字面解析による助詞「が」の抽出

1W－10

菅沼 明† 下園 幸一‡ 牛島 和夫†
†九州大学 工学部 情報工学科
‡九州大学 情報処理教育センター

1 はじめに

我々の研究室では、日本語文章推敲支援ツール『推敲』の開発を行なっている[1, 2]。このツールは、機械可読な日本語文章を字面だけで解析して、推敲に役立つ情報を書き手に提供するものである。

我々は『推敲』に使用するために、推敲に役立つ情報を字面解析だけで抽出する方法を構築してきた[3, 4]。それらの抽出法は実際の日本語文章を調査し、その結果を参考にして構築してきた。字面解析はオンメモリでの処理が可能なので、パソコンでも高速な解析が可能である。

本稿では、活用語尾表と漢字表を利用した字面解析手法を用いて日本語文章を解析する方法について述べる。さらに、この方法を助詞「が」の抽出に適用した結果について報告する。

2 字面解析手法

『推敲』には指示詞、受身、接続助詞「が」、否定表現、とりたて詞(副助詞、係助詞の一部)といった文法的な意味を持つ単語を抽出する機能がある。それらを抽出するために、個々に字面解析手法を構築してきた。それらは、文字列照合を基本としているが、照合の後に文字に関してのいくつかの条件を付加しているものもある。例えば、接続助詞「が」の抽出法では表1のような判定条件を設けている[3, 5]。

字面解析手法による抽出の精度を以下の式で表すと、

$$\text{再現率} = \frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{文章中の抽出すべき対象の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{抽出法で得られる候補の数}} \times 100(\%)$$

これまでに構築してきた字面解析手法では、第一種の誤り(指摘に洩れがある)を犯さないので、再現率は100%である。また、第二種の誤り(指摘すべきでないものまで指摘してしまう)を犯すことはある程度許しているので、適合率は約90%である。

表1: 接続助詞「が」の抽出法

判定条件1	「が」の1文字前が「う、く、す、つ、ぬ、む、る、ぐ、ぶ、い、だ、た、ん」のいずれかである場合、その「が」は接続助詞である。
判定条件2	「が」の1文字後が促音、撥音である場合、その「が」は接続助詞でない。
判定条件3	「が」の1文字前が「つ」であるとき、その「つ」の1文字前が数字または漢数字であれば、その「が」は接続助詞でない。
判定条件4	「が」の1文字前が「う」であるとき、その「う」の1文字前が「ほ」であれば、その「が」は接続助詞でない。
判定条件5	「が」の1文字前が「だ」であるとき、その「だ」が文頭の文字であれば、その「が」は接続助詞でない。

3 活用語尾表および漢字表

用言を語幹と活用語尾とに分けて考える。語幹は漢字で終るものもあれば、ひらがなで終るものもある。漢字は表意文字であるので漢字によって動詞になるもの、形容詞になるものなどの特徴があると考えると、漢字で終わる語幹について品詞の推定を行なうことが可能である。用言のうち語幹が漢字で終わる単語を公用データベース日本語単語辞書[6]を用いて調査し、漢字と品詞の組を表にした(大きさ約8KB)。この表を漢字表と呼ぶ。

活用語尾に関しては、用言の活用語尾と助動詞の活用を登録した活用語尾表を用意した。この活用語尾表には品詞と活用形とを登録している(大きさ約8KB)。

これら2つの表を使用して以下のようにして品詞と活用形を推定する。この処理を始めるに当たって推定を開始する文字の位置をパラメータとして与える。

1. 与えられた文字位置からテキストを遡ってスキャンし、活用語尾表と文章中の文字列とで一致するものがあるかを調べる。
2. 1の結果、一致するものがあれば活用語尾表に登録してある品詞と活用形を一時的に確定しておく。一致するものがない場合には活用語尾ではないとして推定を終了する。
3. 2で活用語尾とした文字列の1文字前の文字を調べ、漢字である場合は漢字表を引く。ひらがなである場合は2で一時的に確定した品詞と活用形を確定させて、推定を終了する。
4. 漢字表に登録されている品詞と、2で一時的に確定

Extraction of Particle "GA" with a Textual Analysis
Akira SUGANUMA†, Koichi SHIMOZONO‡ and Kazuo USHIJIMA†
†Department of Computer Science and Communication Engineering, Kyushu University
‡Educational Center for Information Processing, Kyushu University

した品詞とが一致した場合には品詞と活用形を確定させて、推定を終了する。一致しない場合は活用語尾ではないとして推定を終了する。

4 助詞「が」の抽出への適用

前節で構築した判定法(以下「活用チェック法」と記す)を使用して接続助詞「が」を抽出する方法を構築する。

表1にある判定条件1と3は文字「が」の前の文字が終止形になりうるかを判定するものであるので、この条件を活用チェック法に置き換える。判定条件4と5も文字「が」の前の文字が終止形になりうるかを判定するものであるが、「う、だ」ともに1文字の助動詞が存在するので、これらの条件はそのまま使用する。判定条件2に関しては、文字「が」の後ろの文字に関する条件なので、これもそのまま使用する。

活用チェック法と判定条件を組み合わせた抽出法を使用して、実際の文章中から接続助詞「が」を抽出した。調査対象の文章は我々の研究室で蓄えている以下の機械可読な日本語文章である。

67万字文章: 我々の研究室で書かれた科学技術論文(総文字数669,842文字)。この文章を調査した結果を用いて表1に示した判定条件を構築した。

200万字文章: 朝日新聞記事データ(総文字数1,981,950文字)

表1の判定条件1～5のみで抽出した場合と、活用チェック法と判定条件とを組み合わせて抽出した場合とで抽出の精度を比較した。その結果、どちらの抽出法でも再現率は100%であった。また、2つの抽出法の適合率の比較を表2に示す。この表からもわかるように、判定条件だけでは取り除くことができなかった第二種の誤りを活用チェック法によって取り除けるようになっている。新たに取り除かれたものは「振舞いが、違いが、思いが…」などで、活用チェック法を構築する際に取り除くことを期待した第二種の誤りをふるい落としている。

表2: 接続助詞「が」の抽出結果

調査対象文章: 67万文字文章

項　目	数	適合率
「が」の総数	6,987	—
判定条件1～5を満たす候補	438	90.9%
活用チェック法で抽出する候補	420	94.8%
接続助詞「が」	398	—

調査対象文章: 200万文字文章

項　目	数	適合率
「が」の総数	30,570	—
判定条件1～5を満たす候補	3,799	85.3%
活用チェック法で抽出する候補	3,604	89.9%
接続助詞「が」	3,241	—

文中に出現する文字「が」から助詞でない「が」の候補と接続助詞「が」の候補を取り除くことで、格助詞「が」の抽出を行なった[5]。その際、接続助詞「が」の候補の抽出法は活用チェック法を使用した方法を用いた。また、助詞でない「が」の候補の抽出は文献5で与えた方法を用いた。調査の結果を表3に示す。

接続助詞「が」の抽出法で取り出す候補の中に格助詞「が」が含まれているために、上に示した方法で格助詞「が」を抽出すると第一種の誤りを犯してしまう。そのため、再現率は98.9%であり、適合率は94.4%である。

表3: 格助詞「が」の抽出結果

分　類	数
格助詞「が」の候補	25,746
候補中の格助詞「が」	24,298
第一種の誤り	266
第二種の誤り	1,448
文章中の格助詞「が」	24,564

5 おわりに

活用語尾表と漢字表を用いて活用語の活用形を判定する方法を構築し、それを使用して接続助詞「が」の抽出を行った。その結果、以前の判定条件だけによる抽出に比べて約5%の適合率の向上が見られた。

現在はまだ、助動詞の取り扱いが不十分で、活用形の判定の際に第二種の誤りを多く含んでしまう。特に1文字の助動詞の判定に際しては、その助動詞と前の単語との接続関係を判定する処理を行なう必要があると考える。

謝辞

朝日新聞ニューメディア本部には、新聞記事データの使用を許していただいた。ここに記して謝意を表する。

参考文献

[1] 牛島和夫, 日並順二, 尹志熙, 高木利久: "日本語文章推敲支援ツールのプロトタイピング", コンピュータソフトウェア, Vol.3, No.1, 1986, pp.35–46.

[2] 倉田昌典, 菅沼明, 牛島和夫: "日本語文章推敲支援ツール『推敲』のパソコン上での実用化", コンピュータソフトウェア, Vol.6, No.4, 1989, pp.55–67.

[3] 菅沼明, 牛島和夫: "日本語文章推敲支援ツール『推敲』における字面解析手法とその評価", 情報処理学会自然言語処理研究会, No.68, 1988, 68–8.

[4] 菅沼明, 倉田昌典, 牛島和夫: "日本語文章推敲支援ツール『推敲』における否定表現の抽出法", 情報処理学会論文誌, Vol.31, No.6, 1990, pp.792–800.

[5] 下園幸一, 菅沼明, 牛島和夫: "日本語文章推敲支援ツール『推敲』における助詞「は」と「が」の抽出について", 情報処理学会自然言語処理研究会, No.94, 1993, 94–6.

[6] 吉田将, 日高達, 稲永紘之, 田中武美, 吉村賢治: "公用データベース日本語単語辞書の使用について", 九州大学大型計算機センター広報, Vol.16, No.4, 1983, pp.335–361.