係り受け構造や語の意味情報を利用した日本語テキスト検索システム

立石 健二　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
{tateishi,mine,amamiya}@al.is.kyushu-u.ac.jp

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をより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検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み検索結果としてユーザに提示する。

また、我々が日常的に同意と認識し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る。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章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ある。

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を構成要素とする。名詞句は、名詞と助詞から構成される。助詞として、格助詞 (ガ、ヲ、ニ等) 及び、提題助詞 (ハ)、取り立て助詞 (モ等) 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、価格の低下) 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 (ト、ヤ、カ等) を含む名詞句が存在する場合には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。

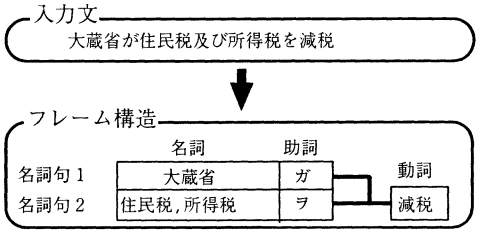

図 1: 入力文からフレーム構造への変換

3 同一性の判定

入力文と検索対象文章内の各文を比較し、同一又は類似のフレーム構造を持つ文を含む文章を正解とする。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も含んでいる。

- 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
  同意の動詞、及びそれに係る同意の名詞句が存在又は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、適用した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を表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律な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内における同意語、照応関係については次の3.1節、3.2節で説明する。

種類	条件	ルール
動詞の態の変化	態が異なる	をーがーに
格助詞の用法に重複	-	にーへ　からーより
格助詞から提題助詞	-	{が,を}ーは
格助詞から「の」	サ変名詞のみ	{が,を,に}ーの

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する。今回は、EDR 電子化辞書 [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、概念間の繋がりを木構造で表現しており、葉に近い節点程概念の具体度は高い傾向にある。各単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) には主にカタカナ揺らぎ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も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。そこで今回は、基の単語が所属する概念の親側の節点に所属する単語 (以下、上位概念語) と、子側の節点に所属する単語 (以下、下位概念語) も同意語の範囲に含めることにした。

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。そこで今回、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。

- 根から数えた深さが 7 番目以下
- 基の単語が所属する概念の深さとの差が N以下

Nの値については、実験で定めることにする。なお、基の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、動詞に係るべき名詞が前方の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。

ゼロ代名詞の照応格を推定する方法としては centering theory[2][3] がある。助詞の判定のみの複雑な処理を行うことなく高い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格は、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。
- cf へのなりやすさは、文に含まれる格要素に基づいて、TOPIC(ハ格) > SUBJECT(ガ格) > OBJECT(ヲ格、ニ格) > OTHERS の順序となる。

この具体的なアルゴリズムを以下に記す。今、入力文「名詞 [A]+助詞+動詞 [B]」(例) マンションを販売する) に対し、検索対象文書内で名詞 A と動詞 B が同フレーム内に存在せず、離れて存在するとする。

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。
2. B の格要素がすでに埋まっている (例.[A] マンション … ホテルを [B] 販売)

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 M番目の文までを調べて 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 の判定には表 1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が多く、照応格の存在する可能性も高いからである [8]。step3 の M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として受取り、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 程度の小規模な辞書しか必要せず、また 700～800 語/秒 (WS、Sun-SS20) という高速性を持つ。

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 概念辞書の規模は約 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、検索対象文書内での位置情報を得る。索引ファイルは検索対象文書内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 パラメータを与えている。単語が出現する文書番号、文書の先頭から数えた文番号、文の先頭から数えた単語のオフセットである。

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

し、検索対象文書を絞り込む。次に、残った検索対象文書内各文と入力文との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。検索対象文書内各文のフレーム構造は、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)(構文解析機能及び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。形態素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本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、Precison 値を用いた。

$$\text{Recall} = \frac{\text{システムが出力した正解文書件数}}{\text{全正解文書件数}}$$

$$\text{Precision} = \frac{\text{システムが出力した正解文書件数}}{\text{システムが出力した文書件数}}$$

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選択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 件が誤りとなっているが、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。そこで、まず同意語の展開時における 3.1節で述べた概念辞書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選択した場合で、かつ同意語の展開のみを行った場合の検索精度を示す。1 入力文当たりの同意語展開語数の平均は、下位概念語で 18 語、上位概念語で 3 語 ($N = 1$) 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語数は 3 語と少ないが、抽象度が高いため出現回数が多く少しでも誤った語を展開すると 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。そのため、同意語の展開は下位概念のみ ($N = 0$) とすることにした。

	フレーム同一	AND 検索
Recall (正解/全正解)	13.7% (46/336)	29.2% (98/336)
Precision (正解/全検索)	82.1% (46/56)	62.0% (98/158)

表 2: フレーム同一と AND 検索の検索精度

	同意語展開なし	上位概念語の適用範囲 $N = 0$ 下位概念のみ	上位概念語の適用範囲 $N = 1$
Recall (正解/全正解)	13.7% (46/336)	14.9% (50/336)	16.7% (56/336)
Precision (正解/全検索)	82.1% (46/56)	79.4% (50/63)	35.2% (56/159)

表 3: フレーム類似 (同意語のみ) 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選択した場合で、かつ照応関係の推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。Precsion 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いて、同一性の判定にフレーム類似を選択した場合の検索精度を 5に示す。Recall 値は、18.6%で表 2のフレーム同一と比較して約 5%上昇

	フレーム類似	AND 検索
Recall (正解/全正解)	18.8% (63/336)	36.0% (121/336)
Precision (正解/全検索)	74.1% (63/85)	42.6% (121/284)

表 5: フレーム類似 (同意語+照応関係) と AND 検索 (同意語の展開あり) の検索精度

	照応解析なし	照応格の推定範囲					
		$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$
Recall (正解/全正解)	13.7% (46/336)	16.1% (54/336)	16.7% (56/336)	17.0% (56/336)	17.0% (57/336)	17.3% (58/336)	17.3% (58/336)
Precision (正解/全検索)	82.1% (46/56)	84.4% (54/64)	84.8% (56/66)	83.8% (56/66)	83.8% (57/68)	84.1% (58/69)	81.7% (58/71)

表 4: フレーム類似 (照応関係のみ) の検索精度

している。Precision 値は、74.1%で AND 検索よりも約 28%上昇しているが、フレーム同一よりも約 8%低い。表 4で示したように照応解析のみでも Recall 値 17.3%が保証されるので、それに同意語の展開を加えても約 1%しか Recall 値が上昇していないことになる。Precision 値の低下を含めて考えると、同意語の展開が有効に機能していなかったと判断できる。BMIR-J2 では展開語として、実際の企業名等の固有名詞を要求する場合が多く、今回使用した概念辞書では対応できなかったと思われる。

以上の検索精度に関する実験結果から、フレームの同一性の判定条件にフレーム類似を選択し、かつ同意語の展開は行なわずに照応解析のみを行なった場合が最も効果的であることがわかる。

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、検索精度での実験と同様の検索要求文で数回測定した値の 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、10MByte 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を比較すると、フレーム同一の 20MByte 時の検索時間は 130mmsec、AND 検索では 32mmsec で約 4 倍の差がある。また 10Mbyte 当たりの検索時間の増加量はフレーム同一で 44mmsec、AND 検索で 4mmsec 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て依存関係の判定を行なう必要があり、AND 検索との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。同意語の展開方法については今後検討する予定である。

検索条件		検索時間 [mmsec]	増加量 [mmsec /10M]
AND 検索		32	4
フレーム同一		130	44
フレーム類似	照応関係のみ	278	120
	同意語のみ	991	294
	同意語+照応関係	1320	512

表 6: 検索時間

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-ROM'94 データ版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、ならびに簡易日本語解析系 QJP を利用した。毎日新聞社ならびに BMIRJ2 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

[1] 益岡 隆志, 田窪 行則, "基礎日本語文法" , くろしお出版

[2] Megumi Kameyama, "A Property-Sharing Constraint in Centering", Proc. of ACL-86, pp.200-206, 1986.

[3] Marilyn Walker, Masayo Iida, Sharon Cote,"Japanese Discourse and the Process of Centering", Computational Linguistics, vol.20, No.2, pp.193-232, 1994.

[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ション BMIR-J2, 情処研究会報告 98-DBS-114-3(1998)

[5] Masayuki Kameda,"A Portable & Quick Japanese Parser:QJP", Proc. of COLING'96, pp.616 - 621, 1996

[6] 太田 千晶, 奥村 学, "EDR 電子化辞書を用いたクエリー拡張による検索支援", 言語処理学会, 第 3 回年次大会論文集, pp.373-376, 1997.

[7] (株) 日本電子化辞書研究所, EDR 電子化辞書仕様説明書 (第 2 版), 1995.

[8] 中岩 浩巳, 池原 悟," 日英機械翻訳における用言意味属性を用いたゼロ代名詞照応解析", 情報処理学会論文誌, Vol.34, No.8, pp.1705-1715, 1993.