7G-1

漢字の意味に着目した類義表現検索の検討

斎藤 珠喜
NTTヒューマンインタフェース研究所

1. はじめに

日本語の情報処理において文章が表現している意味を扱おうとすると、形態素解析、構文解析等の文法的解析に加えて意味のレベルの処理を結合させることが必要となる。情報検索のような日本語処理の適用例を考えても、検索者の意図(検索要求時に使用する言葉)とデータ登録者の認識(データに付与するキーワード)とが一致する保証はなく、したがって検索要求中のキーワードの同義語,類義語への展開は不可欠であり、シソーラスを用意してこの種の対応をとることが一般に考えられている。しかし、類義語,関連語の範囲をどこまでにするか、さらに、語の多義の問題もあり、本質的な解決にはならない。

森岡[1]は漢字を形態素の資格をもつものと考え、語構成上の特徴によって分類している。ここでも語を構成する要素としての漢字の字義に着目し、漢語(漢字の熟語)の意味を、それを構成する各漢字の意味の合成として考えてみた。

2.漢語の意味と漢字の意味

二文字の漢語の意味とその各構成要素の漢字の意味との関係は、例えば新字源によれば(A)共通の意味をもつ二字をかさねたもの,(B)反義語,類義語などの二語を並列したもの,...など、13種類に分類している。実際には二つの漢字の間には種々の格関係を含む多様な関係があるが、ここでは最も簡単に二つの漢字の"and"または"or"と見なした場合について、テキストの検索に適用したときの効果を調べた。実際、ユーザからの検索要求で"青空"という言葉が使われ、テキストデータには"青い空"と記述されていると、類義語辞書中に"青空"と"青い空"とが類義語であるという関係が書かれていなければ、ストリングのマッチングによってこのテキストを検索することはできない。一般に類義語辞書には"青空"に対して"碧空","青天"等の言葉はあっても、"青い空"というような複数の語からなる言葉は登録されないので、このような簡単な言い換えにも対応できないことになる。これに対して漢語をその構成漢字の意味の合成と見なせば、"青空"=("青"and"空")であるから、"青(い)"と"空"の二語に分けたのと等価になり、それぞれの語によるマッチングが可能になる。さらに、"青"の類義語として"碧"等が定義されていれば、"碧"を使った表現("紺碧の空","碧空"...)とのマッチングも可能になる。

3.テキスト検索に関する実験

上記の方法のテキスト検索における効果を調べるため、角川類語新辞典の用例文(6万件以上)を対象としてストリング・マッチングによる検索の実験を行った。結果は、構成要素に分解し各構成漢字とその類義文字とによる検索結果と、類語新辞典の類義語(同一の項目中に記載されている言葉)による検索結果とを、再現率,適合率で比較する。

Similar Expression Retrieval base on Meaning of Kanji Character
Tamaki SAITO
NTT Human Interface Laboratories

なお、個々の構成漢字に関する類義文字の収集は意味から漢字を検索できる漢和辞典が必要であるが、ここでは、岩波国語辞典の漢字母(造語成分として漢字)の情報を逆引き(意味から検索)して求めている。

(1)『美人』の場合

(a)本手法

美:(妖 艶 麗 佳 嘉 綺 嬌 媚 妙)

人:(女 娘 嬢 婦 姫)

<検索結果>

[正解] 容貌の美しい人 顔貌の美しい女性 麗わしい女性 ...等 計 98件

[誤り] 華麗に着飾った女 軽妙な人物 巧妙に人をだます...等 計 23件

[検索もれ] 色女 グラマー シャン 玉 等 6語 9件

[再現率]=98/107= 92%

[適合率]=98/121= 81%

(b)類義語による検索の場合

類語新辞典の類語:女 女子 女性 婦人 等 計 53語

<検索結果>

[正解] 美人(49) 美女(5) 佳人(4) 等 計 72件

[誤り] 女児が誕生する シャンプー 玉砕 玉の輿 ... 等 計908件

[検索もれ] なし

[再現率]=72/107= 67%

[適合率]=72/980= 7%

(2)『青空』の場合

(a)本手法

青:(碧 蒼 紺)

空:(天 宇 宙)

<検索結果>

[正解] 大空が青く澄み渡る 紺碧の空 白雲なびく青い空 等 計 22件

[誤り] 空間を青の絵の具でぬりつぶす 計 1件

[検索もれ] 晴れた空 五月晴れの空 空がすっかり晴れ上がる 計 10件

[再現率]=22/32= 69%

[適合率]=22/23= 96%

(b)類義語による検索の場合

類語新辞典の類語:青天井 青天 碧空 蒼空 ... 等 7語

<検索結果>

[正解] 青空(4) 青天(7) 碧空(2) 蒼空(1) ... 等 計 17件

[誤り] なし 0件

[検索もれ] 同上 計 10件

[再現率]=17/32= 53%

[適合率]=17/17=100%

他の言葉も含め図1に結果を示す。

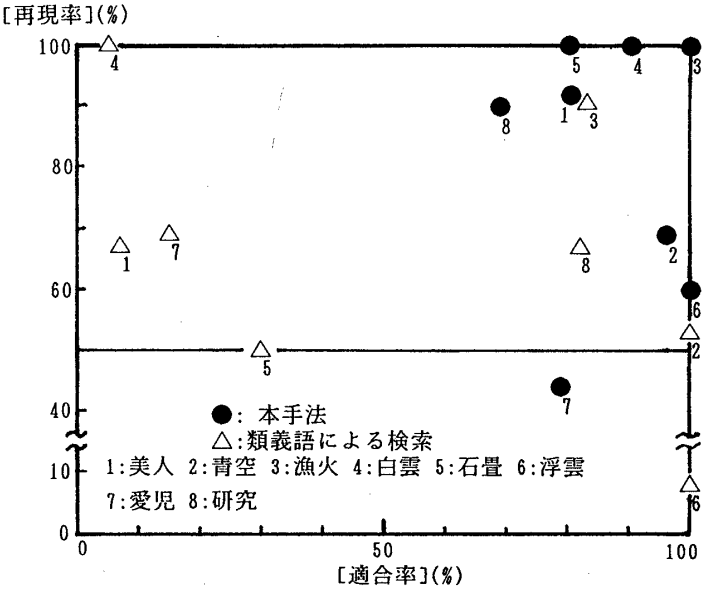

図1.検索結果の比較

4.まとめ

文章の検索において、漢語の類義表現を漢語の構成要素の各漢字の類義文字の合成によって近似する方法の効果を調べ、具体名詞の場合に効果が見込める結果を得た。特に、名詞にそれを修飾する語がついた形の場合は、構成漢字に分割することは名詞と修飾語の二語に分割することの効果が大きいと考えられる。ただし、外来語のように漢字と無関係な言葉、さらに漢字辞書外の文字を含む言葉は類義語辞書によるしかない。類義な漢字同士を集積した類義漢字辞書の構築あるいは漢字辞書の逆引きによる類義漢字の自動収集、さらに検索時の解析の高度化、等について引続き検討していく。

[文献]

[1]森岡健二: 語彙の形成,明治書院