1 はじめに

入力文の持つ言語情報を有効に利用し、ユーザが求める情報を正確かつ高速に提供する情報検索システムの開発が求められている。自然言語で入力可能な情報検索システムは、従来のキーワード入力型のシステムと比較してユーザの検索要求よりも明確に表現できるという利点がある。したがって、入力文が持つ言語情報を有効に利用すれば高い检索精度を得ることが期待できる。

本研究では、検索時の制約条件として単語間の係り受け関係を用いる。構文解析結果から抽出した動詞及びそれに係る名詞句を1つのフレームとして定義し、システムは入力文と検索対象文書内のフレームの同一性を判定する。そして、一致すると認められる文書の检索結果としてユーザに提示する。

また、我々が日常的に認めている単語についてはシステムが自動的に拡張することが望ましい。そのため、同一性の範囲に個々の単語の同意語も含めていきます。さらに、検索対象文書内で名詞が省略される場合を考慮して、照応関係にある名詞も範囲内とした。

本稿では、このフレーム関係を利用した検索手法について提案するとともに、この手法の性能について、従来の論理演算子 (AND, OR 等) を用いた手法との比較実験の結果をもとに、検索精度および検索時間の両点から議論する。

2 文の構造化

入力文及び検索対象文書内各文からフレーム構造を抽出する。フレームは、動詞を中心に、それに係る名詞との繋がりを記述したものです。

個々のフレームは、動詞及びそれに係る一つ以上の名詞句で構成される。名詞句は、名詞と助詞から構成される。助詞として、格助詞 (が, が, に, が) 及び、接続助詞 (を, て, に) を用いる。また、「〜に関して」「〜に対して」のように動詞を含むが慣用的に格助詞のように扱われるものは、格助詞相当句 [1] として扱う。

動詞には、サ変動詞の名詞的用法 (例, 場合の低下) も含める。この場合、直前の助詞 [の] を含む名詞句がフレームの構成要素となる。また名詞は、直前に並列接続助詞 (と, が, が) を含む名詞句が存在する場合に複数存在することになる。図1に入力文からフレーム構造への変換例を示す。

3 同一性の判定

入力文と検索対象文書内の各文を比較し、同一又は類似のフレーム構造を持つ文を含む文書を正解とする。フレーム同一及び類似の定義は以下の通りである。なお、フレーム類似の正解範囲はフレーム同一の正解範囲を含んでいる。

- フレーム同一
  同一の動詞、及びそれに係る同一の名詞句が存在する。

- フレーム類似
  同意的動詞、及びそれに係る同意的名詞句が存在し、照応関係で存在する。

ただし、名詞句内の助詞の同一性については、格助詞の揺れや動詞の態の変化等を考慮して一定の範囲を設けた。助詞相互間のルールを作成し、適用了ルール内に含まれている場合は正解とする。作成したルールの例は表1に示す。厳密には動詞の違いにより取り得る格助詞の種類は変化し、その用法も異なるため今回のような一套のルールは定義できないと思われる。そのため、ルールによる制約は緩いものになっている。

また、複合名詞については部分一致でもよいとする。フレーム類似の定義における同意語、照応関係については次の3.1節, 3.2節で説明する。
<table>
<thead>
<tr>
<th>種類</th>
<th>条件</th>
<th>ルール</th>
</tr>
</thead>
<tbody>
<tr>
<td>種別の数変化</td>
<td>類が異なる</td>
<td>→ もっとも</td>
</tr>
<tr>
<td>情報の数変化</td>
<td>類が異なる</td>
<td>→ へ から へ</td>
</tr>
<tr>
<td>情報の数変化</td>
<td>類が異なる</td>
<td>→ 十 から へ</td>
</tr>
<tr>
<td>動詞の数変化</td>
<td>類が異なる</td>
<td>→ へ から へ</td>
</tr>
<tr>
<td>動詞の数変化</td>
<td>類が異なる</td>
<td>→ 十 から へ</td>
</tr>
<tr>
<td>動詞の数変化</td>
<td>類が異なる</td>
<td>→ 十 から へ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>項目</th>
<th>参考文献</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 同意語</td>
<td>[文献6]</td>
</tr>
<tr>
<td>3.2 照応関係</td>
<td>[文献7]</td>
</tr>
</tbody>
</table>

表1 助詞相互間のルールの例

3.1 同意語

我々が日常近い意味として扱っている単語については、同一の範囲内にあるとする。同意語の検索には、既存の概念辞書を利用してある。今治は、EDR電子辞書[7]内の概念辞書（以下EDR概念辞書）を使用した。EDR概念辞書は、単語の間の関係を木構造で表現しており、みに近い節点概念の具体度は高い傾向にある。同一語はいずれかの節点に所属し、同一の節点に所属する単語（以下、同概念語）には主にハタカナ巡り語や単語の英語読みが登録されている。

EDR概念辞書では、非常に小さい意味の違いを異なる概念として捉えているため、日常的という観点からは同概念語のみではなく、近接の他の概念に所属する単語も含めた方が望ましいと思われる。そこで今治は、基の単語が所属する概念の親の節点に所属する単語（以下、上位概念語）と、子節点の節点に所属する単語（以下、下位概念語）も同意語の範囲を含めることにした。

ただし、上位概念語については文献[6]で示されているように、展開する範囲を制限しないと検索精度が低下する。そこで今治は、次の2つの条件を満たす上位概念語のみを展開することにした。

- 基の単語が所属する概念の深さとの差がN以下
- 基の単語と拡張された単語との間に重みは設定していない。

3.2 照応関係

日本語では、冗長な表現を防ぐために一度表記した要素を次回からはゼロ代名詞として省略する傾向がある。フレーム同一の条件下では、動詞に係る名詞が明示的文でも表記されている場合にはその名詞は省略され、依存関係を構成せずに不一致となる。

ゼロ代名詞の照応文を推定する方法としてはcentering theory[2][3]がある。照応の判定のみの複雑な処理を行うことなく新正解率が確認されている。centering theoryの基本原則は、次の通りである。

- ある文内のゼロ代名詞の照応文は、その前の文内で最も中心的な役割を果たしている事項（以下cf）である。

step1: 以下に該当する場合は、A と B は照応関係にないとして終了する。

1. B の後の文に A が存在する。

step2: A が第1文の TOPIC、SUBJECT、OBJECTの場合、A と B は照応関係にあるとする。

step3: B の前のA番文の文までを調べて A が TOPIC、SUBJECT、OBJECT のいずれかであれば A と B は照応関係にあるとする。

step1 の2の判定には上記で作成したルールを使用する。

step2で第1文を特別視するのは、新聞記事等の文書では第1文にその文書の内容を要約する文が位置することも多く、照応文が存在する可能性も高いからである[8]。step3 の A の値については実験で定める。

4 システムの動作

今回作成したシステムは、まず日本語文を入力して受取り、形態素・構文解析を行う。今回日本語文の形態素・構文解析に(株)リコーラで開発された簡易日本語解析系QJP[5]を用いた。QJP は、約 50KB 程度のノード基盤の辞書が必要で、まず 700～800 語/秒 (WS、Sun-SPS20) といった高速度を持つ。次に、QJP の出力からフレーム構造を作成する。この際、EDR概念辞書を用いて同意語を展開しておく。

EDR概念辞書の規則は約 67MB であり、キャッシュメモリに格納されている。

検索対象文書集合は、検索以前にあらかじめ QJP ですべての文を形態素・構文解析を行い、2次ファイルを作成しておく。

システムの動作は、次に、索引ファイルを参照しフレーム内の名詞・動詞について、検索対象文書文での位置情報を得る。索引ファイルは検索対象文書文でのすべての単語とその出現位置をハッシュ構造で記憶した二次ファイルである。1つの単語の位置を示す情報として3パラメータを与える。単語が出現する文の番号、文の先頭から数えた単語の位置をオフセットである。

検索処理は2段階である。まずフレーム内の名詞・動詞の論理的結合を AND・OR による論理演算式で表記

---318---
し、検索対象文書を数え込む。次に、残った検索対象文書をフレームの同一性を判定する。
フレーム同一と類似どちらの条件を正解範囲とするかはユーザが指定できる。検索対象文書内各フレーム
構造は、あらかじめ2次ファイルとして記憶してある。
システムは出力として、同一と判定されたフレーム
を持つ文を含む文章をユーザに提示する。結果の順位
付けは行わない。

5 実験
5.1 実験方法
実験対象テキスト集合として、情報検索システム評
価用テキストコレクション BMIR-J2 を利用した [4]。テ
キスト集合として1994年の毎日新聞の記事を採用し、
検索対象テキスト件数は5080件、約5MByteからなる。
用意されているテキスト用検索要求文50文は6グループ
に分類されている。今回使用した検索要求文は、グループ
(C)(D)(構文解析機能及言語知識を必要とする)に
所属する検索要求文の中で正解数が6-50の範囲内にあ
る16文である。
比較対象の検索手法としてはAND検索を採用し、こ
れは入力文を単語に分割してすべての単語を含む文書を検
索結果とする手法である。形式解析の精度や索引ファ
イルの構成の違いが結果に影響しないように、AND検
索を行うシステムは各検索システムが中間結果として
出力する論理演算の結果を利用した。
検索精度の評価式は、情報検索の分野で通常利用さ
れているRecall値、Precision値を用いた。

Recall = システムが出した正解文書件数
全正解文書件数

Precision = システムが出した正解文書件数
システムが出力した文書件数

本検索システムは同一性の判定方法としてフレーム
同一及び類似の2条件があるが、それぞれ別個の評価
を行った。以下、結果を示す。

5.2 実験結果
まず、同一性の判定方法にフレーム同一を抽出した場
合の検索精度の結果を表2に示す。Precision値はAND
検索と比較して約20%上昇している。係り受け関係と
いう制約条件が有効に作用していることがわかる。
フレーム同一の全検索数56件の中で10件が正解となっているが、内9件までが入力文の否定表現を原因とす
る(例)赤字国債の発行を避ける)。
Recall値はAND検索よりも低下しているが、フレーム
類似はこの差を縮めることを目的としている。ここ
で、まず同一性の数値計算における3.1節で述べた概念
解釈の適用範囲を定める。表3は、同一性の判定にフレ
ーム類似を抽出した場合で、かつ同一性の展開のみ
を行った場合の検索精度を示す。1入力文当たりの同一
性展開文書数の平均は、下位概念語で18文、上位概念
語で3文(N=1)であった。Recall値では上位と下位
概念語の両方を展開した場合が16.7%で高い。しかし、
上位概念語は平均展開文書数は3文で少ないが、抽象度
が高いため出現回数が多く少しでも誤った文を展開す
るとPrecision値を著しく下げる傾向がある。ここでも
Precision値は35.2%であり下位概念語のみを展開した
場合と44%の差がある。その意味、同一性の展開は下
位概念のみ(N=0)とすることにした。

表2: フレーム同一とAND検索の検索精度

<table>
<thead>
<tr>
<th></th>
<th>フレーム同一</th>
<th>AND検索</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (正解/全正解)</td>
<td>13.7% (46/336)</td>
<td>29.2% (95/336)</td>
</tr>
<tr>
<td>Precision (正解/全検索)</td>
<td>82.1% (46/56)</td>
<td>62.0% (95/158)</td>
</tr>
</tbody>
</table>

表3: フレーム類似(同一性のみ)の検索精度

<table>
<thead>
<tr>
<th></th>
<th>同意語展開なし</th>
<th>同意語展開あり</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall (正解/全正解)</td>
<td>13.7% (46/336)</td>
<td>14.9% (50/336)</td>
</tr>
<tr>
<td>Precision (正解/全検索)</td>
<td>82.1% (46/56)</td>
<td>79.4% (50/63)</td>
</tr>
<tr>
<td>N=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall (正解/全正解)</td>
<td>16.7% (56/336)</td>
<td>16.7% (56/336)</td>
</tr>
<tr>
<td>Precision (正解/全検索)</td>
<td>35.2% (56/159)</td>
<td>35.2% (56/159)</td>
</tr>
</tbody>
</table>

表4: フレーム類似(同一性のみ)の検索精度

次に、3.2節で述べた照応格の推定範囲Mを定める。
表4に同一性の判定にフレーム類似を抽出した場合で、
かつ照応関係の判定のみを行った場合の検索精度を示
す。M=4とした時がPrecision値84.1%、Recall値
17.3%で共に最高値をとる。Recall値の上昇は約4%で
あるが、正解数は46件から57件へ2割程度上昇しており、
しかも新たに検索された文書のすべてが正解である。
Precision値を下げることなくRecall値を高めることの
できる点は評価できると思われる。推定範囲はM=4と
する。
以上から決定した範囲に基づいて、同一性の判定にフ
レーム類似を抽出した場合の検索精度を5に示す。

表5: フレーム類似(同一性+照応関係)とAND検索

<table>
<thead>
<tr>
<th></th>
<th>フレーム類似</th>
<th>AND検索</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (正解/全正解)</td>
<td>18.8% (63/336)</td>
<td>36.0% (121/336)</td>
</tr>
<tr>
<td>Precision (正解/全検索)</td>
<td>74.1% (63/85)</td>
<td>42.6% (121/284)</td>
</tr>
</tbody>
</table>

以上の結果は、フレーム同一と比較して約5%上昇

―319―
表 4: フレーム類似（照応関係のみ）の検索精度

<table>
<thead>
<tr>
<th>照応解析</th>
<th>照応格の推定範囲</th>
</tr>
</thead>
<tbody>
<tr>
<td>なし</td>
<td>M = 0</td>
</tr>
<tr>
<td>Recall</td>
<td>13.7%</td>
</tr>
<tr>
<td>Precision</td>
<td>82.1%</td>
</tr>
<tr>
<td>(正解/全検索)</td>
<td>(46/56)</td>
</tr>
</tbody>
</table>

5.3 検索時間

検索時間の測定は、約 2 万件の新聞記事 20MB に対し
て、検索精度で実験と同様の検索要求文を数回測定し
た値の 1 検索要求文当たりの平均をとることにより行っ
た。使用した計算機は、SUN Ultra-1（主記憶 96MB）で
ある。表 6 に AND 検索及び本検索システムの検索時間
と、10MBbyte 当たりの検索時間の増加量を示す。

AND 検索とフレーム同一致を比較すると、フレーム同
一の 20MBbyte 時の検索時間は 130msec。AND 検索
では 32msec で約 4 倍の差がある。また 10MBbyte 当
たりの検索時間の増加量はフレーム同一で 44msec、AND
検索で 4msec であり、フレーム同一は検索対象
文書集合の規模に対する検索時間の依存度が大きいこ
とがわかる。フレーム同一では、名詞と動詞が同一文
内に存在する場合には、フレーム構造ファイルを読み
出し依存関係の判定を行う必要があり、AND 検索
と検索時間の差はここから生じている。

6 おわりに

今回、構文解析結果から抽出した動詞及びそれに係
る名詞句を 1 つのフレームとして定義し、入力文のフ
レームと同一又は類似のフレームを持つ文書を検索す
る手法を提案した。実験結果から、フレーム同一での
Precision 値は 82% で AND 検索よりも 20% 高いことが、
フレーム類似で照応関係を適用した場合に Precision
値を低下させることなく Recall 値を約 4% 上昇できるこ
とを示すことができた。同一語の展開方法については
今後検討する予定である。

謝辞

本研究では、(社) 情報処理学会・データベースシス
テム研究会が、新情報処理開発機構との共同作業によ
り、毎日新聞 CD-ROM94 データを基に構築した情報
検索システム評価用テストコレクション BMIR-J2、
ならびに简易日本語解析系 QJP を利用した。毎日新聞
社ならびに BMIR2 の開発に携わられた方々に感謝し
ます。また、QJP の使用を許可して頂いた株式会社リ
コーと開発者の亀田氏に感謝します。

参考文献

Discourse and the Process of Centering”, Computational
[4] 木谷ほか：日本語情報検索システム評価用テストコレク
拡張による検索支援”, 言語処理学会, 第 3 回年次大会論
[7] (株) 日本電子辞書研究所, EDR 電子辞書仕様説明
書 (第 2 版), 1995.
[8] 中岩 浩己, 池原 悟, “日英機械翻訳における用言意味
性を用いたゼロ代名詞照応解析”, 情報処理学会論文誌,