自 然 言 語 処 理　146－7
（２００１．１１．２０）

漢字インデックスを利用したパラフレーズの抽出

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

表意文字である漢字どうしが結合して語形成を行う日本語などの言語では，英語などの言語と比べて独特のパラフレーズが生れる可能性があり，パラフレーズを抽出する手法としては，その可能性も考慮する手法が望ましいといえる．そこで我々は，漢字をインデックスの単位とする情報検索を利用したパラフレーズの抽出手法を提案する．本論文では，手法を具体的に述べるとともに，構築したシステムで実際にパラフレーズの抽出を行う．その結果，単語の置き換えや統語的な変形では抽出できないパラフレーズが抽出できた．

Paraphrase extraction by using Chinese character index

KIMURA Kenji, TOKUNAGA Takenobu, TANAKA Hozumi
Tokyo Institute of Technology
{kkimura, take, tanaka}@cl.cs.titech.ac.jp

This paper proposes a Japanese paraphrase extraction method using information retrieval techniques in which a Kanzi (Chinese character) is used as index terms. Since a Kanzi is an ideograph, two expressions sharing the same Kanzis tend to share the same meaning. Using this characteristics of Japanese, our method succeeds to find paraphrases which can not be found by replacing words nor changing syntactic structures. A preliminary experiment is described and its results are discussed.

1　はじめに

近年，パラフレーズ / 言い換え表現（以下，パラフレーズ）が注目され始めている [1]．パラフレーズは，異なる表現に対する意味的等価性の判定を必要とする情報検索，情報抽出，要約などさまざまな分野で応用が期待されるだけでなく，どのようなメカニズムによってパラフレーズが生成されているのかという言語学的関心もあり，これを研究することは非常に意義があると考えられる．

しかし，一方でパラフレーズを行う目的を明確にする必要であるという指摘もある [2]．パラフレーズの研究は得られたパラフレーズをどのように利用するかという目的をはっきりとさせなければ正解が客観的に規定できず，評価が困難となる．

そこで，我々は情報検索において，パラフレーズを使用してクエリと文書の間での意味的等価性の判定し，検索性能の向上を図ることを目的としたパラフレーズの抽出を目指すことにした．

パラフレーズに関する先行研究として，[3] や [4] がある．前者は情報検索において，変形規則によるパラフレーズで索引語を拡張することで，再現率を向上させている．一方，後者はパ

ラレルコーパスからパラフレーズと変形規則を自動的に獲得する手法を提案しており，両者とも興味深い．

これらの手法は主にパラフレーズ (特に純粋に言語の問題として解決できるパラフレーズ) を，

- 語彙的なパラフレーズ (「手法」 ⇌ 「方法」 など)
- 形態–統語的なパラフレーズ (「花子の考え」 ⇌ 「花子は考えた」 など)
- 統語的なパラフレーズ (「太郎のカバン」 ⇌ 「太郎が持っているカバン」 など)

に分けて考えている．しかし，[5] では，複合語の語形成に関して，日本語においては，特に英語に基づく分析に見られるような，独立性のある語どうしの組み合わせだけでなく，表意文字である漢字（一字漢語）が結びついて複合語を形成する特性が挙げられている．したがって，その日本語の語形成の特性から独特のパラフレーズ生成の可能性が考えられる．そこで，本論文では漢字を単位とする，日本語の語形成を考慮したパラフレーズの抽出手法を提案する．

2 概要

2.1 問題の設定

パラフレーズの抽出を行う上で，入力と出力の問題設定を以下のように規定した．

入力 複合名詞，もしくはそれらが「の」で結ばれたもの

出力 入力と意味的に類似する表現のリスト

入力を複合名詞，もしくはそれらが「の」で結ばれたものとしたのは，本手法が長い文のパラフレーズ抽出には向いていないという判断のためだが，実際の情報検索のクエリからみて，それ程不自然なものではないと考えられる．

2.2 手法の概要

図 1 にシステムの構成図を示す．本論文で提案するパラフレーズの抽出システムは 3 つの要素からなる．

- 入力を検索要求文として情報検索を行い，パラフレーズの候補を検索する．
- 各候補に JUMAN+KNP[6] で係り受け解析を行い， 入力の意味内容を保持しているかを調べ，適切な形へ変形する．
- 各候補にスコアを付け，順位付けして出力する．

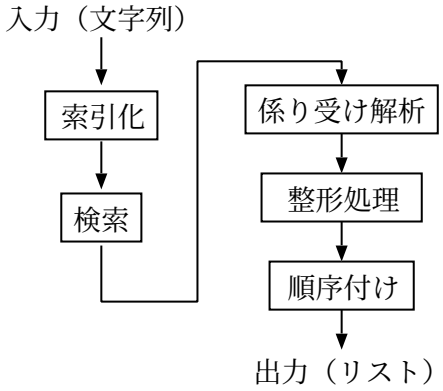

図 1: システム構成

本システムでは，情報検索の索引語として漢字を単位としているのが特徴である．これに関しては次節で詳しく述べる．

3 漢字インデックス

本論文が提案するパラフレーズの抽出手法は，情報検索の技術を利用している．情報検索における索引語の単位としては形態素が用いられることが多いが，本論文では漢字を索引語の単位として用いる．

これは日本語の語形成の特徴と共に，表意文字である漢字は，それ自身である程度の意味を保持しているという考えに基づくものである．

3.1 索引付けの方法

索引語は索引付けする対象を JUMAN[7] で形態素解析し，名詞，動詞，形容詞，副詞，および未定義語と判定されたものの中に表われた漢字のリストとして生成される．ただし，以下にような例外を設ける．

1. 外来語であると思われるカタカナ語は漢字としては表せないが，強い意味内容を持つものとして，そのまま利用する．
2. 数詞は時として情報検索で重要な情報をもたらすが，今回はさまざまな言い方を抽出することが目的であるので，＜数＞という概念で数詞を抽象化する．

3.2 重み付けの方法

次に，索引語に対する重み付けを説明する．一つの単語を複数の漢字に分割した場合，元の単語に対する各漢字が持つ意味の強さは均等ではない場合が考えられる．したがって，元の単語と漢字の意味を考えて適切な重み付けを行う必要がある．

日本語においては，右側主要部の規則が挙げられるが，これは実際の意味支配とは異なるものである．そこで我々はシソーラスを利用して，以下の方法で重み付けを行った．

1. 日本語語彙体系 [8] で，索引語を生成する際に得られた形態素が含まれている意味クラスを探す．
2. その意味クラスに属する他の単語を探し，各単語の漢字の出現回数をカウントする．
3. 分割対象の形態素の各漢字が何回カウントされているかの比を計算し，それを正規化したものをその漢字の重みとする (今回の実験では 1 形態素からできる索引語の重みの和が 20 になるように正規化した)．

3.3 索引語の拡張

また，これに加えて，シソーラスを利用して以下の方法で索引語を拡張した．

漢字の場合

$$T_e = \{c_C | freq(c_C) > freq(c_{maxInput}) \text{ or } c_C \text{ がカタカナ語 } \}$$

カタカナ語の場合

$$T_e = \{c_C | \max freq(c_C)\}$$

$T_e \cdots$ 拡張する索引語集合
$C \cdots$ 元の索引語が属する意味クラス
$c_C \cdots C$ 内に現れる漢字
$freq(t) \cdots t$ の出現回数
$c_{maxInput} \cdots$ 入力中で着目している形態素の中で，最も出現回数の多かった漢字

つまり，漢字の場合は元の索引語が属する意味クラスの中で，元の索引語で最大の出現回数の漢字より多く出現する漢字を索引語に加え，さらに同意味クラス中に出現するカタカナ語も索引語に加える．一方，カタカナ語の場合は意味クラス中で最も多く出現する漢字を索引語に加える．これは，カタカナ語は一つの意味クラスに一回しか出現しないことを考慮したことによる．なお，拡張された索引語の重みは，すべて 1 とした．

最後に‘温泉旅行’という入力があった場合のインデックス作成例を図 2 に示す．‘温泉旅行’をまず，‘温泉’と‘旅行’という形態素に分割する．次にシソーラスで‘温泉’と同じ意味クラスに属する語を集め，それらの漢字の出現回数をカウントする．‘温’と‘泉’の出現回数は 8：22 なので，これを和が 20 になるように正規化して‘温’に 5，‘泉’に 15 の重みを付ける．また，‘温泉’の中で最頻だった‘泉’よりも多く出現している‘湯’と，カタカナ語である‘スパー’や‘オアシス’などを，重み 1 で索引語に加える．‘旅行’についても同様に行い，インデックスを作成する．

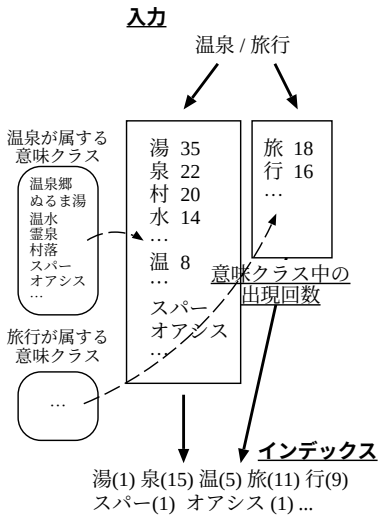

図 2: 索引付けの流れ

4 検索データの作成

パラフレーズ抽出システムを構築するにあたって，まずは候補を検索するためのデータベースが必要である．我々はこのデータベースを新聞記事から自動作成することにした．これらのデータは汎用連想計算エンジン GETA[9] を利用して検索する．作成の手順は以下の通りである．

句読点による記事の分割　まず，新聞記事を句読点あるいは記号で区切り，それを情報検索における検索対象とする．句読点や記号は JUMAN を用いて形態素解析を行って判別している．ただし，読点や記号は並列表記にも用いられることを考え，‘今日，明日’ のように，その前後の品詞が同一である場合は並列表記であるとして区切らないこととした．

索引付け　索引付けは 3 章で述べたように漢字をインデックスとして行うが，その重みはすべて同値 (すべて 1) としている．これは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 [10] が必ずしも成り立たず，また，入力中のそれぞれ索引語に対応する語が文書中に含まれているかということを重視しているためである．

5 パラフレーズの抽出

本節では，本研究が提案する手法の各々の段階の詳細と実際の実装方法について述べる．

5.1 パラフレーズ候補の検索

パラフレーズの抽出にあたり，まずは情報検索を行ってパラフレーズ候補を獲得する．

入力は，3 節で述べたとおり，インデックス化，重み付け，索引語の拡張を行い，漢字インデックスとして表現される検索要求に変換する．そして，検索要求は GETA に入力し，複数のパラフレーズ候補を得る．

ここで，検索はベクトル空間法を利用している．類似度の計算は，文書と入力に出現する索引語の集合を T，$T \ni t$ である t の重みを $w(t)$ とし，以下の式で計算する．

$$Sim_{qd} = \sum_{t}^{T} w_q(t) \times w_d(t)$$

ここで，$w_q(t), w_d(t)$ はそれぞれ，入力と文書の t に対する重みである．

獲得した候補は，入力と全く同じ表現が含まれている場合はそれを除き，整形処理を行う．

5.2 整形処理

次に，検索によって得られたパラフレーズ候補を JUMAN + KNP によって係り受け解析し，以下の手順で整形処理を行う．

1. 入力の各形態素の概念を含んでいる候補中の文節を同定する．これは，入力のインデックスとパラフレーズ候補中の漢字の対応により行う．
2. パラフレーズ候補が入力の各形態素が表す概念を含んでいるかをチェックし，マッチ

ングを行う．すべての概念が含まれていない候補は除く．
また，否定表現に留意し，‘非’，‘未’，‘無’，‘不’ 及び ‘ない’ を否定表現カテゴリとみなして，これに関してもマッチングを行う．ただし，二重否定の可能性もあるので，入力とパラフレーズ候補双方の出現数の偶奇も考慮した．

3. 入力の表す概念を含む文節とそれらを結ぶのに必要な文節を切り出す．
4. 切り出された文字列の終端が適切な形になるように修正する．具体的には，助詞の削除，動詞の終止形への変更などが挙げられる．ただし，‘調印が見込まれる’ など受動態の動詞の場合は，変更は行わない．

この整形処理が行われる例を図 3 に示す．

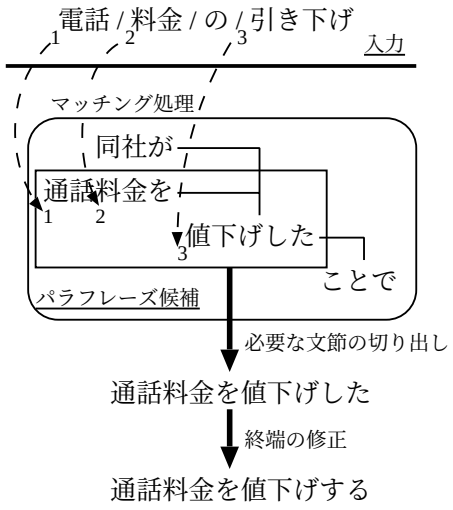

図 3: 整形処理例

5.3 順序付け

最初に行った検索の返す順序だけでは，パラフレーズとしての内容の良さを考えた順序とは言い難い．そこで，検索でのスコア ($S_{initial}(p)$) と，入力で隣接していた形態素と対応する表現の係り受け構造での距離 ($Dist(i, i+1)$) を用いて以下の式でスコアを修正した．

$$Score(p) = \frac{S_{initial}(p)}{\sum_i Dist(t_i, t_{i+1})}$$

このスコアで順序付けを行い，それを最終出力とする．

6 実験と考察

6.1 実験

作成したシステムを用いて，新聞記事よりパラフレーズを抽出する実験を行った．実験には，新聞記事として毎日新聞 3 年分と，入力は BMIR-J2[13] コレクション中より，人手で任意に指定した入力を使用した．図 4 に入力の一例を示す．

収益の減少
塾通いの子供
株価低迷
倍額の増資
個人所得課税の減税

図 4: 使用した入力の一例

一方，出力は順位付けされた候補のうち，上位 10 フレーズを最終的な出力とした．ただし，整形にかかる計算量を考慮し，検索を行った時の上位 50 フレーズまでをパラフレーズ候補として整形処理することにした．

6.2 結果

このようなパラフレーズの抽出に対し，どのような評価方法をもって性能を測ればよいかは，非常に判断が難しい．そこで，本論文では実験によって抽出されたパラフレーズを主観的ではあるが，以下の 4 タイプに大別した．ここでは，それぞれのタイプを簡単に記す．

1. 語彙的，統語的，もしくは形態–統語的なパラフレーズ
   例)

- 収益の減少 → 経常利益が減少
- 塾通いの子供 → 進学塾に通う子供
- 株価低迷 → 株価の低迷
- 個人所得課税の減税 → 個人所得課税の負担軽減

2. 複数の形態素が表わす概念が一つの形態素へと凝縮されるなど，単語の置き換えでは抽出できないパラフレーズ
   例)
   - 収益の減少 → 減収減益
   - 塾通いの子供 → 通塾の子供
   - 倍額の増資 → 出資額倍増
3. 場面によって，パラフレーズとして適格か否かが分かれるもの
   例)
   - 収益の減少 → 二期連続減収・減益
   - 塾通いの子供 → 子どもは大手塾に通わせる
   - 株価低迷 → 低迷株価
   - 倍額の増資 → 貸付限度額の倍増
4. パラフレーズとはいえないもの
   例)
   - 収益の減少 → 学生人口減少分の収益
   - 倍額の増資 → 高額商品の輸出増加
   - 個人所得課税の減税 → 所得税減税など個別

6.3 考察

実験の結果，いくつかの興味深いパラフレーズが抽出できたと考えられる．特に，2 に分類したパラフレーズは，日本語における語形成の特徴に沿うもので，本手法における漢字インデックスが有効に機能していると考えられる．

一方，3 に分類したものは判断が難しい．例えば，‘塾通いの子供’ と ‘子供は大手塾に通わせる’ では，前者は ‘子供’ という人間に力点が置かれているのに対し，後者は ‘通わせる’ という行為に力点が置かれている．他にも，この辺りの適合性の判断は場面や個々によって異なることが予想され，パラフレーズに関する客観的な評価を困難なものにしている．

本手法には未だ不完全な部分もあり，改良すべき点も多い．今後の課題を以下に示す．

入力の長さ　本論文では，入力を複合名詞，もしくはそれらが「の」で結ばれたものと規定しているが，規定を満していても，入力の文字数が長くなると，全く関係のない，誤ったパラフレーズを抽出することが多く見られた．これは，平坦な重み付けと索引語の増加により，さまざまな意味をもつ候補が検索されてしまっているためだと思われる．

これについては，文脈情報の類似性の利用などで性能を改善できると考えられる．

重み付けの問題　3 年分の新聞記事に対して，本手法の重み付けでは，多くの候補が同じスコアとなってしまう状況がよく見られた．従って，評価尺度の見直しなど，重み付けの方法を洗練する必要がある．

マッチング　本手法で行っているマッチングは，おたがいの漢字が出現しているかを調べるだけの極めて単純なもので，性能がよいとは言えない．しかし，マッチングは重要な要素と考えており，より性能のよいマッチングを行うことが望ましい．

これについては，[12] などを参考にすることで，性能を改善することができるのではないかと考えている．

整形処理　修飾語がついたために厳密にはパラフレーズとはいえない出力が多数見られた．これは，部分切り出しを文節以下の単位で行っていないためである．我々は将来的にパラフレーズを情報検索に応用しようと考えているが，この時，パラフレーズに厳密さが要求されるかどうかは定かではない．この点に関しては更なる観察が必要であると考える．

7 まとめ

本論文では，情報検索を利用したパラフレーズの抽出手法を提案した．この際，日本語の語形成の特徴を考え，表意文字である漢字をインデックスとして情報検索を利用した．

また，提案手法を実装したシステムを構築して実験を行った．その結果，いくつかの興味深いパラフレーズが抽出でき，本手法の有効性が確かめられた．また，誤った出力に対して，その原因と解決するための対策を考察した．

参考文献

[1] 言語処理学会第 7 回年次大会ワークショップ論文集, 2001.

[2] 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.

[3] Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of 35th Annual Meeting of the Assosiation for Computational Linguistics*, 1997.

[4] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 50–57, 2001.

[5] 影山太郎. 文法と語形成. ひつじ書房, 1993.

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

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

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

[9] 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.

[10] 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.

[11] CD–毎日新聞 '91–'93.

[12] 黒橋禎夫, 酒井康行. 日本語表現の柔軟な照合. 言語処理学会 第 7 回年次大会発表論文集, pp. 343–346, 2001.

[13] 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.