2N-7

自由語検索のための同義語・異表記展開方式

畠山　敦、川口久光、加藤寛次、藤澤浩道
(株)日立製作所　中央研究所

1．はじめに

従来の文書データベースの検索システムでは、あらかじめ定められたキーワードや分類コードを用いて登録されたインデクスファイルに基づいて検索が行われてきた。しかしながら、データベースの普及に伴い、広く一般のユーザが検索システムを使用するようになるにしたがって、キーワードや分類コードのように統制された言葉ではない自由語（非統制語）を用いて、文書の登録、検索が行えるシステムが必要とされてきている。このような自由語検索では、登録者や検索者の意図により自由に検索語が指定されるために、同義語や表記のゆれ（異表記とよぶ）が検索もれの原因となる。本研究の目的は、上記の同義語や異表記を検索システムの内部で自動的に展開して、検索もれのない検索システムを実現することにある。

2．同義語・異表記展開方式

本方式での自由語検索では、文書検索時にコンソールより入力された検索語に対して同義語展開及び異表記展開を施した後、展開された複数の言葉を新たな検索語として検索を行う。ここでは、検索者が同義語辞書の見出し語以外の異表記を検索語として入力する場合があるため、入力された検索語をまず異表記展開し、次に同義語展開する構成とした。同義語辞書には単一の表記で表現された言葉が登録されているために、さらにこの異表記をも得る必要がある。したがって、同義語展開により新しく得られた言葉についても異表記展開処理を加える方式とした。そして、これらの展開した全ての言葉について、検索処理部で文書データベース（ＤＢ）を検索する。（図１参照）

こうすることにより、登録者と検索者が使用する言葉の違いを検索システムで吸収でき、検索もれを抑えることができる。

2．1　異表記展開処理

異表記展開処理では、文字種毎に対応した異表記に展開する。

・カタカナ文字種の場合
外来語のカタカナ音節表記に関して異表記へ展開する。
例：‘ピアノ’→‘ピヤノ’

・漢字文字種の場合
新旧字体、送りがなに関して異表記へ展開する。
例：‘斎藤’→‘齊藤’

・アルファベット文字種の場合
大文字、小文字に関して異表記へ展開する。
例：‘ＡＬＰＨＡＢＥＴ’→‘Ａｌｐｈａｂｅｔ’

2．2　同義語展開処理

同義語辞書を参照して、異表記展開処理で得られた言葉に対応する同義の言葉へ展開する。
例：‘計算機’→‘コンピュータ’，
‘ＣＯＭＰＵＴＥＲ’

3．今後の課題

・同義語辞書の整備
一般的な同義語を登録した辞書を作成する必要がある。

・高速辞書探索方式の開発
大規模な辞書に対応するため、高速な同義語辞書アクセスアルゴリズムとそのための辞書構造について検討する必要がある。

参考文献

(1)畠山他、自由語検索における異表記、異表現解消法、第33回情処全大
(1)第20回　国語審議会報告、外来語の表記(昭和29年)
(2)伍井他、カタカナ異表記処理、第38回情処全大

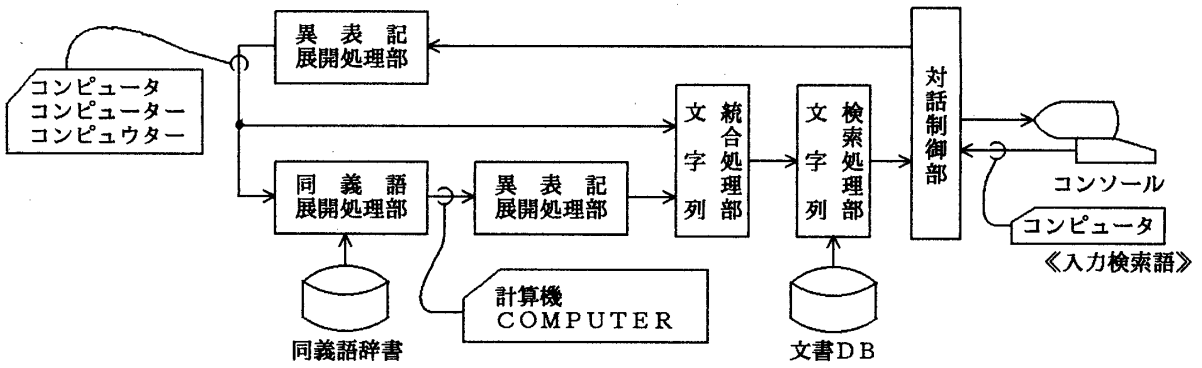

図１　同義語・異表記展開処理による文書ＤＢ検索の概略

Processing of Synonym and Variant Notations for Free Term Search

Atsushi Hatakeyama, Hisamitsu Kawaguchi, Kanji Kato, Hiromichi Fujisawa
HITACHI, Ltd.