漢字の持つ意味情報を利用した日本語の検索

2B−3

大須賀勝美，寺澤登紀江，朝木由貴，鈴木佐和，國井佐和子，黒川一夫

東京理科大学

1．はじめに

ある質問に対してその意味を理解し、知識を利用して必要な答を導くためには、大量な情報の中から答を導くために利用できる情報を集めることが必要である。知識として自然言語で書かれた文書を利用し、その文書中から必要な情報を含む文章を検索することを考えた場合、日本語では表意文字である漢字が用いられているため、厳密という訳でなければ、膨大な辞書データを利用したり、複雑な解析処理を用いなくても、比較的簡単に意味的な情報を扱うことができるものと考えられる。本研究では、漢字の持つ意味情報に着目し、全文サーチによる情報検索システムとして、利用する方法を検討する。

2．日本語文章の意味情報

日本語文章の場合、漢字が意味を持ち単語としての役割を果たすので、意味情報のほとんどは漢字の部分によって表現されている。したがって、漢字の部分だけを残して平仮名を除いても、元の文章の意味を大体知ることができる。厳密に意味を調べるには、平仮名も含めて文法的な解析をする必要があるが、すべての文章について解析する必要はなく、検索によって得られたものを更に詳しく調べる方が効率的である。また、同じ意味の文章でも、多少語順が変ることもある。同じ概念を表現しようとすれば、同じような意味を表す漢字が用いられるはずあり、二つの文章で使用されている漢字の関連性から、意味的なつながりをある程度知ることができる。また、漢字によって表される熟語の意味を見てみると、個々の漢字の持つ意味を含んでいる。

例えば、

実測する　実際に測定する
測定する　計測する　測る　計る　など、

それぞれの文字の意味によって色々なつながり方があり、意味的には分解して考えることもできる。同じ意味を持つ熟語でも色々と表記できる。また、意味的に関連性のある単語は、同じ漢字が利用されているか、関連性のある字が利用される場合がほとんどである。

3．漢字の持つ意味情報

漢字が持つ意味情報を、次に述べるような形で整理して利用する。

(1) 関連性の強い文字の組

漢字同士の間の関連性として、次のような文字の集合を、リスト化して利用する。

・新旧字の対応（同じものとして扱う）
・同義文字、類義文字
　異なる・違う、隔てる・離れる、など
・反意文字、対意文字
　大・小、長・短、左・右、前・後など

(2) 部首による分類

部首は漢字の持つ意味の主要な部分を占め、同じ部首を持つ漢字の間には、ある共通の概念が存在する。しかし、部首による概念を含むものと含まないものや、意味を持たずに形状で分類される部首もある。そこで各文字について、部首の表す意味を含む度合いの評価付けしておく。さらに、自然に関するもの、人間に関するもの、などのようにいくつかの部首に共通な概念として存在するものは、部首同士の関連性として結び付けておく。

(3) 意味による分類

部首とは関係なく、ある共通の概念を含んで、意味的に関連性の高い漢字の集合を、細分化して分類する。意味だけでなく、名詞、動詞、形容詞など、文法的に働く機能も考慮する。1つの漢字がいくつかの分類に属することもある。

Retrieval for Japanese using Semantic Information of Chinese Character.
Katsumi OSUGA, Tokie TERASAWA, Yuki ASAKI, Sawa SUZUKI, Sawako KUNII, Kazuo KUROKAWA
Science University of Tokyo

・人を表す、物を表す、など（名詞的なもの）
・手の動作、物の動き、など（動詞的なもの）
・色、形状、量、方向、など（形容詞的なもの）

漢字の持つ意味をなるべく詳しく表現することにより、文章の意味を知る上で効果がある。実際には文字同士の関連性が必要であり、各文字同士の関連度を数値化して、各文字について関連度の強いものから順番に何文字分かの関連する文字とその強さを示した一覧表を作成して利用する。

4．文章の関連度の評価

二つの文章で使われている漢字に着目して、それらの文字の間の関連度を利用することにより、文章の表す意味の関連性を評価値として算出する。各文字に対して一致したり、関連する文字の意味のつながりによる評価値を加え合わせる。一つのキーワードから発生した関連文字が対象となる文章中にいくつか含まれるが、一番評価値の高い値だけを加算していく。また、関連性を調べるだけなら、語順や、単語や熟語の文字の並び方を調べなくても、文章中でどのぐらい関連性のある文字が使用されているか調べるだけで十分である。

5．情報検索への利用

意味的に関連性のある文章を、全文検索で探そうとした場合に、すべての文章に対して関連性を調べていたのでは、必要な情報を持つ文章数は少ないため効率が悪い。直接的に近いものから順にたどって、必要なものを見つけ出せることが望ましい。漢字だけに着目しているので、事前に文書中でどの漢字がどこで使われているかを調べておき、検索の際に利用する。また、どの文字がどのくらい使用されているか調べてあるので、索引としての役割も持ち、検索を行う文書としての文献を選択する際にも利用する。

検索の手順としては、キーワードとして入力された漢字、一文字ずつ順番に、各文字について関連性の高い文字を一覧表によって調べ出す。次に事前に調べてある各文字が使用されている文章の情報から、それぞれの文章に対して関連性の評価値を与えていく。入力されたすべての文字について関連性を調べた後に、評価値がある値を越えているものを集めて、評価値の高い順に検索結果として文章を出力する。

6．実際の検索システムでの例

日常よく使われる漢字約3000字に対して、上述のような分類により意味の関連性を表したリストを作成し、これを用いて、全文サーチによる情報検索を行った。対象となる文章としては、教科書などの文書を用いた。一つの文書の文章数は700～900文で、1文章当たり漢字数は平均15～20文字である。使用されている漢字数は 1000～1500文字、漢字の種類数は500～600である。キーワードとして、漢字10文字前後を与えて検索を行った。結果として、漢字の意味による関連性を考慮しているために、同じ文字が使用されていなくても関連性のある文章が検索できることが確認された。しかし、それと共に不要なものも取り出されるようになってしまったので、キーワードとしての各文字の重要度に応じて、検索の際に重み付けを行うことにより改善を図った。また、対象となる文書の内容によって、検索方法を変えた方がよいこともわかった。また、関連性のない文書を利用した場合には、どの文章も評価値が小さく、該当する文章はないと判断された。

7．まとめ

漢字の部分に着目することにより、日本語において全文サーチで文書中から、文章単位で必要な情報を効率良く得る方法を調べた。現段階では、質問に対して適切なキーワードを与えたり、検索の対象とする文献を選んだり、出力結果を判断して最終的に答を導き出す作業は人間が行っている。支援システムとしての役割を果たすのには、十分な効果が得られるものと考えられる。出力されてきた文章を詳しく解析して関連性を評価する事により、本当に必要なものだけに、絞り込めるようにする。更にこの作業を自動化することにより、人工知能的な処理へと応用させていく。

《参考文献》

(1) 斎藤珠喜：“漢字の意味に着目した類義表現検索の検討”，第39回情処全大，7G-1，pp.742（1989.10）.
(2) 加藤寛次 他：“全文検索用テキストサーチマシンの開発”，信学技報，DE89-38，pp17-24（1989.12）.