日本語名詞句のパラフレーズ検索に関する研究

手塚 芳樹　　橋本 泰一　　徳永 健伸　　田中 穂積

東京工業大学 大学院 情報理工学研究科

{**tezuka , taiichi , take , tanaka**}**@cl.cs.titech.ac.jp**

1 はじめに

Web を中心として存在している膨大な量の情報を有効活用するためには，高度な検索の技術が求められる．その技術の 1 つとして，パラフレーズが挙げられる．

近年，パラフレーズの研究は盛んに行なわれており，それらの研究をまとめた文献も執筆されている [1] が，まだ十分な成果があげられていない分野である．パラフレーズの技術としては，大量のコーパスを用いる方法が多く提案されている．計算機の発展により，大規模なデータに対し，わずかな時間での処理が可能になったことが一因に挙げられる．

本論文では，木村 [2] が行なった研究を基に，新聞記事コーパスを用いたパラフレーズ検索の手法を提案する．対象とする入力は日本語名詞句である．漢字をインデックスとした検索を用いることにより,「学校に通う」⇔「通学する」のような単語の置き換えなどでは不可能な言い換えを行なうことが可能になる．

2 システムの概要

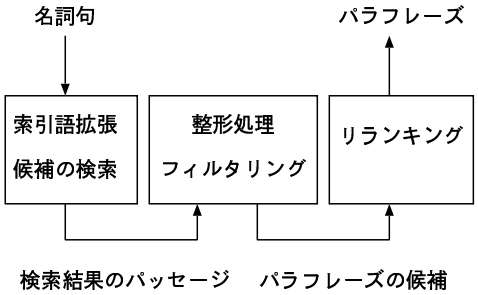

図 1: 処理の流れ

木村のシステムによる処理の流れを図 1 に示す．主に，以下の 3 つのステップから構成される．

1. パラフレーズの候補となるパッセージを検索する．
2. それらの候補に対して，構文的，意味的制約に基づいてフィルタリングを行なう．
3. 候補をリランキングする．

それぞれのステップについて，順に概説する．

2.1 候補の検索

検索対象となるのは，新聞記事を句読点や記号で区切ったパッセージである．それぞれのパッセージに対し，JUMAN[3] によって名詞，動詞，形容詞，副詞，未定義語とされた形態素に出現する漢字をインデックスとする．ただし，カタカナ語およびアルファベットに関しては，そのままの形でインデックスとする．また数詞に関しては，<su>という特別なインデックスを用いる事により，そのパッセージに数詞が含まれているという情報だけを残す．

入力名詞句に対しては，NTT シソーラス [8] を利用して索引語拡張を行なう．これは，入力に含まれる各形態素 *m* について，次の手順を実行することにより行なう．ただし，ここで言う形態素とは，助詞や助動詞などは含まない．また，カタカナ語に対しては，索引語拡張を行なわない．

1. 形態素 *m* が属する意味クラスを参照する．
2. その意味クラス内に出現する漢字の頻度をカウントする．
3. 形態素 *m* 中のどの漢字よりも頻度が大きい漢字を索引に追加する．

各インデックスに対する重みは，意味クラス中での頻度の対数により 100 を分配した値を付ける．

2.2 フィルタリング

検索された候補に対して，次に説明する意味的制約と構文的制約を用いてフィルタリングを行なう．

- **意味的制約：** 候補は，入力中の概念をすべて含んでいなければならない．候補に対して KNP[4] を用いて係り受け解析を行ない，入力中の各概念が，どの文節に現れているかを調べ，入力中のすべての形態素に対応するインデックスを持っているかチェックする．
- **構文的制約：** 検索されたパッセージは，新聞記事を句読点などで区切っただけなので，ここから必要な文節だけを切り出す必要がある．そのとき，切り出された句は，係り受け関係が 1 つの木構造

で表されなければならない．またこのとき，終端が適切な形となるように整形する．

2.3 リランキング

漢字インデックスによる検索により多数の候補が得られるが，この順位がパラフレーズとしての良さを表しているとは言い難い．そこで，以下の 3 つの情報を用いて，候補をリランキングする．

- **検索時のスコア：** これだけではパラフレーズとしての良さを表しているとは言い難いが，各インデックスに対して重み付けを行なっている点から，評価尺度の 1 つとして用いることは妥当である．
- **係り受け距離：** 2 つの文節の係り受け距離が短いほど，両者の意味的結合度は強いと考えられる．したがって，入力中で隣り合っている形態素の概念を含むそれぞれの文節の係り受け距離を評価尺度の 1 つとして用いる．
- **文脈情報：** 候補の文脈は，その元となっている記事とする．入力に対しては，元となっている記事は存在しない．しかし，検索結果中に，入力文を完全に含むパッセージが存在することがある．そのパッセージの元となっている記事を入力の文脈とする．ただし，そのようなパッセージが多数見つかった場合，最大 10 件の記事を利用する．それぞれの文脈となる記事中の形態素を $tf \cdot idf$ で重み付けをし，それらの余弦の値を文脈の類似度とする．この値を評価尺度の 1 つとして用いる．

2.4 実験，評価

入力として，BMIR-J2[5] の検索要求文から選んだ 53 の名詞句を利用する．検索対象には，1991 年から 1993 年までの毎日新聞記事 [6] を使用した．検索には，GETA[7] を利用した．1 入力に対し，最大 50 件を出力し，その結果について，2 人の判定者が評価を行なったところ，精度は 10%程度であった．

3 システムの評価と改良

本節では，木村の手法の問題点を挙げ，それらを改良する方法について述べる．本研究で改良を行なった点を以下に挙げる．

- 検索対象のパッセージ作成
- 漢字拡張とインデクシング
- 係り受け関係によるフィルタリング
- 文脈の利用
- 候補を用いたフィルタリング

以下の項で，具体的な改良法について，それぞれ説明する．

3.1 検索対象のパッセージ作成

木村の手法では，記事の分割は単に句読点や記号で区切るだけである．この方法では，本来係り受けが成立していないパッセージが作成される可能性がある．後の処理で，何らかの形に構文解析されるが，このようなパッセージからは意味のあるパッセージが得られないと考えられる．したがって，構文解析を検索されたパッセージに対して行なうのではなく，パッセージを作成する前に実行する．区切りの中でその係り先が現れない文節が存在した場合は，その文節の直後でさらに分割する．これにより，作成されたパッセージは，必ずその内部で係り受けが成立していることになる．

3.2 漢字拡張とインデクシング

漢字拡張を行なう際，木村の手法では，参照する意味クラスを入力中の形態素が属するクラスに限定している．その意味で，これを局所的漢字拡張法と呼ぶことにする．この方法では，拡張をされることが期待される漢字が，参照される意味クラス中に出現しないため，有効な拡張を行なえない場合がある．そこで，すべての意味クラスを参照する大域的漢字拡張法を提案する．これは，漢字レベルでのシソーラスを構築することによるものである．その手順を以下に示す．

1. NTT シソーラス [8] の固有名詞を除いた意味クラス c について，それぞれ以下の処理を行なう．
   (a) 意味クラス c 中に現れる漢字 k の数 n をカウントする．
   (b) 漢字 k に対し，意味クラス c を n の重みでインデックスとする．
2. 作成されたデータを漢字ごとに整理する．

この処理により，漢字に対し，意味クラスでの頻度を要素とするベクトルが作成される．なお，NTT シソーラスの固有名詞を除いた意味クラス中に出現する漢字の総数は 4,718 個である．このデータを利用し，漢字の類似度を式 1 で定義する．

$$sim(k_1,k_2) = \frac{\overrightarrow{k_1} \cdot \overrightarrow{k_2}}{|\overrightarrow{k_1}||\overrightarrow{k_2}|} \cdot \frac{cf(k_1 \wedge k_2)}{cf(k_1)} \qquad (1)$$

ここで，$cf(k)$ は，漢字 k が出現する意味クラスの数を表し，漢字 k_1 を入力中の漢字と想定している．大域的漢字拡張法では，この類似度がある閾値より大き

表 1: 漢字拡張法の評価実験結果

	w/o	L	L+	G(0.1)	G(0.07)	G(0)
R(%)	75.7	81.3	87.3	83.7	87.3	98.0
Av.O	1.4	2.9	6.4	5.0	7.8	27.8
Av.E	0	2.4	77.7	4.9	9.6	3063

い漢字 k_2 をインデックスに追加する．重みは，式 1 で与えられる値を用いる．

それぞれの漢字拡張法について，EDR 辞書 [9] を用いた評価実験を行なった．実験は，見出し語をクエリとして漢字拡張を行ない，概念説明を検索することにより行なった．これにより，クエリである見出し語に対応する概念説明が検索できれば正解であると自動的に判定することができる．その際，見出し語に含まれる漢字の数が重要な要素となる．本論文では，その数が 2 つの見出し語についての実験結果について示す．該当する見出し語からランダムに約 1000 個を抽出して実験を行なった結果が，表 1 である．それぞれの項目は以下の内容を表す．

- **w/o:** 漢字拡張なし
- **L:** 局所的漢字拡張法を利用
- **L+:** 局所的漢字拡張法において，参照する意味クラス中のすべての漢字をインデックスに追加
- **G(θ):** 大域漢字拡張法において，類似度 θ 以上の漢字をインデックスに追加
- **R:** 正解を検索できた割合
- **Av.O:** 正解の平均順位
- **Av.E:** 見出し語に対し，拡張された漢字数の平均

まず，Recall を比較してみると，大域的拡張法において，閾値を適切に設定する事により，局所的拡張法より大きな値が得られることが分かる．L+,G(0) はそれぞれの拡張法における上限値と言える項目である．局所的拡張法は，87.3%が限度であるのに対し，大域的拡張法では，100%に近い値となっている．このときの漢字拡張数から，閾値を 0 とすることは現実的ではないが，その値を 0.07 としたとき，L+の Recall と等しくなっており，拡張力に優れていることが分かる．一方，平均順位に関しては，大域的拡張法を用いた場合は局所的拡張法を用いた場合より下がる傾向にある．これは，大域的拡張法では，拡張される漢字の数が多いため，それだけ検索のインデックスにマッチする概念説明が多くなるためと考えられる．しかし，実際のシステムでは，1 つのインデックスにマッチしただけでは候補とはされず，入力中のすべての概念に対応するインデックスを含んでいなければならない．よって，この平均順位の悪化が，システムの性能を下げる要因とはならないと考えられる．したがって，Recall の値を重視し，大域的拡張法を用いることが有効であると考える．

3.3 係り受け関係によるフィルタリング

木村の手法では，入力文の構文構造を考慮していないため,「高速道路の建設」に対して,「建設中の高速道路」のように，対応する文節の係り受け関係が反転してしまっている候補が出力される．そこで，入力文に対しても構文解析を実行し，入力と候補において，対応する文節の係り受け関係が反転している候補は除去する処理を 2.2 項のフィルタリングに追加する．

3.4 文脈の利用

評価に用いた 53 名詞句のうち，1 つでも出力のあった 46 入力について，何件の記事からの文脈情報を利用できたかを調べたところ，上限の 10 件まで利用できた入力は 13 個，1 件から 9 件まで利用できた入力が 10 個，1 件も利用できなかった入力が 23 個であった．したがって，文脈情報を十分利用できる入力はあまり多くない．これは，入力文を完全に包含するパッセージが多数見つからなかったことによるものである．したがって，文脈情報を十分利用するためには，別の方法を用いる必要がある．

そこで，形態素レベルでの文脈情報を利用する方法を提案する．それは，形態素に対し，それとよく共起する形態素を文脈と考えるものである．例えば,「手本」と「本棚」はどちらも “本” という漢字を持つが,「手本」は「示す」や「見せる」,「本棚」は「置く」や「並べる」などの形態素とよく共起することから，これらの意味的な類似度は低いとみなすことができる．この考えに基づき，コーパス中の文を構文解析し，互いに係り受け関係にある形態素同士を互いの文脈とする．

このような処理を，1991 年から 1999 年までの毎日新聞記事 [6] を用いて行った．ただし，ある形態素に対し，一度しか文脈として現れない形態素は除去する．また，文脈となっている各形態素は $tf \cdot idf$ によって重み付けをする．これにより，94,332 形態素についての文脈情報が得られ，1 つの形態素につき，平均 72.3 形態素が文脈として登録された．入力と候補の文脈情報は，それぞれに含まれる形態素に対応する文脈を足し合わせたものとし，文脈の類似度はそれらの余弦の

表 2: 平均順位の比較

	対象 1	対象 2
[木村]	12	12.4
実験 1	4.4(5.7)	4.6
実験 2	2.6	4.2

値とする．この文脈情報を利用したときの評価については，次節で述べる．

3.5 候補を用いたフィルタリング

「高速道路の建設」という入力に対し,「高速道路建設」という出力が得られるが，これ以外に「湾岸高速道路建設現場」のように,「高速道路建設」という文字列をその中に含む候補も出力される．しかし，このような候補は入力に対するパラフレーズとは言い難い．そこで，2.2 項で述べた整形を行なった形が，他の候補を包含している候補は除去するフィルタリングを追加する．

4 パラフレーズ抽出実験と評価

実験に際し，入力としては，木村が用いたものと同じ名詞句を利用する．検索対象は，1991 年から 1993 年までの毎日新聞記事を 3.1 項で述べた修正を加えて作成し直したパッセージである．評価の対象は，木村のシステムの出力のうち，二人の判定者がともに正解と判定した 23 個の出力とする．実験は以下の 2 つを行なった．

- **実験 1:** 大域的漢字拡張法を用い，文脈の利用は木村の方法を用いる
- **実験 2:** 大域的漢字拡張法を用い，文脈の利用は形態素レベルでの文脈情報を用いる．

実験 1, 実験 2 で得られた出力は，評価対象 23 個のうち 20 個であった．得られなかった 3 つの評価対象は，3.5 項で述べたフィルタリングにより除去されたり，作成し直した検索対象のパッセージ中に該当する記述がなかったことによるものであった．

それら 3 つ以外の 20 個の評価対象について，その平均順位を比較する．その結果を表 2 の対象 1 の項目に示す．括弧内の数値は，係り受けによるフィルタリングを行なわなかったときの値である．

この表から分かるように，大域的拡張法を用いることにより，正解と判定された出力を上位で得ることができている．また，形態素レベルでの文脈情報を用いたときが最もよい結果となっている．しかし，ここで使用した評価対象 20 個は,「株価の動向」⇔「株価動向」のように，助詞「の」が除去されたり，逆に挿入された形のものが多い．そこで，別の評価対象として，二人の判定者のうち，少なくとも一人が正解とし，入力中のすべての形態素をそのままの形で含んでいない 17 個の出力について評価した．その結果を，表 2 の対象 2 の項目に示す．この場合でも，対象 1 と同様の結果となっている．

以上の結果と 3.2 項での実験結果から，正解を上位で出力でき，拡張力に優れている大域的拡張法を用いることが有効であると言える．また，文脈情報については，実験 2 が最も良い順位であることから，形態素レベルでの情報が文脈として利用できるものであると考えられる．しかし,「電話料金の値下げ」に対し,「通話料金の値下げ」よりも「電気料金の値下げ」の方が大きいスコアが付けられるなど，検討しなければならない問題がある．

5 おわりに

本論文では，木村が行なった研究を基に，漢字インデックスによるパラフレーズ検索の手法を提案した．その結果，木村のシステムによる出力のうち，正解と判定された評価対象をより上位で出力させることができた．大規模な評価は行なっていないが，これらは，システムの改良を行なうための有効な手法であると考えている．

参考文献

[1] 乾健太郎．言語表現を言い換える技術．言語処理学会第 8 回年次大会チュートリアル，2002

[2] 木村健司，徳永健伸，田中穂積．日本語名詞句に対するパラフレーズ事例の自動抽出に関する研究．言語処理学会第 8 回年次大会，2002

[3] 黒橋禎夫、長尾真、日本語形態素解析システム JUMAN version 3.61 使用説明書、京都大学大学院情報学研究科 (1998)

[4] 黒橋禎夫、日本語構文解析システム KNP version 2.0 b6 使用説明書、京都大学大学院情報学研究科 (1998)

[5] 木谷強ほか．日本語情報検索システム評価用コレクション bmir-j2. 情報処理学会研究報告，Vol.98,No.2,pp.15-22,1998

[6] CD-毎日新聞 '91-99' 年版

[7] 西岡真吾、今一修、汎用連想計算エンジン GETA とそれに基づく連想検索システム、情報処理学会研究報告 53,p.93(2000)

[8] 池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩巳、小倉健太郎、大山芳史、林良彦、日本語語彙体系、岩波書店 (1997)

[9] EDR.EDR electronic dictionary version 1.5 technical guide,1996