

# 文や句による日本語テキストの検索

## —語と語の係り受けを用いた検索の試み—

5 T-1

那須川哲哉

日本アイ・ビー・エム株式会社 東京基礎研究所

### 1. はじめに

柔軟なテキスト検索を行なうためには、ユーザの検索要求としてキーワードでなく文や句をそのまま入力できることが望ましいが、日本語においては、

- 形態素が分かち書きされていない
- 語順の自由度が高い

などの特徴があることから、日本語テキストの検索システムでは、文や句による検索は、全文検索における完全一致の場合を除いては、実現が困難である。本稿では、システムの頑健性を維持したままで文や句による柔軟な検索を可能にする手法として、語と語の係り受けを用いた検索の試みを示す。

検索要求をキーワードでなく文や句で入力させることにより、

- 検索式を意識させない
- 語間関係を反映した検索を行なう

というメリットが得られるが、本稿では特に語間関係を反映した検索に焦点を当てる。

例として「計算機で開発」という内容の検索を考える。一般的なキーワード検索システムでは、附属語の助詞は検索対象にならず、「計算機」と「開発」のANDをとっただけでは「計算機を開発」も検索されてしまうため、語間関係を考慮した検索は困難である。全文検索システムを用いれば附属語も検索対象になるが、検索要求として「計算機で開発」を指定すると、「計算機で○を開発」のように挿入を含む文は検索されず、「計算機で」と「開発」のANDをとると、「計算機で…、○を開発」のように、「計算機で」が「開発」を修飾しないパターンも検索されてしまう。

このように従来の手法では、語と語の関係の扱いの問題から、文や句による検索は困難である。語間関係を考慮した手法は、研究レベルでは存在[1]するが、意味的な語間関係を重視してインデックスの構築に人間の処理が介在するため、大規模な文書データベースへの適用は困難であり、実用性を欠いている。

実用性を考えた場合、高い精度で自動的にインデックスを作成するのが望ましい。現在の自然言語処理技術で安定して高い精度を期待できるのは形態素解析レベルま

であり、省略の多い日本語の場合は特に、構文解析レベルで文全体を正しく解析するのは困難である。特に、文が長くなると構文解析の精度は落ち、長い文に対して完全な解析結果を得ることは困難である。ただし、長い文の場合でも、解析に失敗するのは、複文や重文における距離の長い係り受けの部分であることが多く、隣接した主語述語関係のような局所的な係り受けの解析は正しい場合が多い。また、日本語の場合は係りの方向が一方方向であることから、形態素解析で文節の認識に成功していれば、少なくとも文末の述部に近い部分での係り受け関係に関しては、比較的高い精度で解析できると考えられる。

そこで我々は、検索対象となる文書中の各文の構文構造を解析し、構文解析結果全体は用いずに、『係る語—係り関係—係られる語』という3項関係の集合体に変換して文書のインデックス情報として蓄積した上で、この3項関係を利用した検索を試みた。

### 2. 語と語の係り受けを用いた文書検索

本手法では、あらかじめ、検索対象となる複数の文書に対し、文中の各文を構文解析した上で、『係る語—係り関係—係られる語』という係り受けの3項関係を抽出する。そして検索の際には、同じ構文解析器を介して検索要求文から『係る語—係り関係—係られる語』という係り受けの3項関係を抽出し、基本的に同じ3項関係を含む文書を検索することで語間関係を反映した検索を行なう。

例えば「IBMがCADでロボットを開発。」という文からは、構文解析器を用いて、

```
開発【動詞】
├が【格助詞ガ】— IBM【名詞】
├で【格助詞デ】— CAD【名詞】
└を【格助詞ヲ】— ロボット【名詞】
```

のように、文中各語の品詞及び係り受け関係情報を得ることができる。この解析結果から、

```
IBM【名詞】— が【格助詞ガ】— 開発【動詞】
CAD【名詞】— で【格助詞デ】— 開発【動詞】
ロボット【名詞】— を【格助詞ヲ】— 開発【動詞】
```

という係り受けの3項関係を抽出し、これらの3項関係に対し、この文を含む文書の識別番号を結びつけた上で、検索用のインデックスとして保持する。

| 日付            | 選好度 | 記事のタイトル                 |
|---------------|-----|-------------------------|
| 940927 (1.15) |     | 中小の技術支援施設、都の計画見送り相次ぐ——  |
| 940905 (1.15) |     | 雇用悪化、短期景気判断と分離——企画庁、「運行 |
| 940915 (1.00) |     | CNC旋盤のタカハシキカイ、ASEAN軸に受  |
| 940906 (1.00) |     | エヌテック——LCD製造機械が新たな柱に、ア  |
| 940925 (0.50) |     | 円、波乱含みの展開——包括協議の結果で左右(金 |

図1:『円高で加速』の検索結果

| 日付            | 選好度 | 記事のタイトル                 |
|---------------|-----|-------------------------|
| 940925 (1.30) |     | 円、波乱含みの展開——包括協議の結果で左右(金 |
| 940903 (1.00) |     | 内外価格差の謎(6)米国と物価比較——割高招  |
| 940927 (0.35) |     | 中小の技術支援施設、都の計画見送り相次ぐ——  |
| 940905 (0.35) |     | 雇用悪化、短期景気判断と分離——企画庁、「運行 |
| 940930 (0.30) |     | 日銀岡山支店が管内企業の動向まとめる、原材料・ |

図2:『円高が加速』の検索結果

### 3. 新聞記事を用いたインデックス抽出実験

本手法の有効性を検討するため、日英機械翻訳システム JETS の構文解析器 [2] を用いて、1994 年 9 月及び 10 月の日経新聞の記事で実験を行なった。人事異動やスポーツ欄など、文としての体裁をなさない文が多数混在しているため、最終的に構文木が得られる割合は、9 月の記事 (14516 件) の場合 79.8%、10 月の記事 (15484 件) の場合 83.0%と、全体の 8 割程度であった。(記事あたりの平均文数は、9 月分が 10.7 文、10 月分が 12.5 文であった。) 抽出された語と語の係り受け関係は、のべにして、9 月分で約 100 万 (1005919) 件、10 月分で約 92 万 (919,454) 件であった。また、同じ係り受けの 3 項関係が、同一記事内で複数回出現する割合は 2.6%程度、同一月の記事内で複数回出現する割合は 20%程度であった。さらに、このようにして得られた係り受けの精度は、無作為抽出した 50 文を調べた範囲では 80%程度であった。

### 4. 選好度を用いた検索

前節の通り、構文解析で得られた係り受け関係だけをを用いるのでは、構文解析に失敗して係り受け関係の得られない記事が検索できなくなるため、「A - (関係 R) - B」という係り受けを含む記事を検索をする際には、

- 「A」を含む記事
- 「B」を含む記事
- 「A - (関係 R) - B」を含む記事

各々に対して選好度を与え、与えられた選好度の総和の高い記事から順に検索結果を表示するようにした。

1994 年 9 月の記事 14516 件から、『円高で加速』と『円高が加速』で検索された結果の上位 5 記事を各々図 1 と図 2 に示す。14516 件中で『円高』あるいは『加速』を含む記事は 525 件あり、いずれの検索でも 525 件の記

| 日付            | 選好度 | 関連する格内容及び記事のタイトル                                                      |
|---------------|-----|-----------------------------------------------------------------------|
| 940927 (1.15) |     | 【移転が/円高で → 加速 → して】<br>中小の技術支援施設、都の計画見送り相次ぐ——                         |
| 940905 (1.15) |     | 【円高で/海外移転が → 加速】<br>雇用悪化、短期景気判断と分離——企画庁、「運行                           |
| 940915 (1.00) |     | 【ASEAN進出が/円高で → 加速 → 、伴】<br>CNC旋盤のタカハシキカイ、ASEAN軸に受                    |
| 940906 (1.00) |     | 【メーカーは、/円高で/海外進出を → 加速】<br>エヌテック——LCD製造機械が新たな柱に、ア                     |
| 940925 (0.50) |     | 【失望感から/円高が/円買い・ドル売りが → 加速<br>→ する可能性/し、更新】<br>円、波乱含みの展開——包括協議の結果で左右(金 |

図3: 係り受け情報を加えた『円高で加速』の検索結果

事に選好度が与えられたが、検索文中の文間関係を含む記事により高い選好度が与えられている。

### 5. 係り受け情報の提示

従来の手法では、検索結果を表示する際に、タイトルのみを表示する場合が多く、ユーザが検索結果から欲しい情報を得るためには、検索された各記事の内容を逐一見る必要があった。そこで、検索された文書の中で検索要求に関連している情報を把握し易くするために、あらかじめインデックスとして抽出してある文書中の語の係り受け情報をタイトルと併記するようにした(図 3)。

### 6. まとめ

語と語の係り受けをインデックスとして用い、検索要求文中の各語句の検索と係り受け関係の検索を、選好度という形で融合させることで、システムの頑健性を保ちながら語間関係を反映した検索を実現することができた。また、検索結果を表示する際に、インデックス化した係り受け情報を併記することで、検索文書の内容把握を容易にすることができた。

本手法では、検索要求文中に含まれる全ての語句と係り受け関係について、同じ語句や係り受けを含む文書を検索し、より多くの語句と係り受けを共有する文書により高い選好度を与える。従って、既出の例のように短い句だけでなく、長い文や複数の文を検索文として入力しても同様に動作する。検索文が長くなると検索対象の要素数が多くなるため、選好度を与えられる文書の数が増大するが、最終的に選択する文書の選好度のしきい値を調整することで、検索結果を絞り込むことができる。

### 参考文献

- [1] 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,” 情報処理学会論文誌, Vol.35, No.5, pp.908-916 (1994).
- [2] 荻野, 丸山: “日英機械翻訳システム JETS における日本語解析,” 自然言語処理研究報告 (91-NL-84), 84-17, pp.127-134 (1991).