固有名詞分類属性を利用した情報検索

5N－3

竹元義美　　山田洋志　　福島俊一
NECヒューマンメディア研究所

1. はじめに

インターネットが広く普及し大量のテキストデータへのアクセスが可能となった。テキストデータが爆発的に増大するにつれ、高精度な情報検索技術への要求が高まっている。

検索キーが多義語の場合、ユーザの意図に適合しない大量の検索結果を得る問題がある。これに対し、検索精度を高めるために、共起語を多義の絞り込みに用いる手法がよく知られている。また、検索対象テキストを意味構造化し、単語の役割を考慮してクエリとの照合を行い検索意図の多義を解消する研究[3]がある。しかし、検索キーが多義の場合、多義を絞り込む制約情報をユーザが与えないと精度の改善は厳しい。従来法では、例えばテニス選手の「グラフ」を検索する場合に、単に「グラフ」という検索キーでは検索結果に多義によるゴミが発生するので、ユーザは「グラフ選手」や「グラフ AND テニス」のように検索条件を増やすことになる。これとは別な視点から制約情報を与える方法も考えられる。例えば「グラフ：人名」のような与え方である。

本稿では、検索キーに多義を絞り込むための制約情報として分類属性を与えることによる高精度な検索方式を提案する。設定すべき分類属性として今回は、固有名詞分類属性を取り上げた。固有名詞は文章の特徴になりやすく検索時の絞り込みの効果が期待できること、固有名詞分類属性はユーザにとって比較的容易に付与できてシステム面でも実現できる見込みが高いことなどから固有名詞に着目した。

2. 属性を利用した検索の方式

図１に本検索方式の構成を示す。検索用の単語インデックスを作成する登録系と検索キーを単語インデックスとマッチングする検索系とに大きく分かれる。登録系では、検索対象テキストの形態素解析結果に対して、分類属性付与部により分類属性を付与する。インデックス作成部は、単語の見出しとその

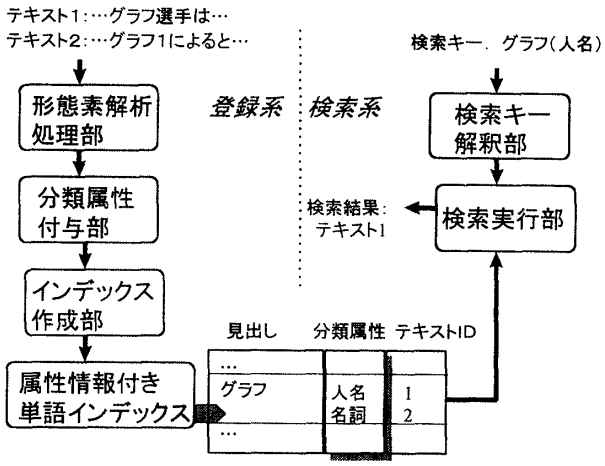

図１：属性利用検索方式の構成

単語を含むテキストＩＤ情報に分類属性情報を加えて、属性情報付きの単語インデックスを作成する。検索系では、ユーザが属性付きの検索キーを入力すると、検索キー解釈部は見出しと属性情報とに解釈し、検索実行部に渡す。検索実行部では、その見出しと属性情報とで属性情報付き単語インデックスを検索し、テキストＩＤ情報を検索結果として返す。

多義語の検索キーに意味や分類に関する属性情報を与えることにより、ユーザは検索キーに検索意図を表現することができる。3 節以降では、属性情報として固有名詞属性（人名・地名・組織名）に限定した場合について述べる。この場合、分類属性付与部には、固有名詞抽出技術[1]を利用する。

3. 本方式の評価実験

本方式の評価のため、新聞記事を対象とした検索で固有名詞が検索キーに使われる場面を想定し、次の二つの実験を行った。第一の実験では、固有名詞が検索キーに使われたときに、それが多義をもつ率（提案方式が作用する率）がどれくらいかを調べた。本来は実際の検索ログの分析をすべきであるが、入手困難のため、検索対象テキスト中の固有名詞の出現率を検索キーに使われる率と同じであると仮定した。第二の実験では、本方式が作用するケースで、検索精度に対する最大の効果を見積もるために、形態素解析および固有名詞分類属性付与の精度を100%と仮定し、属性を指定した検索と属性を指定しない通常の検索との適合率を比較評価した。なお、仮定から再現率は 100%となる。

An Information Retrieval Method Using Proper Noun Classification Attributes
Yoshikazu Takemoto, Hiroshi Yamada,
and Toshikazu Fukushima
Human Media Research Labs, NEC Corp.

3.1 検索キーにおける多義の発生率

RWC テキストデータベース[2]（RWC-DB-TEXT-95-2；毎日新聞94年版3000記事に正解品詞情報が付与されたもの）から一般名詞および固有名詞を抽出・カウントし、表1にまとめた。

表1で、多義発生率以外の数字は、それぞれの語の表記レベルでの異なり数・のべ数を表す。多義語Aは一般名詞と固有名詞とで表記が共通の語である。多義語Bは固有名詞間で表記が共通で、人名・地名・組織名の分類情報が異なる語である。多義発生率は、固有名詞のうち多義語A,Bの占める割合であり、のべ数で42.7%を占めた。

表1：固有名詞多義語の発生率

	表記異なり数	表記のべ数
一般名詞	20,357	182,231
固有名詞	12,560	53,428
多義語A	664	14,985
多義語B	374	7,819
多義発生率(%)	8.3	42.7

表2：通常検索と属性指定検索の適合率の比較

	全体	人名	地名	組織名
件数	337	117	136	84
適合率1	53.4	55.7	57.8	43.2
適合率2	100	100	100	100

3.2 属性利用検索方式の評価

固有名詞多義語（多義語A,Bの合計1,038件）のうち、高頻度語（上位200）と表記が共通する語を集めて検索キーのセット（337件）を作成した。図2に作成した検索キーの例を示す。

これらの検索キーの見出しのみで検索を行った場合（通常検索）と属性を含めた検索を行った場合（属性利用検索）との適合率の比較を表2に示す。表2で、適合率1は通常検索の、適合率2は属性指定検索の平均適合率を示す。通常検索では、適合率は全体で53.4%である。つまり検索キーが多義語の場合、46.6%が検索過剰となることがわかる。

4. 考察および今後の課題

3.1節の実験により、新聞記事を対象とした固有名詞検索の場面で多義が発生するケースが42.7%存在し、3.2節の実験により、多義が発生するケースで、形態素解析および固有名詞抽出精度100%を仮定した場合に46.6%適合率が向上することを見積もることができた。

以下、今後の課題・展望について述べる。

```
タイ　　地名
横浜　　組織
横浜　　地名
巨人　　組織
社会　　組織
米　　人名
米　　地名
羽田　　人名
羽田　　地名
```

図2：作成した固有名詞検索キーの例

(1) 作成した検索キーのセットは、検索対象テキスト中の出現頻度を検索キーの利用頻度に当てはめるという仮定のため、実際の検索キーとのギャップがあると考えられる。今後は実際の検索ログで本方式の有効性を検証したい。

(2) 本方式の精度見積もりのため、品詞・分類属性付与がすべて正確にできたとして評価した。しかし、とくに現状の固有名詞抽出精度[1]は十分ではなく、抽出の誤りや洩れが起きる。今後は固有名詞抽出技術の高精度化が必要である。

(3) 検索キーに属性を付加するという従来より余分な入力をユーザに求めることになる。実用面ではユーザインターフェースの設計が重要となる。

(4) 今回は固有名詞検索に限定したが、一般名詞についての属性を考えて枠組を拡大してゆく。また、固有名詞も人名・地名・組織名の3つの分類以外に、同一人物・企業などのレベルで検索する要求もあり得るため、より詳細な適合を考えてゆく必要がある。

(5) 本検索方式の同義語展開への発展も考えたい。図2に示した「米」（地名）のようなクエリは直接的にはないにしても、例えば「アメリカ」という検索キーを同義展開して地名を意味する「米」を検索する要求はあり得る。

5. おわりに

検索キーに多義を絞り込むための属性情報を与えることにより、適合率の高い検索手法を提案した。

新聞記事を対象とした固有名詞検索の場面で多義が発生するケースが42.7%存在し、多義が発生するケースで、形態素解析および固有名詞抽出精度100%を仮定した場合に46.6%適合率が向上することを見積もることができた。

今後は今回の実験で得られた知見をもとに、本方式を組み込んだ検索システムを設計・試作する。

参考文献

[1] 竹元・山田・若尾,"日本語新聞記事からの固有名詞情報抽出",情処第53回全大,1996
[2] RWCテキストデータベース報告書,平成7年度版
[3] 岸本・須之内・塚田・千葉・石川,"テキストの構造化に基づく検索システム",情処論 Vol.35 No.5,1994