デジタル・ドキュメント 13－7
（1998. 7. 17）

係り受け情報を利用したWeb上の日本語テキスト検索システム

立石 健二　　大庭 直行　　峯 恒憲　　雨宮真人

九州大学大学院 システム情報科学研究科 知能システム学専攻
E-mail:{tateishi,,ohba,mine,amamiya}@al.is.kyushu-u.ac.jp

本稿では、情報検索システムのランキング手法として、「単語間の係り受け構造」を利用する手法を提案する。本手法を用いることにより、従来のシステムでは捨てられていた文法的な制約条件を生かすことができ、より正確なランキングの実現が期待できる。

実際、本手法に基づいた情報検索システムを作成し評価を行なったところ、従来の統計的手法によるシステムよりも、全体として約11%の適合率の向上が実現できた。また、候補ペアの含有率が70%以上のランキング順位においては、適合率約94%と非常に高精度であり、本手法の有効性を確認できた。

Japanese Information Retrieval System using Syntactic Dependency Relations on the Web

Kenji TATEISHI　　Naoyuki OHBA　　Tsunenori MINE　　Makoto AMAMIYA

Dept. of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University
E-mail:{tateishi,ohba,mine,amamiya}@al.is.kyushu-u.ac.jp

This paper proposes a Japanese Information Retrieval(IR) system which uses syntactic dependency relations between words .We believe this method makes good use of the grammatical conditions that have been ignored.

Experimental results say that our IR system obtains 11% higher precision rate than a convertional one adapting tf · idf method,and also, the precision rate reached about 94% in the case that candidate pairs in a query were included more than 70% in a sentence of a retrieved document.

1 はじめに

今日では、インターネットやCDROMを始めとする電子媒体を通して、大量の情報を手に入れることができるようになった。情報検索システムは、その中からユーザの欲しい情報を高速かつ的確に取り出すことのできるものとして重要度を高めている。

情報検索システムの検索手順を簡単に言うと、クエリーからキーワードを取り出し、文字列マッチングを行いキーワードを含むテキストをユーザに提示する、ということになる。ただし、このように得られた検索結果には、ノイズが多く含まれているため、システムはある一定の評価基準に基づきランキングを行なってから、結果として返す必要がある。従来のシステムでは、評価基準として主にtf・idf手法を採用している。これは統計的手法であり、キーワードの出現数に注目してランキングを行なう。

しかしtf・idf手法を用いた検索システムは、必ずしもユーザにとって使いやすいものとは言えない。その理由として、(1) ランキング上位であっても検索結果にむらがあること、(2) ランキング手法に、通常ユーザにとって未知である、検索対象テキストに依存する統計量を用いていること、が挙げられる。

そこで本研究ではこれらの問題点をふまえ、検索の際「単語間の係り受け構造」を利用する手法を提案する。本手法を用いることにより、(1) 従来のシステムでは捨てられていた「単語間の係り受け構造」という文法的な制約条件を生かすことができ、より正確なランキングが実現できること、(2) クエリー中の文法的な条件に従い行なうことによって、ユーザに対してわかりやすいランキングが実現できる、といった利点が考えられる。

また、本研究では、検索時間についても考慮し、基となる検索対象テキストからあらかじめ索引ファイルと係り受け情報ファイルを作成し、検索時にはこれらを参照することで高速化を実現している。

以下、2節にて、単語間の係り受け構造を利用した情報検索システムの基本動作について述べ、3節で本研究で作成したシステムの構成について説明する。さらに、4節で実験を行ないその結果を考察する。

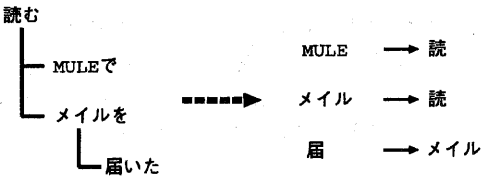
2 基本動作

本検索システムでは、構文解析による係り受け関係を基にして、2段階のランキング手法により検索対象テキスト集合をランク付けする。以下、検索の流れに従い、例を踏まえながらシステムの基本動作を示す。

1. ユーザにクエリーを自然言語で入力してもらう。

   届いたメイルをMULEで読む

2. クエリーを構文解析し、係り受けペアを抽出する。



   構文解析木を、係る単語と係られる単語のペア(係り受けペア)の集合に分解する。
   =(MULE→読)(メイル→読)(届→メイル)

3. 検索対象のテキスト集合から候補ペアを含むテキストを抽出する。
   テキストA ⊃ メイルを家で読んだ。
   =(メイル, 読)
   テキストB ⊃ 以下にメイルをMULEで読む方法について述べる。
   =(メイル, 読)(MULE, 読)
   テキストC ⊃ 本を読んだので、メイルを書いています。
   =(メイル, 読)

ここで抽出するのは、クエリー中の任意の係り受けペアを構成する2単語が1センテンス内に含まれる時の、その2単語(候補ペア)を含むテキストである。すなわち、候補ペアとは、実際に係り受け関係を構成しているかは問わないまでも、その可能性がある単語ペアを言う。なお、説明を容易にするため、ここでは1テキストに、候補ペアを含むセンテンスは1つのみ出現することと仮定している。

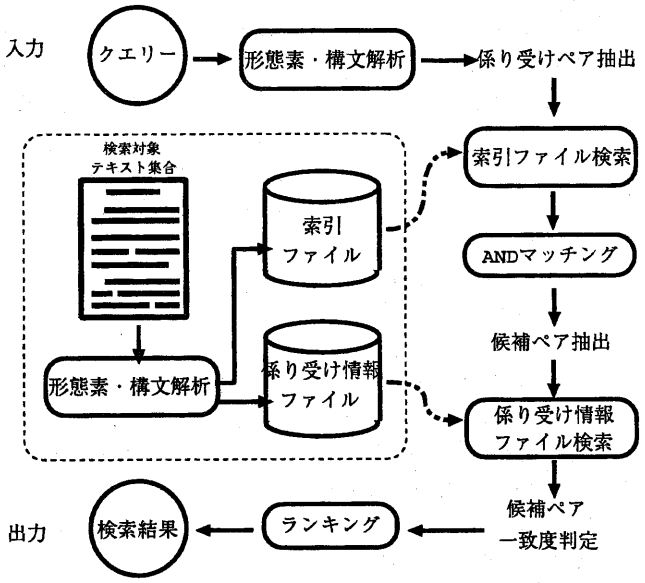


図 1: システムの構成

4. 候補ペアの含有率の高い順にテキストをランク付けする。

   {テキスト A,C}:33% < テキスト B:66%

   第1段階目のランキング手法である。係り受け候補ペアの含有率とは、クエリー中の係り受けペアの個数に対してのセンテンス中に含まれる候補ペアの割合である。

5. 含有率が同じ場合は、係り受け候補ペアの一致度によりランク付けする。

   テキスト C < テキスト A

   第2段階目のランキング手法である。係り受け候補ペアの一致度とは、クエリー中の係り受けペアに対する係り受け候補ペアの一致の度合を言う。実際に候補ペアを構成する2単語が係り受け関係になっていれば、一致の度合は高くなり、そうでなければ低くなる。

以上5つのステップの結果として、B,A,C の順にテキストがランク付けされ、ユーザに提示される。

3 システムの構成

図1にユーザがクエリーを入力してから出力を返すまでの動作を示した。システムはまず、クエリーを構文解析して係り受けペアを抽出する。次に、検索対象テキストから候補ペアを含むセンテンスを抽出して、係り受けペアに対する一致度を判定する。最後に、ランキングを行ない検索結果を HTML 形式で出力する。

3.1 係り受けペアの抽出

クエリーとして自然言語の文を採用し、形態素・構文解析により係り受けペアを抽出する。構文解析木は、必ず係る単語に対し係られる単語は1つである。そこで今回、構文解析木そのものを使うのではなく、係る語・係られる語に注目した2単語の関係を利用している。また、係り受けペア抽出時には、以下の点を考慮している。

- 格助詞を省略 例)「メイルを読む」⇒(メイル→読)「を」を省略
- 活用語は語幹のみ 例) 読む→読

なお、形態素・構文解析を行うため、今回、(株)リコーで開発された簡易日本語解析系 QJP を用いた。QJP は約 50KBytes 程度の辞書しか必要としない、高速な日本語解析系である。[1]

3.2 候補ペアの抽出

次に、検索対象のテキスト集合から、係り受けペアを一つ以上含む可能性のあるセンテンスを抽出する。可能性というのは、係り受けペアを構成する2単語が、あるテキスト内のあるセンテンスに含まれることが条件であり、必ずしも係り受け関係になっている必要はないということである。ここで、抽出されたセンテンス内の2単語を候補ペアと呼ぶとにする。したがって、ここでは、候補ペアを含むセンテンスを抽出すれば良いことになる。

この過程を、全てのテキストに含まれる全てのセンテンスについて全文走査により行うことは、大きなテキスト集合を扱う場合には現実的ではない。そこで、あらかじめ基のテキスト集合から索引ファイルを作成し、検索時には索引ファイルを参照することにしている。以下、3.2.1節で索引ファイルの構成を、3.2.2節で索引ファイルを用いて候補ペアを抽出する方法について述べる。

3.2.1 索引ファイルの構成

まず、基となるテキスト集合から、一つ一つセンテンスを取りだし形態素解析をおこない単語に分割する。その単語に位置情報を付加して登録したのが、索引ファイルである。位置情報は3つの要素からなる。一つは、テキスト ID、次にセンテンス番号、最後にセンテンス先頭からのオフセットである。

図 2に、一つのセンテンスから索引ファイルに登録を行なう例を示した。なお、索引ファイルの容量については、8MBytes のテキストファイルに対して、約 19MBytes である。ただし、ユニバーサル符合化により圧縮を行なうと、約 7MBytes まで圧縮できる。

3.2.2 索引ファイルを用いた候補ペアの抽出方法

候補ペアの抽出は、索引ファイルに登録されている位置情報について、センテンス番号までの AND マッチングを行なうことにより抽出できる。図3 にその例を示す。

```
#1:1=ここでは多くの研究者が日夜、研究を重ねている。
[0,0] ここでは [1,8] 多くの
[2,14] 研究者が [3,22] 日夜、
[4,28] 研究を [5,34] 重ねている。[6,48]
                        ([文節番号:オフセット])

<索引ファイル>
    研究者 → #1:1:14
    日夜   → #1:1:22
    研究   → #1:1:28
    重     → #1:1:34
      (#テキスト ID:センテンス番号:オフセット)

<係り受け情報ファイル>
       [0] [1] [2] [3] [4] [5] [6](文節番号)
#1:1 → 0   8  14  22  18  34  48(文節オフセット)
        5   2   5   5   5   -   -(係り先文節番号)
```

図 2: 索引ファイルと係り受け情報ファイルの構成

```
 研究   → #1:1:28,#2:7:6,#3:3:4,#5:4:2, …
 重     → #1:1:34,#3:4:6,#5:4:12, …
AND マッチ #1:1:(28,34),#5:4:(2,12), …
                              ← 候補ペア
```

図 3: 候補ペアの抽出

3.3 候補ペアの一致度判定

3.2節で抽出した候補ペアの、クエリー中の係り受けペアに対する一致度を判定する。候補ペアが、実際に係り受けを構成していれば、一致度は高く、そうでなければ低くなる。具体的には、タイプ I〜III による3段階の一致度の基準を設けている。表1に、クエリー「メイルを読む」の係り受けペア (メイル→読) に対する候補ペアの、タイプ別一致例を示す。

なお、候補ペアの一致度を判定するためには、係

タイプ	センテンス	含まれる候補ペア
I	この前読んだ雑誌に電子メイルのことが…	-
II	電子メイルを読み書きするには、メイラーを…	(電子メイル , 読み書)
III	新しく届いたメイルを読むには、メニューから…	(メイル , 読)

表 1: 係り受けペア (メイル → 読) に対する候補ペアの一致例

り受け情報を利用する必要がある。しかし、すべての候補ペアを含むセンテンスを構文解析していたのでは検索時間の遅延を招いてしまう。そこで 3.2節と同様に基となるテキスト集合から、あらかじめ構文解析を行なった情報を記憶したファイルを作成することにする。このファイルを係り受け情報ファイルと呼び、その構成を図 2 に示す。なお容量については、8MBytes のテキストファイルに対して、約 4MBytes である。

3.4 ランキング

ランキングは、前節までの過程で抽出されたセンテンスに対し、

1. まず、クエリーに対するセンテンスのスコアを定め、
2. 次に、それを用いてクエリーに対するテキストのスコアを定める。

をすることにより行う。

以下、3.4.1節で「クエリーに対するセンテンスのスコア」について、3.4.2節で「クエリーに対するテキストのスコア」について説明する。

3.4.1 クエリーに対するセンテンスのスコア

「クエリーに対するセンテンスのスコア」、は 2 節で示したように、次の 2 段階の評価基準から求められる。

1. センテンスに含まれる候補ペアの含有率
2. センテンスに含まれる候補ペアの一致度のベクトル

今、クエリーを Q として、センテンス S_iに含まれる候補ペアの含有率を $irate_{s_i}$、候補ペアの一致度のベクトルを $mlevel_{s_i}$ とする。

$irate_{S_i}$は以下の式で定義される。

$$irate_{S_i} = S_i\text{に含まれる候補ペアの数}/|Q|$$

$|Q|$ は、Q に含まれる係り受けペアの総数である。なお、同じ候補ペアは 1 回のみ数えることとする。

次に、候補ペアの一致度については、3.3節で説明している。そのベクトル、$mlevel_{S_i}$については、

$$mlevel_{S_i} = (ma^{I}_{S_i}, ma^{II}_{S_i}, ma^{III}_{S_i})$$

$ma^{I}_{S_i}$: S_iに含まれるタイプ Iの候補ペアの数/$|Q|$
$ma^{II}_{S_i}$: S_iに含まれるタイプ IIの候補ペアの数/$|Q|$
$ma^{III}_{S_i}$: S_iに含まれるタイプ III候補ペアの数/$|Q|$
で定義される。2 つのベクトル $(ma^{I}_{S_i}, ma^{II}_{S_i}, ma^{III}_{S_i})$ と $(ma^{I}_{S_j}, ma^{II}_{S_j}, ma^{III}_{S_j})$ の大小については、

1. $ma^{III}_{S_i}$ と $ma^{III}_{S_j}$ を比べ大きい方が全体として大
2. $ma^{II}_{S_i}$ と $ma^{II}_{S_i}$ を比べ大きい方が全体として大
3. $ma^{I}_{S_i}$ と $ma^{I}_{S_i}$ を比べ大きい方が全体として大
4. 2 つのスコアは等しい。

という手順によって求められる。

これら $irate_{S_i}$ と $mlevel_{S_i}$ を用いると、S_iのスコアを、

$$SC_s(Q, S_i) = (irate_{S_i}, mlevel_{S_i})$$

という 2 次元ベクトルで定義することができる。このとき、$SC_s(Q, S_i)$ と $SC_s(Q, S_j)$ の大小は、

1. $irate_{S_i}$ と $irate_{S_j}$ の大きい方が全体として大
2. $mlevel_{S_i}$ と $mlevel_{S_j}$ の大きい方が全体として大
3. 2 つのスコアは等しい。

という手順によって求められる。

3.4.2 クエリーに対するテキストのスコア

クエリーに対するテキストのスコア $SC_t(Q,D)$ は、前節で求めたクエリーに対するセンテンスのスコア $SC_s(Q,S)$ の集合として、

$$SC_t(Q,T) = \{SC_s(Q,S_i) : T \in S_i\}$$

と定義する。

2つのスコア $A = SC_t(Q,T_1)$ と $B = SC_t(Q,T_2)$ の大小については、

1. AとBからそれぞれ一番大きな要素、aとbを取り出す
2. $a = b = \phi$ なら、2つのスコアは等しい
3. aとbの、大きいほうが全体としても大きい
4. 1. に戻る

という手順によって定める。

以上の方法を用いて、ランキングを実施する。

4 実験

4.1 実験方法

本システムの性能を評価するため、検索実験をおこなった。

検索対象テキストとして、

- 九州大学情報処理教育センターの手引 [2]
- 九州大学情報処理基礎教育の講義資料 [3]

を用いた。その内容は、約8MBytesのHTML形式のテキストファイルである。

クエリーとして、150の自然言語の文を採用した。それらは、

- ユーザが実際にシステムに検索文として入力した43文と、
- 検索対象テキスト中から任意抽出した107文

からなる。

実験方法として、検索結果の上位10位までの適合率を計算する方法を用いた。適合率とは、検索されたテキストのうちでユーザの要求と合致するテキストの割合をいう。この場合、例えば上位10個のテキストの中でユーザの要求に見合うテキストが5個あれば、適合率は50%ということになる。

また、従来のシステムとの比較という観点から、ランキング手法にtf・idf法を使用したシステムの適合率も計算した。次の4.2節で、tf・idf法について説明する。

4.2 tf・idf法

tf・idf法は、現在多くの検索システムで採用されている代表的な手法であり、統計量により検索対象テキストをランク付けする。

tfは、term frequencyの略であり、テキスト内に指定の単語がいくつ出現するかを示した値である。一方、idfは、inverse document frequencyの略で、指定の単語が出現するテキストの割合、の逆数である。

テキスト t におけるクエリー中の単語 x の重要度 $w_{t,x}$ は、tf・idfの積として以下の式で表せる。

$$w_{t,x} = f_{t,x} \times \log \frac{N}{f_x}$$

ここで、$f_{t,x}$ は t 内で x の出現する頻度であり、f_x は x の出現するテキストの数である。Nはテキストの総数である。

$w_{t,x}$ を用いて, クエリーに対する、テキスト t のスコア SC_t は、

$$SC_t = \frac{\sum_{i=1}^{X} w_{t,i}}{\sum_{i=1}^{X} {w_{t,i}}^2}$$

で表せる。X はクエリー中の単語の総数である。

4.3 実験結果

- 本検索システムの適合率 58.04%
- tf・idf 法を使用した検索システムの適合率 46.55%

検索結果の内分けとして、ペア候補の含有率に注目した実験の結果を表2に示す。表中の「単文」とは、本システムがクエリーを解析した結果、唯一の係り受けペアしか抽出しなかった文、「複文」とは2つ以上の係り受けペアを抽出した文を表す。また、検索結果は「総数：正解数」で表した。

	単文 (91 文)		複文 (59 文)		合計 (150 文)	
ペア候補含有率	検索結果	適合率	検索結果	適合率	検索結果	適合率
90%～	411 : 383	93.19%	110 : 108	98.18%	521 : 491	94.24%
70%～90%	0 : 0	-	2 : 2	100%	2 : 2	100%
50%～70%	0 : 0	-	239 : 123	51.46%	239 : 123	51.46%
0%～50%	304 : 108	35.53%	270 : 48	17.78%	574 : 156	27.18%
0%～100%	715 : 491	68.67%	629 : 289	45.95%	1344 : 780	58.04%

	検索結果	適合率
tf・idf 法	1422 : 662	46.55%

表 2: 候補ペアの含有率に注目した実験結果

ランキング順位	検索結果	適合率
上位	523 : 422	80.69%
下位	899 : 240	26.70%
合計	1422 : 662	46.55%

表 3: tf・idf 法を利用した場合のランキング順位に注目した実験結果

表より、ペア候補の含有率が 70%を境に、適合率が大幅に変化していることがわかる。計算すると、70%以上での適合率は 94.26%、70%以下での適合率は 34.32%である。本検索システムではペア候補の含有率が高い程大きなスコアが付くため、ランキング上位では従来のシステムよりも検索精度が高いことが予測できる。

このことを示すために、tf・idf 法による検索システムの結果を、ランキング順位に従い分類した。表 3に示す。表中の上位とは、同クエリーで検索した場合に、本検索システムにおいてペア候補の含有率が 70%以上で検索されるランキング順位をいう。下位についても、ペア候補の含有率が 70%以下であることを示す以外同様である。この表より、上位における適合率は約 80.69%である。本検索システムと比較すると 13.57%適合率の差がある。したがって、本検索システムは、候補ペアの含有率が 70%以上のランキング上位で、従来のシステムと比較して優位性があることがわかる。

4.4 検索例

本システムにクエリーとして「届いたメイルを MULE で読む」と入力し、検索した様子を図 4に示す。

検索結果はマッチしたテキスト名が優先順位順に並べられ、その頭にクエリーとテキストとのマッチ度に応じて「金・銀・銅」のメダルが表示される。テキスト名の隣には「score」とあり、これは候補ペアの含有率を表している。また、各テキストの下には、実際にクエリーとマッチしたセンテンスが表示される。

5 おわりに

今回、ユーザにとってより使いやすい情報検索システムという観点から、単語間の係り受け構造を利用した検索手法を提案した。そして、システムを実際作成して実験を行ない、tf・idf 手法との比較をした。実験結果からは、本検索システムについて、次の 2 つのことがわかった。

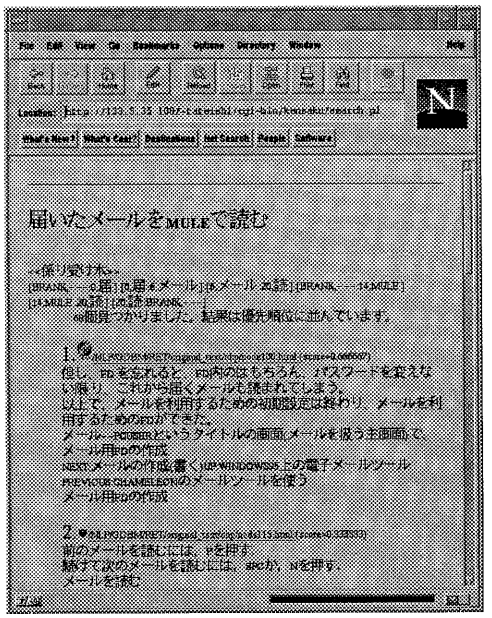


図 4:「届いたメイルを MULE で読む」の検索結果

1. システム全体で、tf・idf 法よりも約 11%適合率が向上

2. ペア含有率が 70%以上のランキング上位では適合率が約 94%であり、tf・idf 法と比較して高い検索精度を実現

今後は、係り受けの情報をより効率的に利用する方法を考え、さらなる検索精度の向上を達成するよう努めていきたい。

参考文献

[1] Masayuki KAMEDA, "A Portable and Quick Japanese Parser - QJP", Proc. of COLING'96, pp 616–621, 1996

[2] 九州大学情報処理教育センター 編," 九州大学情報処理教育センター利用の手引 (1997 年版)", 九州大学出版会,1997

[3] 廣川 佐千男, 宮原 哲浩, 峯 恒憲, 正代 隆義, "12 回で学ぶ情報処理", 九州大学生協, 1997

[4] 麻生和昭, " 単語間の係り受け構造を利用した WWW 上での日本語テキスト検索システム ", 九州大学大学院システム情報科学研究科 修士論文,1997

[5] 原田晶紀 : " サーチエンジン徹底活用術", オーム社開発局,1997