日本語の知的情報検索に於ける辞書を用いた異表記処理

春遍雀來(Jack Halpern)
日中韓辭典研究所(CJKI)

要旨

日中韓各語の表記法の複雑さは、計算言語学処理用の各種ツール、特に知的情報検索用ツールの開発者にとって、大きな課題となっている。この難しさは、これら各語に標準的な正書法がないことで増幅され、特に表記のゆれが著しい日本語に於て顕著である。本稿では日本語に於ける表記のゆれの類型論に焦点を合わせ、その言語学的な諸問題について手短に分析し、異表記処理に於て語彙データベースが中心的役割を果たす理由について論じる。

1. 序論

日本語に於ける情報検索の難しさには種々の要因がある。真の「知的」検索を実現するには、多くの課題を克服しなければならない。主な課題には以下のようなものがある：

1. 標準的正書法の欠如。極めて多数の異表記体と文字種を処理するには、**表記法間検索**(Halpern 2000)のような高度な情報検索技術を備えていることが必要である。
2. 形態論上の複雑さ。これは、正確な形態素解析ツールの開発にとって非常に困難な課題となっている。形態素解析ツールは、形態素レベルで、基準化、語幹抽出(活用語尾の除去)、基本形復元(複数活用形を単一の語に還元)を行う。
3. 正確な単語分節が困難であること。これは辞書検索や索引作成の際、生のテキストを有意の意味単位に分解し、各単語の境界を識別するものである。
4. 様々な検索技術が要求されること。これは語彙素に基づく検索、統語表現(「研究をした」から「研究する」のような)を突きとめること、同義語の拡張、言語間情報検索(CLIR)(Goto *et al.* 2001)等の技術を含む。
5. 各種の技術的な要求。これは複数文字セットと符号化方式間の相互変換、ユニコードやIMEのサポート等を含む。これらの問題の殆どはLunde(1999)で報告されているように十分に解決されている。
6. 数多い固有名詞。これは情報検索ツールにとってとりわけ大きな問題である。つまり、固有名詞は極めて数が多く、辞書無しには発見が困難であり、また表記がゆれている等がその原因である。

本論の焦点は異表記処理に置く。異表記処理とは日本語の表記のゆれの認識、基準化、変換のことである。本稿では日本語に於ける表記のゆれの類型論を概観し、その言語学的諸問題を簡略に分析し、異表記処理に於て、どのように語彙データベースが中心的役割を果たすかについて論じる。

2. 日本語に於ける表記のゆれ

2.1. 1つの言語、4つの文字種

日本語の表記法は非常に不規則的である。異表記や間違えやすい同訓異字語が数多くあるため、日本語の書記体系は、中国語を含む他のどの主要言語よりもはるかに複雑である。その主な要因として、日本語を書き表すために用いられる4つの文字種の複雑な相互作用がある。(Halpern 1990, 2000)。表6は「取り扱い」の異表記の様々なパターンを示している。

表６「取り扱い」の異表記

トリアツカイ	異表記のタイプ
取り扱い	「標準」表記
取扱い	送り仮名異表記
取扱	全漢のみ
とり扱い	漢字を平仮名で代用
取りあつかい	漢字を平仮名で代用
とりあつかい	平仮名のみ

日本語の情報検索がいかに困難であるかという例に、「きんのたまごをうむにわとり」がある。「標準」表記は「金の卵を産む鶏」であろうが、実際、「たまご」には４つの異表記(卵、玉子、たまご、タマゴ)があり、「にわとり」には３つ(鶏、にわとり、ニワトリ)、「うむ」には２つ(産む、生む)がある。「金の卵を生むニワトリ」「金の玉子を産む鶏」等並べ替えると異表記が 24 通りにもなる。ウェブの検索で容易に確認できるように、これらの表記のゆれはウェブページで頻繁に見られる。アプリケーションが異表記処理機能を備えていなければ、ユーザに表記のゆれを見つける望みがないのは明らかである。

2.2. 送り仮名のゆれ

日本語の表記のゆれで最もよく見られる類型の一つは、「送り仮名」と呼ばれる仮名文字の語尾で起こる。動詞(「飛出す」)から派生した名詞(「飛出し」)のような一部の送り仮名の異表記をアルゴリズムで生成することは可能ではあるが、一般に異表記のデータベースが必要である。送り仮名はその使用法がしばしば予測不可能であり、異表記が数多くあることから、日本語の異表記処理に於て重要な役割を果たすのである。

表７ 送り仮名の異表記

英語	読み	「標準」表記	異表記
Publish	*kakiarawasu*	書き表す	書き表わす 書表わす 書表す
Perform	*okonau*	行う	行なう
Handling	*toriatsukai*	取り扱い	取扱い 取扱

2.3. 文字種間の表記のゆれ

日本語は４つの文字種の組み合わせで表記される。「漢字」、「平仮名」と「片仮名」と呼ばれる２種の音節文字、それに「ローマ字」である。日本語情報検索に於て大きな役割を果たす文字種間の表記のゆれは、極めて頻繁に起こり、殆ど予測不可能である。そのため同一の単語が平仮名、片仮名もしくは漢字で書かれ、更には２種の文字の混ぜ書きの可能性もある。表８は日本語の文字種間の表記のゆれの主なパターンを示している。

表８ 文字種間の表記のゆれ

漢字 対平仮名	大勢　おおぜい
漢字 対 片仮名	硫黄　イオウ
漢字 対平仮名 対 片仮名	猫　ねこ　ネコ

片仮名 対 混ぜ書き	ワイシャツ　Yシャツ
漢字 対片仮名 対混ぜ書き	皮膚　ヒフ　皮フ
漢字 対混ぜ書き	彗星　すい星
平仮名 対 片仮名	ぴかぴか　ピカピカ

2.4. 仮名表記のゆれ

近年、主として借用語を書くための文字として、音節文字である片仮名を使用する傾向が急激に高まっている。日本語情報検索に於ける主要な問題の一つは、この片仮名の表記がしばしば不規則なことである。つまり、同一単語に対して、アルゴリズムでは生成不可能な、複数の予測し難い書き方をする場合が、極めて普通に見られる。一方、平仮名は主として文法要素及び和語を書くのに用いられる。平仮名の表記法は概して安定しているが、少数ながら不規則な異表記が存在する。仮名表記のゆれの主な例をいくつか表９に挙げる。

表９ 片仮名と平仮名の異表記

類型	英語	読み	「標準」表記	異表記
長音符号	Computer	*konpyuuta* *konpyuutaa*	コンピュータ	コンピューター
長母音	Maid	*Meedo*	メード	メイド
複数仮名	Team	*chiimu* *tiimu*	チーム	ティーム
旧仮名遣いの継承	Big	*Ookii*	おおきい	おうきい
づ 対 ず	Continue	*Tsuzuku*	つづく	つずく

上記の表は、仮名表記のゆれの主要な類型を簡単に紹介したに過ぎない。他にも数多くあり、例えば、任意で中に使用する中黒、小文字の片仮名の異表記「クォ)対「クオ」、また伝統的仮名遣い(じ対ぢ)や歴史的仮名遣い(い対ゐ)の使用等がある。

2.5. 雑多な表記のゆれ　日本語には、他にも様々なタイプの表記のゆれがあるが、それらは本稿の範囲を超えており、ここでは二、三の主なものに触れておくに留める。詳細な記述が Halpern(2000) にある。

2.5.1. 漢字表記のゆれ　日本語の書記体系は戦後に大きな改革を経て、漢字の字形は既に標準化されてはいるが、まだかなりの数の異体字が一般に使用されている。例えば、現代日本語の省略形(「歳」に対する「才」や「幅」に対する「巾」)、固有名詞や古典作品に残る伝統的な形(「島」に対する「嶋」や「発」に対する「發」)等である。

2.5.2. 同訓異字語　日本語の書記体系が複雑である理由として、多数の同訓異字語(発音は同じだが表記が異なる語)とその様々な表記のゆれの存在がある(Halpern 2000)。ひとつひとつの漢字に多くの訓読みがあるだけでなく、多くの訓読みの単語が驚くほど様々な書き方をされる。同訓異字語の中には各々が近似又は同一の意味を持つために混同され易いものも多い。例えば、「のぼる」は「上る」と書くと「上に行く」という意味だが、「登る」と書くと「(手足を使って)登る」を意味する。また、「やわらかい」は「柔らかい」もしくは「軟らかい」と書くが、意味は同じである。

3. 語彙データベースの役割

日中韓各語では表記法が不規則なので、異表記処理のような語彙素レベルの処理を、例えばバイグラミングのような確率的手法だけに基づいて行うことはできない。Brill(2001)及び Goto *et al.*(2001)等

の多くの試みがこの方針に沿って行われ、ある研究者は辞書を用いた手法と同等の成果があったと主張している。一方、Kwok(1997)は非常に小さな辞書と単純な分節ツールで良い成果を上げたと報告している。

このような方法は純粋な情報検索には十分かもしれないが、異表記処理を行うには不十分である。Emerson(2000)や他の研究者は、語彙素を処理できる強力な形態素解析ツールは、バイグラムやNグラムよりむしろ大規模な計算機辞書(10万語でもまだ非常に小さすぎる)を備えていなければならないことを示している。

日中韓辭典研究所(CJKI)は日中韓各語のコンピュータによる辞書編纂を専門にしており、包括的な日中韓各語語彙データベース(現在約600万語)を編纂するために、特に異表記処理と固有名詞に重点を置いて、たゆまぬ研究と開発を展開している。知的情報検索用ツールと異表記処理に役立つ主要なデータベースには、a.包括的日本語表記法データベース、b.同訓異字語意味分類データベース、c. 言語間情報検索用英日辞書があり、未登録異表記の識別ルール集については現在開発中である。

結論

日本語の情報検索用ツールは、情報検索に於ては特に、また情報技術一般に於ても、益々重要になりつつある。これまで述べてきたように、日本語の表記法が不規則であるために、知的情報検索には高度な形態素解析ツールだけでなく異表記処理のために細かく設計された語彙データベースが必要である。

表記の曖昧性を解除する日本語情報検索ツールがあるとしても、その数は非常に少ない。なぜならば、真の「知的」情報検索を実現するには、辞書を用いた異表記処理機能を備えているだけでなく、言語間情報検索同義語拡張処理、同音異義語間検索といった新しい技術も備えていなければならないからである。

現在、当研究所は知的日中韓各語情報検索ツールの構築や、精度の高い分節処理技術を支えるのに必要な語彙資源の、更なる拡張を図っている。

参考文献

Brill, E. and Kacmarick, G. and Brocket, C. (2001) *Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs.* Microsoft Research, Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

Emerson, T. (2000) *Segmenting Chinese in Unicode.* Proc. of the 16th International Unicode Conference, Amsterdam

Goto, I., Uratani, N. and Ehara T. (2001) *Cross-Language Information Retrieval of Proper Nouns using Context Information.* NHK Science and Technical Research Laboratories. Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan

Halpern, J. (1990) *Outline Of Japanese Writing System.* In "New Japanese-English Character Dictionary", 6th printing, Kenkyusha Ltd., Tokyo, Japan (www.kanji.org/kanji/japanese/writing/outline.htm)

Halpern, J. (2000) *The Challenges of Intelligent Japanese Searching.* Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

Kwok, K.L. (1997) *Lexicon Effects on Chinese Information Retrieval.* Proc. of 2nd Conf. on Empirical Methods in NLP. ACL. pp.141-8.

Lunde, Ken (1999) *CJKV Information Processing.* O'Reilly & Associates, Sebastopol, CA.