異表記処理能力を備え持つ情報検索システムの評価

久保村 千明† 亀田 弘之‡
山野美容芸術短期大学 美容保健学科† 東京工科大学 工学部 情報工学科‡

1 はじめに

高度な情報検索システムはその機能の一つとして，文字表記のゆれ（異表記）に対応する能力を備え持つことが重要である．例えばキーワード“バイオリン”を検索した場合，システムはこのキーワードの異表記である“ヴァイオリン”をも検索できなければならない．しかしながら現在利用されている情報検索システムでは，入力されたキーワードに対する異表記を十分に処理することは不可能である．このような状況から，筆者らは異表記の一つである片仮名異表記を処理することができる情報検索システム[1]を構築した．本稿では構築したシステムの評価とその考察について述べる．

2 書き換え規則

片仮名単語の処理には，すべての可能な表記（語形）を予めシステム辞書に登録する方法と，データベースに異表記を登録する方法と，辞書項目に記載されている表記（語形）を書き換え規則により，毎回推定し処理する方法とがある．

本研究では，辞書またはデータベースに異表記を予め登録しておいても，すべての異表記を予め網羅的に知ることが一般的には不可能であり，またシステム辞書が不用意に増大しないようにとの配慮から，入力文字列（単語）からその異表記を毎回生成し，ユーザが選択した異表記の検索を行う処理方式を採用した．この考えに基づき，文字列書き換え規則を，小型文字異表記，長音記号異表記，その他の異表記と分類し，合計 258 種類作成した．

以下３種類の書き換え規則について述べる．

2．1 小型文字異表記の書き換え規則

文字列中の文字（「ア」，「イ」，「ウ」，「エ」，「オ」，「カ」，「ケ」，「ツ」，「ヤ」，「ユ」，「ヨ」，「ワ」）と（「ァ」，「ィ」，「ゥ」，「ェ」，「ォ」，「ヵ」，「ヶ」，「ッ」，「ャ」，「ュ」，「ョ」，「ヮ」）との相互の書き換えにより生じる異表記を小型文字異表記とし，その書き換え規則を 24 種類設定した．以下にその例を挙げる．

＜フイルム，フィルム＞

2．2 長音記号異表記の書き換え規則

長音記号（ー）の有無，および長音記号をほかの空ではない文字列に書き換えることにより生じる異表記を長音記号異表記とし，その書き換え規則を 28 種類設定した．以下その例を挙げる．

＜コンピューター，コンピュータ＞

2．3 その他の書き換え規則

小型文字異表記，長音記号異表記以外の異表記をその他の異表記とし，その書き換えの規則を 206 種類設定した．以下にその例を挙げる．

＜ギリシア，ギリシャ＞

3 システムの構成と処理の概要

システムの構成と処理概要について述べる．

3．1 システムの構成

図１に示す IFA(Interface Agent)におけるキーワード生成部，IRA(Information Retrieval Agent)における検索システム選択部と検索式生成部，対話インタフェースと検索インタフェースから構成されるシステムを構築した．

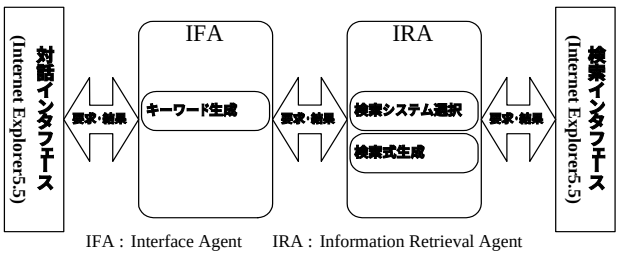

図１ システムの構成

3．2 処理の概要

本システムは対話インタフェースから片仮名文字列を入力し，それに対する異表記を IFA のキーワード生成部で生成する．次に生成した異表記を人間が選択し，入力文字列と選択した異表記の論理和検索式を IRA の検索式生成部で生成する．生成した検索式を用いて infoseek Japan，Yahoo! JAPAN，フレッシュアイの検索システムを検索システム選択部により人間が選択し，検索結果を検索インタフェースに表示する．

Evaluation of Information Retrieval System with Abilities of Processing Allographs
†Chiaki KUBOMURA・Department of Aesthetics and Health Science, Yamano College of Aesthetics
‡Hiroyuki KAMEDA・Department of Information Technology, School of Engineering, Tokyo University of Technology

4 評価

異表記処理能力を備え持つ情報検索システム（以下，評価システムと記す．）の有効性を確認するために三項目の評価実験を行った．実験環境は計算機 ARMADA M300(COMPAQ 社製，CPU：pentiumIII 500MHz，主記憶：256MB)，オペレーティングシステム Windows98 を用いた．以下評価実験について述べる．

4．1 評価実験1

（1）概要　評価システムにより生成される異表記が有効であるかどうかを確認する為の実験である．評価用単語として，片仮名辞書[2]の見出し語（合計 10,470 語）から無作為抽出した単語 20 語を準備した．

評価用単語を評価システムの入力として異表記を生成し，それらを“自分が異表記として利用しても良い”（以下，“視点１”と記す．）と“異表記として使われているかもしれない”（以下，“視点２”と記す．）という二つの視点により被験者に判断してもらい，それらが生成された全異表記うちの何%かを確認する．なお被験者は工学系４年生大学の学生 11 名と工学系教員１名の計 12 名である．

（2）手順　**a)**評価システムにより，評価用単語 20 個の異表記をすべて生成する．**b)**評価用単語毎に生成された異表記を提示し，生成された異表記が視点１と視点２の観点から適当かどうかを被験者に判断してもらう．この判断は評価用単語 20 語のすべての異表記に対して行う．**c)**上記の **b)**を計 12 名の被験者に対して行う．

（3）結果　評価用単語 20 語のうち，視点１の割合の平均は 7.1%であり，標準偏差は 4.1%であった．また視点１と視点２を併せた割合の平均は 17.6%であり，標準偏差は 8.3%であった．

4．2 評価実験2

（1）概要　評価システムを適用した場合と，適用しない場合との検索結果の差から，評価システムの優位性を確認する為の実験である．検索実験の際のキーワード（以下，“評価用キーワード”と呼ぶ）として，前述の片仮名辞書の見出し語（合計 10,470 語）から無作為抽出した単語 100 語を準備した．比較対象は，平成１４年８月の中旬から下旬にかけての時点で，検索エンジン infoseek Japan（以下，“基準システム１”と記す．）とフレッシュアイ（以下，“基準システム２”と記す．）が蓄積していたインデックスデータ群とした．

（2）手順　**a)**基準システム１と基準システム２に評価用キーワードを入力し，ヒット件数をそれぞれ調べる．**b)**評価システムに同じ評価用キーワードを入力し，ヒット件数を調べる．**c)**すべての評価用キーワードに対して上記の **a)**と **b)**を実行する．**d)**評価システムにおいて，100 個のキーワード中何個のキーワードが基準システム１，基準システム２よりもヒット件数の点で増加したか確認する．

（3）結果　基準システム１は 100 個の評価用キーワードのうち，ヒット数がともにゼロであったものが 32 個，ヒット数がゼロ以外でかつ同数であったものが 33 個であり，残り 35 個に対しては，評価システムのヒット数が大きかった．基準システム２は 100 個の評価用キーワードのうち，ヒット数がともにゼロであったものが 30 個，ヒット数がゼロ以外でかつ同数であったものが 23 個であり，残り 47 個に対しては，評価システムのヒット数が大きかった．

4．3 評価実験3

（1）概要　評価実験２においては，前述の片仮名語辞典から無作為に 100 語を選び出して実験を行ったが，評価実験３では片仮名語辞典に載っていない新出片仮名表記 6 語，異表記を数多く有する片仮名表記 2 語，特殊な表記を有する片仮名表記 2 語の計 10 語に対して，評価実験２と同じ実験を行った．

（2）結果　基準システム１は 10 個の評価用キーワードのうち，ヒット数がゼロ以外でかつ同数であったものが 2 個であり，残り 8 個に対しては，評価システムのヒット数が大きかった．基準システム２は 10 個の評価用キーワードのうち，ヒット数がゼロ以外でかつ同数であったものが 3 個であり，残り 7 個に対しては，評価システムのヒット数が大きかった．

4．4 考察

３つの評価実験より，評価システムにより生成される異表記の有効性及び評価システムの有効性が確認できた．

5 おわりに

本稿は異表記処理能力を備え持つ情報検索システムの評価実験について述べた．

参考文献

[1] 久保村千明,亀田弘之,“片仮名異表記処理能力を備えもつ情報検索システム,”電子情報通信学会論文誌 D-II，(in printing)．

[2] 現代言語研究会,“カタカナ語新辞典,”アストロ教育システムあすとろ出版部,1994.