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

木村 健司　徳永 健伸　田中 穂積
東京工業大学
{kkimura, take, tanaka}@cl.cs.titech.ac.jp

1 はじめに

「ある表現を意味内容を保持したまま別の表現に変換する」というパラフレーズ [1] の技術は表現の不一致によって性能の低下が起りうる情報検索にとって有用である.

また，可読性の向上させるために，パラフレーズを応用する研究も行われてきている [2][3]. これらの多くは，元となる表現に対して，ユーザに最適な表現を一つ返すことを目的としているが，情報検索では，一つの表現からできるだけ多くの類似した意味を持つ表現を生成できることが重要である. したがって，情報検索に特化したパラフレーズの応用としては，従来の手法とは別の手法を用いることが望ましい.

本稿では，以上の背景から情報検索で検索要求文としてよく用いられる名詞句を対象として，新聞記事からそのパラフレーズ事例を抽出する手法を提案する.

提案する手法は情報検索技術を利用する. 1 字の漢字どうしが結合して語を形成するという日本語の特徴を考慮し，漢字を索引とすることで広い範囲での事例の抽出が可能である.

2 パラフレーズ事例の自動抽出

2.1 本手法の概要

本手法の処理の流れを図 1 に示す. 本手法は大きく分けて，新聞記事からの候補の検索，係り受け解析を利用して冗長な部分を削り落す整形処理，そして，意味的に近い表現の順に並べるリランキングという 3 つの過程からなっている. 以下の項では各過程で行われる処理について述べる.

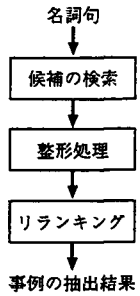

図 1: 処理の流れ

2.2 候補の検索

新聞記事の中から，入力された名詞句に対してパラフレーズとなっている表現を情報検索技術を利用して取り出す. なお，検索対象は新聞記事自体ではなく，新聞記事を句読点や記号で区切った節とする. 情報検索はベクトル空間モデルを使用し，漢字を索引とする.

ただし，以下の例外を設けた.

- カタカナ語に関しては重要な意味を持つと考え，それ一語で索引語とする.
- 具体的な数量だけの違いによる不一致を防ぐため，数詞は抽象化した＜数＞という記号を索引とする.
- 「～に対する」などの助詞相当語に出現する漢字の持つ意味は小さいと考え，EDR 電子化辞書で助詞相当語と定義されている表現に入力がパターンマッチした場合には，その部分の索引付けを行わない.

また，できるだけ様々な表現を収集することが本研

究の目的であるので，索引語拡張を行い，類似した意味を持つ漢字を入力の索引語に追加することを試みた．索引語拡張はシソーラスを参照して行い，入力中の各形態素が属する意味クラス中で，閾値以上の頻度で出現する漢字あるいはカタカナ語を索引語に追加する．

重み付けの方法は検索要求文 (入力側) と文書 (新聞記事側) とで異なる．検索要求文にはシソーラスを用い，各形態素が属する意味クラスでの各漢字の出現頻度に比例して重みを付ける．これにより，入力の表現中で意味支配が強い漢字に大きな重みを付けることが期待できる．ただし，拡張した索引語が大きい重みを持つと悪影響が出ることを確認したため，拡張した索引語には最低の重み 1 を付けた．この重み付けの例を図 2 に示す．

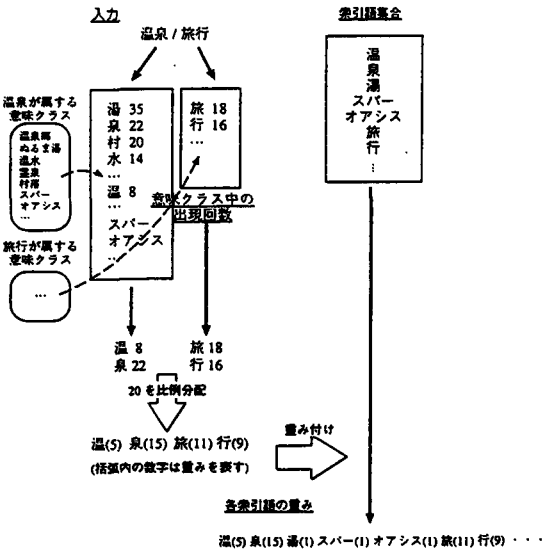

図 2: 入力に対する重み付けの例

一方，文書には全ての索引語で均等に 1 という重みを付けた．これは，以下の理由による．

1. 一般の情報検索と違い，パラフレーズを抽出するという観点では，「何度も繰り返し言及される概念は重要な概念である」という仮定 [4] が必ずしも成り立たない．
2. 検索対象は句読点や記号で区切られた節なので非常に短く，文書内で出現頻度を計算するだけの十分な量のテキストが存在しない．
3. 検索要求文と同じ重み付けを行うのが望ましいが，大規模な記事には計算時間が非常にかかる．

以上の処理により，パラフレーズ事例の候補を得る．

2.3 整形処理

検索された候補は新聞記事を句読点や記号で区切っただけの節であるので，そのままでは冗長な部分や文法的に不適切な箇所がある可能性がある．従って，より自然な出力となるよう，各候補に整形処理を行う．

図 3 に整形処理の例を示す．整形処理はまず，入力に対応する概念が候補中のどこに現われているかの対応付けをとる (マッチング処理)．これは，索引語がどこに現われているかということを基に決定する．概念の対応付けがとれない場合は，不適格であるとして，その候補を棄却する．次に，係り受け解析を行い，文節係り受け構造から各概念を全て含むような最小の構造を取りだす (必要な文節の切り出し)．そして，最後に終端にある助詞や活用された動詞を削る，終止形に直すなどして適切な表現に修正する (終端の修正)．

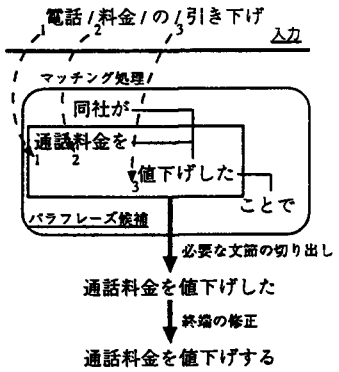

図 3: 整形処理の例

2.4 リランキング

前項までの処理により，システムが出力とする表現のリストが得られる．ここでシステムは，入力と意味的に近い「よい」パラフレーズ事例の順に事例を出力することが望ましいが，検索時の類似度の順だけでは図 4 に示すような入力と意味的に無関係な候補が上位

で出力される可能性がある．そこで，システムが「よい」候補から順に出力するよう，各候補に対してスコアを計算してリランキングを行う．

市[1]内[2]の駐車[3]場[4] → 普通車[3]の国内[2]市[1]場[4]

数字は索引の対応を表わす

図 4: 高いスコアがつくが意味的に無関係な例

スコアは以下の 3 つの尺度で計算された値を掛け合わせることで計算する．

- 検索時のスコア
- 係り受け距離
- 文脈情報

ここで，係り受け距離は入力中で隣接する各概念がどれだけ短い係り受けとなっているかを評価したものである．一方，文脈情報は各候補が記述されていた元記事と，入力名詞句と完全一致する表現のある記事との類似度により計算される．これにより，図 4 のような不適切な候補の順位を低く抑えることが期待できる．

3 実験

我々は本手法を実装したシステムを用いて抽出実験を行った．実装には，検索エンジンとして GETA[5]，形態素解析器として JUMAN[6]，係り受け解析のための統語解析器として KNP[7]，そして，索引付けのために必要なシソーラスとして日本語語彙体系 [8] を使用した．

また，実験で入力となる名詞句として，BMIR-J2 テストコレクション [9] に含まれる検索要求文のうち，本手法の評価に相応しくないと考えられる，社名などの固有名詞を含む名詞句や，カタカナ語・数詞を多く (1 語以上) 含む名詞句を除いた，53 の検索要求文を用いた．各検索要求文は名詞句になっている．抽出する対象となる新聞記事には毎日新聞 3 年分 (1991 ～ 1993) を使用した．

3.1 実験結果

実験の結果，入力に使用した 53 の名詞句のうち，44 の名詞句で少なくとも一つの出力がシステムから得られ，表 1 に示すような，パラフレーズが抽出できた．実験結果の中には「製販一体化」に対して「製造・販売を一体化」を出力するなど，統語構造の変化や単語の置き換えだけでは抽出できないと思われる事例も見受けられた．

入力	出力
農薬	農業用薬剤
製販一体化	製造・販売を一体化
冷夏の被害	冷害問題
株価動向	株価の値動き
経営陣刷新	経営陣の一新
企業の社会貢献	社会に貢献する企業活動

表 1: パラフレーズ抽出例

しかし，一方でパラフレーズとして相応しくないと思われる表現も多く見受けられた．

3.2 評価

二つの表現に対する意味的等価性を客観的に判定することは難しいと考えられる．そこで，二人の作業者に

○ ··· ほぼ同義
例: 所得税の減税 → 所得税減税

△ ··· 関連性あり (一方の表現の意味の範囲がもう一方の意味の範囲を包含している場合など)
例: 所得税の減税 → 大型所得減税

× ··· 関連性なし
例: 所得税の減税 → 申告納税額の対前年比を所得者区分別で見る

という基準に従って，システムが出力した表現を分類してもらい (表 2)，その一致度を調べた．

その結果，κ 統計量は $\kappa = 0.544$ で，十分な一致が見られるとは言えず，パラフレーズの客観的な評価が困難であることが分った．

		作業者 A		
		○	△	×
作業者 B	○	31	12	1
	△	5	234	61
	×	4	168	622

表 2: 二人の作業者による評価

3.3 考察

本項では，抽出実験において確認した本手法の問題点について述べる．

入力の長さ

実験では入力の索引語数が 10 語 (拡張含まず) を超えると，明らかな性能の低下が見られた．これに対しては，名詞句の意味的まとまりを解析し，各部分に分割して処理を行うことで対応できるのではないかと考えている．

並列表記

「日本人と留学生」と「日本人留学生」は意味的に全く異なるが，本手法ではこの 2 つの違いの見分けはつかない．これは，入力の構造を解析していないことが原因として挙げられる．従って，入力に係り受け解析を施し，並列となっている箇所を特定することが必要である．

文脈情報

実験では，入力の約半分が完全一致する表現のある記事が存在しないため，文脈情報を利用できなかった．したがって，そのような場合には「全ての形態素が一致する」など，条件を緩めて文脈のための記事を選定する必要があると考えられる．

4 おわりに

本稿では，日本語特有の語形成の特徴を捉えるため，漢字を索引とした情報検索を利用して，新聞記事から日本語名詞句に対するパラフレーズ事例を抽出する手法を提案した．

実験では，日本語名詞句に対して，統語構造の変更や，単語の置き換えといった手法だけでは抽出できないような興味深いパラフレーズ事例が抽出できた．

しかし，一方で精度が低いという問題もあるため，今後はより高い精度で抽出を行うことができるよう手法を改良していく予定である．また，実際に情報検索の性能に与える影響を調査することも考えている．

参考文献

[1] 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945, 1999.

[2] 新森昭宏, 斎藤豪, 奥村学. 特許請求項の可読性向上のための自動言い換えについての考察. 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 65–70, 2001.

[3] 乾健太郎. コミュニケーション支援のための言い換え. 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 71–76, 2001.

[4] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *j-IBM-JRD*, Vol. 1, No. 4, pp. 309–317, October 1957.

[5] 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, Vol. 2000, No. 53(2000-NL-137), p. 93, 2000.

[6] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.61, 1999.

[7] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6, 1998.

[8] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. Goi-taikei — a japanese lexicon, 1997.

[9] 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(98-DBS-114), pp. 15–22, 1998.