言い換え処理技術の文書検索システムへの適用

斎藤　佳美、 加納　敏行、 倉田　早織
東芝ソリューション（株） IT技術研究所
{ Saito.Yoshimi, Kano.Toshiyuki, Kurata.Saori}@toshiba-sol.co.jp

1. はじめに

本稿では、筆者らが現在開発に取り組んでいる、言い換え処理を利用した文書検索システムについて述べる。従来の単語をベースとした文書検索では、検索要求に含まれる単語間の概念関係は利用されておらず、検索キーワードが多数の文書で出現するような場合、利用者が意図する文書のみを網羅的に検索することがむずかしい。この問題の解決には、単語間の概念関係を考慮した検索方式が有効であると考えられるが、その際、同じ概念関係を示す多様な表現に対処することが課題となる。

このような、同じ意味内容を示す多様な表現を取り扱うための技術が言い換え処理技術であると考えられる。[1][2] 我々は、言い換え表現の生成、表現の類似度の判定などの言い換え処理技術と、それを利用した文書検索方式を開発している。本稿では、本方式、および従来の言い換え処理技術との関連について述べる。

2. 言い換え処理技術を利用した文書検索方式

言い換え処理技術の研究には、図1に示したような2つの方向性があると考えられる。

1つは、ある表現を入力として、それを別の表現に言い換える技術、すなわち、言い換え表現の生成技術の研究（A）である。

もう1つは、文書などに含まれている表現の中から、ある条件の下、言い換えと見なせる表現を見つける技術、すなわち、言い換えの発見技術の研究（B）である。

（A）においては、言い換え表現を生成する規則などを用いて言い換え表現を生成し、生成された表現が言い換えとして妥当であるかを評価するというアプローチが多く用いられる。

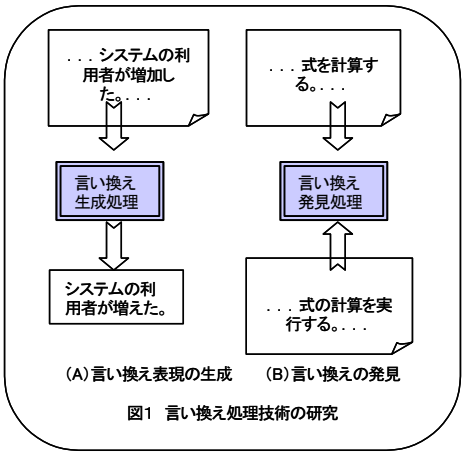

図1 言い換え処理技術の研究

一方、（B）においては、2つの表現の間の共通性を用いて言い換えの候補を発見し、候補となる表現間の対応関係を調べることで言い換え表現であるか否かを判断するというアプローチが多く用いられる。

我々の文書検索方式は、この2つのアプローチを融合し、言い換え処理を文書検索に適用するものである。図2に本方式の構成を示した。本方式は次の 3 段階の処理で構成される。

①言い換え候補の発見
検索要求文に出現する単語をキーワードとして、従来の検索処理により、検索対象文書集合から、検索要求にマッチする文を検索し、これを言い換え候補とする。

②言い換え表現の生成
言い換え候補に対して言い換え規則を適用し、言い換え表現を生成する。

③表現間の比較処理
生成された言い換え表現と検索要求とを比較し、言い換え結果として妥当であるかを判定する。妥当な文（を含む文書）を検索結果とする。

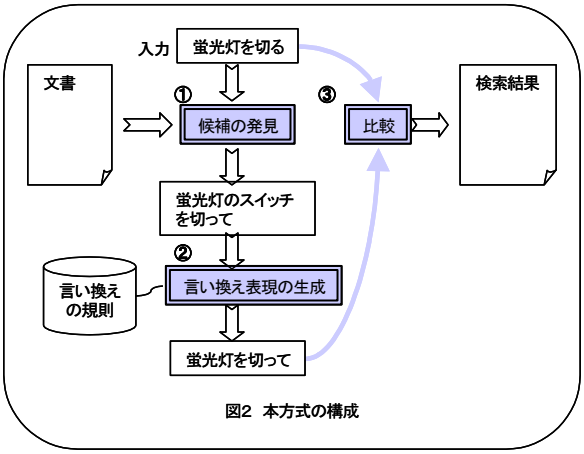

図2 本方式の構成

本方式の特徴の1つは、言い換え候補の言い換えとしての妥当性の評価を、②、③の処理の組み合わせで実現する点である。

これまで筆者らは、言い換え現象を分析するための実験システムを構築し、言い換え表現の収集と分析を行ってきた。[3] その結果、文を構成する単語間の対応関係に基づいて言い換え表現間の対応関係を記述する方式では、複雑かつ多岐にわたる対応関係の記述が必要であることがわかってきた。

（例）
（a）新製品をプレゼンテーションする。
（b）新製品の説明資料を作成し、会議にてプレゼンテーションを実施した。

上記の言い換え表現の例では、(a)と(b)の文に「新製品」と「プレゼンテーション」という2つの共通する単語が含まれている。しかし2単語の係り受け関係は(a)と(b)で大きく異なっているため、対応関係の記述が複雑となる。

一方で、上記例を詳細に分析すると、複数の言い換え現象が含まれていることがわかる。

- 「プレゼンテーションする」⇔
  「プレゼンテーションを実施する」
  （機能動詞結合による言い換え）
- 「新製品」⇔「新製品の説明資料」
  （換喩表現による言い換え）

これらのことから、言い換え候補の評価を、②、③の処理の組み合わせで実現する手法が有効であると考え、本処理方式に基づく文書検索システムを開発した。

3. 各処理の詳細

3－1. 言い換え表現の生成処理

言い換え表現の生成処理では、検索要求文に基づき発見された言い換え候補文に対して、言い換え規則を適用し、言い換え表現を生成する。言い換え規則の種類としては次のようなものを利用している。

- 機能動詞結合による言い換え
  （例）「印刷を実行する」⇒「印刷する」
- 換喩表現による言い換え
  （例）「蛍光灯のスイッチ」⇒「蛍光灯」
- 格共有構造による言い換え
  （例）「野菜を切って鍋に入れる」
  ⇒「野菜を鍋に入れる」

機能動詞結合による言い換えや、格共有構造による格要素の補完は、これまでに広く用いられており、言い換え規則として高い安定性が期待できる。また、換喩表現による言い換えも、例えば[4]において高い妥当性が報告されている。

また、言い換え規則の適用にあたっては、検索要求文の情報を適用条件として利用することとした。即ち、検索要求文に含まれる単語を言い換えによって削除しないよう、言い換え規則の適用を制限している。これにより、生成される言い換え表現の数を抑制することが可能となる。同時に、検索要求文が異なれば生成される言い換え表現も異なることになり、検索要求文に合った言い換え表現の生成が実現される。

また、言い換え規則の適用は再帰的に実行する。このことにより、言い換え規則を簡素化することができる。例えば、前記の例文（b）については、言い換え規則の再帰的適用により、以下のような言い換え表現が生成される。

「新製品の説明資料を作成し、会議にてプレゼンテーションを実施した。」
↓
＜機能動詞結合による言い換え＞＋
＜格共有構造による言い換え＞＋
＜換喩表現による言い換え＞
↓
「新製品を会議にてプレゼンテーションした。」

3－2. 表現間の比較処理

表現間の比較処理では、検索対象文書中の言い換え候補文から生成された言い換え表現の構文木と、検索要求文の構文木を、文を構成する単語間の対応関係に基づいて比較し、類似度を判定する。

検索対象側の文は言い換え表現の生成処理により単純化されているため、類似度判定の負荷が軽減される。例えば、前記の例文(a)と例文(b)においては、次の2文の類似度を判定すればよい。

(a)「新製品をプレゼンテーションする。」
(b)「新製品を会議にてプレゼンテーションした。」

3－3. 言い換え処理の文書検索への適用の効果

以上に述べたように、本方式により、従来の文書検索と比較し、検索要求文で表現される概念に則した適切な文書検索を実現することができる。図3は、本方式による文書検索の1例を示したものである。図3の入力例1では文書(Ⅰ)が、入力例2では文書(Ⅱ)が、本来望ましい検索結果である。従来の検索では、入力例1に対し、文書(Ⅰ)、(Ⅱ)が共に検索される。これに対し、本方式による検索では入力例1に対する検索結果は文書(Ⅰ)のみとなる。一方、入力例2に対しては従来どおり文書(Ⅱ)が検索される。

表1に本方式に基づく文書検索システムでの評価実験結果を示す。対象は技術文書で、検索要求文 15 文に対し①の検索結果として得られた 906 文において、機械の判定結果の適切性を人間の判定と比較評価した。

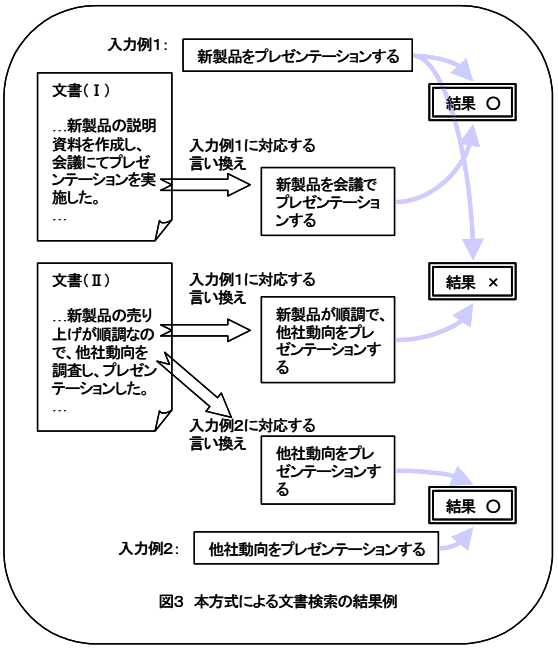

図3 本方式による文書検索の結果例

表1 評価実験結果

検索結果(906文)	人間が言い換えと判定	人間が言い換えでないと判定
機械が言い換えと判定	426	45
機械が言い換えでないと判定	44	390

4. おわりに

言い換え表現の生成と、言い換え表現の発見という2つのアプローチを融合することにより、言い換え処理技術を文書検索システムに適用した。本方式により、従来よりも高い機能の文書検索を実現できる見通しが得られた。

言い換え処理は、解析的観点に加えて、生成的観点で言語を取り扱う点に特徴があると考えている。今後は、処理精度の向上を図るとともに、複数文に亘る言い換え処理など、より高度な言い換え処理の文書検索への適用を検討していきたい。

[1] 佐藤理史(2001). “なぜ言い換え／パラフレーズを研究するのか.” 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 1-2

[2] 乾健太郎, 藤田篤(2004). ”言い換え技術に関する研究動向.” 言語処理学会誌, 11(5), pp. 151-198.

[3] 斎藤佳美, 住田一男(2006). “言い換え現象分析のための実験システムの開発.” 言語処理学会第 12 回年次大会発表論文集, pp. 783-784

[4] 清田陽司, 黒橋禎夫, 木戸冬子(2004). “自動抽出した換喩表現を用いた係り受け関係のずれの解消.” 言語処理学会誌, 11(4), pp. 127-145