E-031

Web 新聞記事と携帯端末用記事における語句言い換え表現の対応付け

Matching of Paraphrased Expression in News Articles on the Web and Mobile Terminals

米村　良介†　　山田　剛一†　　絹川博之†
Ryosuke Yonemura　Koichi Yamada　Hiroshi Kinukawa

１．はじめに

インターネットの普及などにより我々は多くの情報を手に入れることが可能となった。その出力デバイスとして携帯電話などの小画面のものも一般的になってきている。これらに文字情報を表示する場合、画面の大きさにより出力文字数は制限される。よって、携帯端末で文字情報を扱う場合、あるオリジナルのテキスト情報全体ではなく、情報の中核となる意味内容を保持したままで、文字数が抑えられたテキストを生成する技術が求められる。

手法としては抄録や不要文字列の削除といったものがあるが、読みやすさなどの観点から、アプローチとして主に「言い換え（換言）」により文字数を削減する方法を検討することにした。方式としては、言い換え辞書などを利用し、短く言い換えることのできる部分を変換することによって文字数を抑えたテキストを生成するというやり方を考えている。

目的達成の第一歩として、２つの同一内容を表す記事を比較し、語句言い換え表現の対応付けをしようと考えた。その際、プレーンな記事よりも、いくつかの情報を持った記事のほうが処理し易いと考え、記事自体に情報を付加させるようなプログラムを開発することにした。

２．使用するデータ

使用する記事データとして、インターネット上の電子新聞記事（以降、Web 記事）と携帯端末用記事（以降、携帯記事）の対応付けコーパス[1]を用いることにした。これは、Web 記事と、その Web 記事を人手で 1～2 文に要約した携帯記事が対応付けられたものである。事象を詳しく表現している Web 記事と、簡潔に表現している携帯記事を比較することにより、短く言い換えることの出来る表現を抽出できると考えた。

３．記事 XML 化

言い換え表現の対応をとるには、1 単語に着目するのではなく、ある曖昧な部分に着目しなければならない。構造化されていない文書の場合、その曖昧な部分というものを処理するのが困難である。XML は文書要素をツリー構造で表すことが出来る。文書を細かな部分に分割してツリー化し、各々の要素に固有の情報を持たせ、また、要素の親子・兄弟関係といった点も考慮していくことで、曖昧な部分の選定に大いに役立つと考える。

Web 記事と携帯記事を XML で構造化するためのアプローチを以下に示す。また、図１にその構成を示す。

(1)　今回使用したデータである対応付けコーパスは、HTML 形式で表されており、内部のデータは Web 記事・携帯記事といった情報が混在している。これを Web 記事と携帯記事とに分解する。また、不要な HTML タグなどはすべて消去し、純粋な文字情報だけを抽出する。この際、HTML タグの特徴を利用することにより（
など）各々の記事に対して、段落，文といった単位に文章を分割して（Web 記事の場合は「見出し」要素についても分割する）、それらの情報を各記事に付加する。

(2)　文単位に分けられたそれぞれの記事に形態素解析システム茶筌[2]を使用し、文を構成している形態素の情報を得る。形態素とは、語のなかで変化しない最小単位のことをいう。

(3)　それぞれの形態素に茶筌で得た情報を XML タグという形で付加する。茶筌は形態素ごとに ①本文中での語，②読み方，③語の基本形，④品詞情報，⑤活用の種類，⑥活用形 という６つの属性を与える。例えば、"伸びていた"という部分があった場合、「伸び」という形態素に対してのタグ付けは、<chasen word="伸び" read="ノビ" basic="伸びる" part="動詞-自立" con="一段" form="連用形">伸び</chasen>などとする。

(4)　形態素解析により分割され過ぎた形態素を統合し、より構造化された処理し易いデータとする。人名（性，名，～さん）、数詞（金額，年齢，個数）、地域（都道府県，区，市，町，村）などの小さなかたまりを要素として統合し、続けて名詞句，動詞句，連体修飾句，連用修飾句などといった、より大きな要素へと順次統合していく。

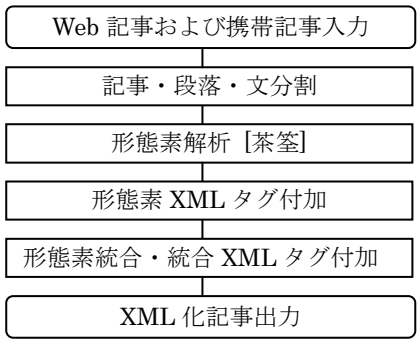

図 1. 記事 XML 化方式

図２は、(1)～(4)の結果により出力された文章の一部分である。「前回のフランス大会前には、・・・」という部分の結果であるが、読点「、」での区切れ目を一つのまとまった句の区切れ目としている。さらに、子要素として連体詞で修飾された名詞句、さらにその子要素として名詞ブロック（名詞連続部分）を含むツリー構造となって出力されている。

†東京電機大学大学院　工学研究科

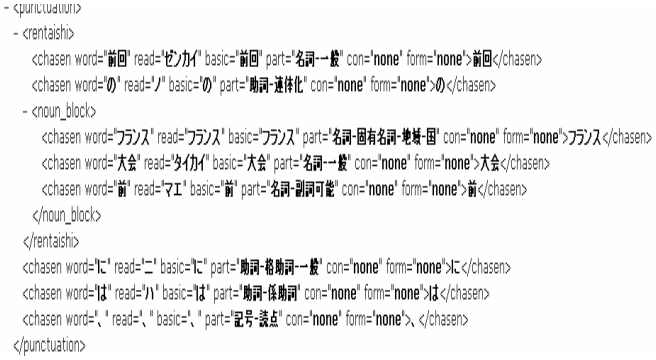

図 2. XML タグ付加の結果

４．言い換え表現の対応付け

同一内容を表す Web 記事と携帯記事を比較し、言い換えられていると思われる部分の対応付けを行う。

現在検討中の方法は、自立語がそのまま使用されていることが多いことに着目し、自立語連接が最長に一致する箇所を対応付けるというものである。それにより対応付けられた部分の自立語以外の箇所で言い換え表現を抽出することが出来ると考えている。

例えば、図 3 のように二つの表現があるとする。上部は携帯記事であり、下部は Web 記事である。なお、<a>, <b>, …<o>の記号は 3.の(3)で示した<chasen・・・>を簡略化して示したものであり、<a>数値といった形をとっていることを表している。<ab>, <hi>等は 3.の(4)の形態素統合の結果得られる名詞句などの統合タグを示しており、<hi><h>ランク</h>付け</hi>といった形を表す。

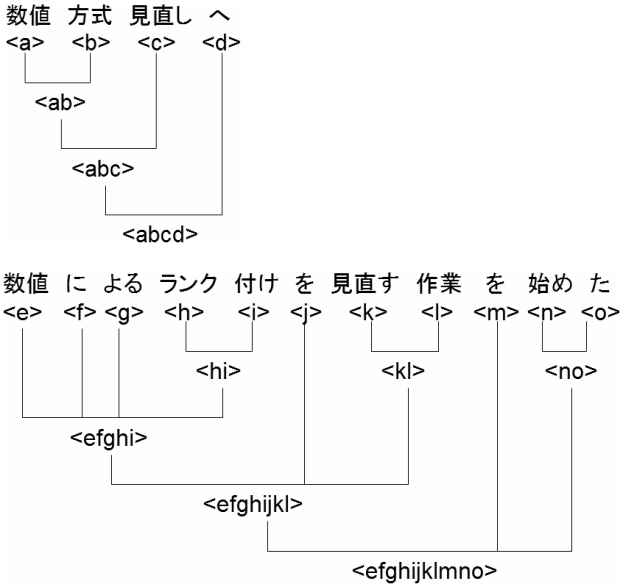

図 3. 言い換え表現対応手順

(1) 携帯記事から自立語である<a>「数値」, <b>「方式」, <c>「見直し」を取り出す。<a>が Web 記事に対して一致するかを見ていくと、<e>「数値」で一致することがわかる。次に、<b>を見ていくが一致する箇所はない。最後に<c>を見ていくと、<k>「見直す」で一致する。

(2) 一致した<a>と<c>を含む句、<e>と<k>を含む句を探すと、それぞれ<abc>と<efghijkl>が得られる。よって、「数値方式見直し」と「数値によるランク付けを見直す作業」を対応付けることが出来る。

(3) (2)で得た句に対して、(1)で一致した形態素を取り除く。これにより、「方式」と「によるランク付け」を対応させることが出来る。

(4) <l>「作業」は<k>「見直す」が連体修飾しており、<kl>で句を成す。そこで、<c>と<kl>とが対応すると判断する。

(5) <d>と<m>, <n>, <o>は文末に位置している。<m>, <n>, <o>中の述語が<n>である事から、<d>と<no>を対応させる。

６．考察および検討

・XML 化について

(a) 記事 XML 化プログラムの開発では、新聞記事（プレーンテキスト）に対して XML タグを付加することにより、文書を構造化することが出来た。

(b) XML タグとして多くの情報を記事に付加した結果、元の記事と比較してデータ量が大幅に増加した。タグ内部の形式について検討が必要である。

(c) 人名辞書, 地名辞書との照合により、人名や地名を表す文字列を一つの句にまとめることが可能と考えられる。

(d) XML 化の精度を実験的に確かめることが必要である。

・言い換え表現対応付けについて

(a) 形態素解析結果を XML 表現に変換し、言い換え表現を対応付ける方式を考案した。

(b) 言い換え表現対応付けの精度について実験的に確かめることが必要である。

７．まとめおよび今後の予定

記事を XML で構造化することによる、言い換え表現の対応付け方式を考案した。記事の形態素一つ一つに情報を与え、さらに形態素統合により文の構造を示すことは、今後の研究にとって大いに役立つと考える。

今後は、記事の XML 化について、より扱いやすい形式を考案していこうと考えている。また、言い換え表現対応の手順を明確化し、言い換え表現対応プログラムの作成および精度・再現率の評価を行うとともに、言い換え表現のデータベース作成を目指す。また、要素及び属性を利用することによる、構文的な観点からの言い換え規則データベースについて検討する予定である。

８．参考文献

[1] 大森岳史, 金田崇宏, 増田英孝, 中川裕志：携帯端末向け記事とインターネット新聞記事の対応付け, 第 64 回情報処理学会全国大会講演論文集, pp.147-148(2002).

[2] 形態素解析システム茶筌：http://chasen.naist.jp/hiki/ChaSen/

[3] 徳永健伸：情報検索と言語処理

[4] 言語処理学会第 7 回年次大会ワークショップ論文集 (2001).