクラスタリングする。

6.2 学術情報検索への適用実験

提案手法の有効性を定量的に検証するため、小規模な学術情報検索実験を行った。NTCIR-1から任意に抽出した500件の論文情報を検索対象とし、対象中のキーワードに対する同表記異義の処理(クラスタリング)および未知語の処理(概念推定と辞書登録)は、オフラインで予め行った。

10件の要求に対する検索実験において、(1)提案手法を全く適用しない場合、(2)未知語処理以外の処理を行った場合、(3)未知語処理を行った場合について、異表記同義・同表記異義の処理を単独で行った場合、あるいは組み合わせて行った場合の検索もれ率および誤検索率を表4に示す。(1)の場合と比べると、(2)では、異表記同義の処理により、誤検索率は僅かに増加するものの、検索もれ率は約1/5にまで減少し、また、同表記異義の処理により、検索もれ率は変化しないものの、誤検索率は約1/2にまで減少する。当然のことながら、異表記同義と同表記異義の両方を処理した場合には、両者の利点が得られる。さらに、(3)では、異表記同義の処理により、検索もれ率は(2)場合の1/2、すなわち、(1)場合の1/10にまで減少し、また、同表記異義の処理による効果は(2)の場合と比べて変化しないものの、異表記同義・同表記異義・未知語の全てを処理することにより、検索もれ率は(1)の場合の約1/10にまで、また、誤検索率は(1)の場合の約1/3にまで減少する。

表4 各場合における検索もれ率および誤検索率

	検索もれ率	誤検索率
(1) 提案手法を全く適用しない場合	28.6%	3.8%
(2) 未知語を処理しない場合		
(a) 異表記同義を処理した場合	5.7%	4.3%
(b) 同表記異義を処理した場合	28.6%	2.0%
(c) 異表記同義と同表記異義の両方を処理した場合	5.7%	1.5%
(3) 未知語を処理した場合		
(d) 異表記同義を処理した場合	2.9%	4.2%
(e) 同表記異義を処理した場合	28.6%	2.0%
(f) 異表記同義と同表記異義の両方を処理した場合	2.9%	1.4%

7. おわりに

本稿では、異表記同義・同表記異義・未知語の処理を統合して学術情報検索に適用し、その有効性を実験的に検証した結果について述べ、異表記同義の処理が検索もれの軽減に、同表記異義の処理が誤検索の軽減に、未知語の処理がキー概念検索の高度化に有効であることを示した。なお、現在は、検索実験の規模をさらに拡大して提案手法の有効性を検証している。

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] 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).

[5] K. Abe, M. Iijima, K. Katami, M. Suzuki, K. Kurokawa, K. Taketa, S. Ohno and H. Fujisaki: “Concept-based search and user modeling in information retrieval based on human-machine dialogue,” *Proceedings of RIAO2000*, vol.2, pp.1728-1743 (2000).

[6] 藤崎 博也, 武田 和也, 阿部 賢司, 堀越 修平, 猪股 尚典, 山崎 潤: “キー概念検索のための異表記同義・同表記異義の処理,” 情報処理学会第61回全国大会講演論文集, vol.3, pp.161-162 (2000).

[7] 阿部 賢司, 片見 憲次, 武田 和也, 藤崎 博也: “同表記異義の処理とその情報検索への応用,” 言語処理学会第6回年次大会発表論文集, pp.459-462 (2000).

[8] 鈴木 匡芳, 中村 宏, 阿部 賢司, 藤崎 博也, 亀田 弘之: “既知形態素からなる未知複合語の概念推定とその辞書登録,” 言語処理学会第6回年次大会発表論文集, pp.423-426 (2000).

[9] 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).

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

[11] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸: 日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版 (1999).

[12] 藤崎 博也, 阿部 賢司, 飯島 岐勇, 武田 和也, 大野 澄雄: “検索結果の適合度の定量的評価,” 情報処理学会第60回全国大会講演論文集, vol.3, pp.113-114 (2000).