言い換え現象分析のための実験システムの開発

斎藤　佳美、 住田　一男
（株）東芝　研究開発センター　知識メディアラボラトリー
{yoshimi.saito, kazuo.sumita}@toshiba.co.jp

1.　はじめに

近年注目を集めている言い換え技術は、質問応答、情報検索、文書要約、校正支援、読解支援など、様々な応用場面を持つと考えられているが、扱う対象が複雑であり、言い換え技術の発展のためには、多様な言い換え現象の分析と、言い換えに関する知識の蓄積が必要である。

乾ら[1]によれば、言い換え現象には、語彙・構文的言い換え、参照的言い換え、語用論的言い換えの3種類があり、これらのうち、これまでに主に扱われているのは、語彙・構文的言い換えである。しかし、実際に言い換えを扱うような応用システムを構築しようとする場合、下記のような事例に遭遇することがある。

(I)　雪で交通が麻痺する
(II)　大雪の影響で新幹線が運転を見合わせる

この2つの表現を言い換え現象として捉えようとする場合、語彙的言い換え、構文的言い換え、参照的言い換え、といった複数の要素を同時に含んでいると考えられる。

上記の2つの表現中では、

- 語彙は全て異なっており、必ずしも類義語の範囲内での言い換えには収まっていないが、関連性の高い語彙で構成されている。
- 構文は異なっているが、部分的には同一性がある。
- 内包的意味が同一であるとは言えないが、文脈を参照しなくても意味的な同一性が理解可能な事例である。

このように、上記のような事例を取り扱おうとする場合、これまでに得られている語彙的言い換え、構文的言い換え、参照的言い換えに関する知見に加えて、さらに詳細な言い換え知識が必要であると考えられる。

そこで我々は、上記のような事例をターゲットとして、ノンパラレルコーパスに含まれる言い換え現象を抽出し、分析することを目的とした実験システムを開発した。

この実験システムは、日本語の類義表現辞書の編集機能、構文解析結果の比較機能を備え、言い換え候補の分析、類義表現辞書の作成、抽出する言い換え表現の洗練をそれぞれ支援する。

本稿では、実験システムの概要を報告し、日本語新聞記事での分析プロセス例について紹介する。

2.　実験システムの概要

図1は、実験システムの構成を概念的に示したものである。

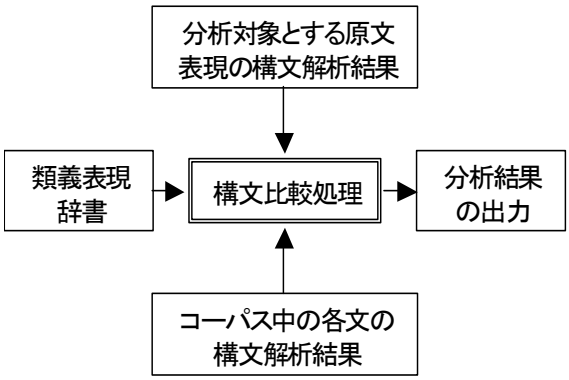

図1

図に示したように、対象となるコーパス中の各文は、あらかじめ構文解析を施した上でDBとして実験システムに登録されている。一方、分析の対象とする原文表現も構文解析され、結果は同じく実験システムに保持される。

構文解析結果はそれぞれ依存構造で表現されている。

構文比較処理では、この2つの構文解析結果を比較し、条件を満たして一致した文を選び出し、比較結果と共に出力する。構文比較処理としてはパラレルコーパスからの言い換え表現の抽出方法として、ノード間の対応関係をベースに構文を比較する方法が知られている[2]。本実験システムではノンパラレルコーパスを対象としているが、コーパスの規模があまり大きくないという前提で、次の3つのステップにより比較を行っている。

(1) 2つの文を比較して同じ単語の対を見つける。その際、類義表現辞書に登録されている語句は同一単語と見做す。
(2) 複数の単語対が見つかった文を対象に、単語対を共有するような部分木を見つける。その際、依存構造上のノードの距離が2以上となるような部分木も一定の条件を満たすものについて対象とする。
(3) 見つかった部分木の数を集計し、一定数以上のものを選び出す。

類義表現辞書には、原文表現中に含まれている単語のリストに対して、各単語の類義表現を自由に編集して登録することができる。

3. 分析プロセス例

次に、上記の実験システムを用いて、比較的小規模のノンパラレルコーパスに含まれる言い換え現象について分析するプロセスについて述べる。ここでは、毎日新聞記事(1998年1月分)を対象とした分析例を紹介する。原文として先に挙げた(I)の例文「雪で交通が麻痺する」を用いた。分析は、以下のステップを繰り返すことにより行う。

(1) 原文を解析した結果の単語リストに対して類義表現を登録する。
例)「雪」→「大雪」
「交通」→「交通機関」「電車」「バス」
「麻痺」→「混乱」
(2) 構文比較処理によって文を選び出す。
(3) 選び出された文集合から言い換えにあたる文を目視で探し、該当する分析結果を調べる。
(4) 単語対が得られていない部分について、類義表現辞書に登録できる表現を探す。
例)「雪」→「降雪」「積雪」
「交通」→「鉄道」「ダイヤ」「高速」
「麻痺」→「運休」「不通」

上記のステップを繰り返した結果、例文(I)に対する次のような言い換え事例を得た。

例)
「首都圏に大雪、交通大混乱」
「雪の影響で新幹線はダイヤが混乱」
「大雪の影響で首都圏の鉄道ダイヤの乱れが続く」
「大雪で電車が不通になる。」
「大雪のために渋滞する首都高速」
「積雪の影響で、私鉄を含めた一部線区は運休や徐行運転を強いられた」
「JR 総武線は降雪の影響で運転を見合わせた」

上記の事例については選び出した際の類義表現辞書や構文比較結果を得ることができ、これらの言い換え現象についての分析を容易に行うことができる。

4. おわりに

こ本稿で報告した実験システムを用いることにより、複雑な要素を持った言い換え事例を一定の基準の元で効率よく選び出し、分析することが可能である。このような実験を通して得られる言い換え事例を蓄積することにより、より複雑な言い換え現象を処理する際に必要となる類義表現や構文変換など、詳細な言い換え知識を獲得していくことができると考えている。

今後は、単語間の対応関係が1対多となるような言い換え事例、複数文にまたがる言い換え事例の処理などに取り組んでいきたい。

[1] 乾健太郎, 藤田篤(2005). ”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.
[2] 関根聡(2001). ”複数の新聞を使用した言い換え表現の自動抽出. ” 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 9-14