日本語文章推敲支援ツール『推敲』における助詞「が」の抽出について

3L−2

下園 幸一　菅沼 明　牛島 和夫
九州大学工学部情報工学科

1. はじめに

我々の研究室で研究開発している日本語文章推敲支援ツール『推敲』は、機械可読な日本語文章を字面だけで解析して推敲作業に役立つ情報を書き手に提供するツールである[1]。本稿では、以前構築した接続助詞「が」の抽出法[2]を基に、格助詞「が」を抽出する方法について述べる。その際、字面解析だけでなく、用言のみを要素として持つ辞書を用いることを検討した。

2. 助詞「が」の問題点

助詞「が」には接続助詞と格助詞がある。接続助詞「が」は、順接、逆接、ただ2つの句をつなぐだけ、という3つの用法を持っている。つまり、接続助詞「が」は、どのような関係にある2つの句でも接続することができる。この性質により、接続助詞「が」を文章中に用いると書き手の意図が読み手に正しく伝わらないことが起こりうる。

格助詞「が」は体言に付いて、その体言が用言に対して主格の関係にあることを示す。この格助詞が1文中に複数出現することは、主語と述語の関係が2つ以上存在することになる。このような文は読みにくくなる可能性がある。また、助詞「は」と「が」の混在する文も分かりにくい文となることが多い[3]。

3. 格助詞「が」抽出法の構築

3.1 格助詞「が」抽出の問題点

現在までに我々が構築してきた字面解析手法は、文中のある特定の文字列(格助詞「が」の抽出の場合では文字「が」)に注目して、その文字列の前後の文字に条件を付けることによって、候補であるかどうかを決定する。この条件は学校文法にある単語の接続条件や実際の文章での調査結果を利用して設けてきた。

格助詞「が」は体言につく。つまり、格助詞を表す文字「が」の前に来る文字は体言の末尾となりうる文字ならば何であってもよいので、その文字から格助詞かどうかを判定することはできない。また「が」の後の文字からも同様に判定することはできない。そこで、文章中に現れる文字「が」を**格助詞**、**接続助詞**、**その他**に分類し、**接続助詞**と**その他**を落すことによって**格助詞**「が」を抽出することを考えた。

3.2 文字「が」の調査

まず、日本語文章中における文字「が」を調査した。調査に使用した文章は我々の研究室で貯えている以下の機械可読な日本語文章である。

67万文字文章: 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

350万文字文章: 朝日新聞記事データ。総文字数3,494,993文字

調査結果を表1, 2に示す。これより、文章中に現れる文字「が」の8割〜9割が、格助詞「が」であることがわかった。

表1: 文字「が」の分類(67万文字文章)

品詞	個数	割合(%)
格助詞	6,249	89.6
接続助詞	401	5.7
その他	328	4.7
合計	6,978	100.0

表2: 文字「が」の分類(350万文字文章)

品詞	個数	割合(%)
格助詞	42,579	80.0
接続助詞	5,717	10.7
その他	4,899	9.2
合計	53,195	100.0

表3: その他の「が」の内訳(350万文字文章)

分類	個数
「〜ながら」	913
「が」の後が促音、撥音	738
「わが」	437
文頭の「だが」	360
「ところが」	212
「上がる」の活用形	195
文頭の「が」	140
その他	1,904
合計	4,899

3.3 その他の「が」の除去

350万文字文章中にはその他の「が」が、4,899個ある。その内訳を表3に示す。このうち文章中の「が」の後に促音、撥音がくる場合は、明らかにその「が」は助詞でない。したがって、この「が」は助詞の候補から外すことができる。

今回の調査に使用した文章では、文字列「ながら」として出現する文字「が」のうち助詞であるものは1つもなかった。そのため、この文字列が出現した場合それを助詞の候補から外すことにする。また、文頭の「が」は、接続詞「が」または、単語の先頭の「が」と考えられるので候補から外す。文頭にある「だが」は、全てが接続詞「だが」であった。これも助詞の候補から外すことにする。

3.4 接続助詞「が」の抽出

以前、我々の研究室では、接続助詞「が」の候補を字面だけの解析で抽出する方法を構築した[2]。しかし、接続助詞「が」の抽出法が完全でない(接続助詞でない「が」も候補に含んでしまう)ため、そのまま使用したのでは抽出すべき格助詞「が」も候補から落してしまう。このため、接続助詞「が」の抽出法を再検討した。

350万文字文章から、まず、以前に構築した抽出法を使用し

Extraction Method of a Particle "が" (GA) in the Writing Tools for Japanese Documents.
Koichi SIMOZONO, Akira SUGANUMA and Kazuo USHIJIMA
Dept. of Comp. Sci. and Comm. Eng., Kyushu University

て接続助詞「が」の候補を抽出した。その結果、候補中に接続助詞「が」が5,717個、格助詞「が」が883個、その他の「が」が193個含まれていた。この抽出法では接続助詞「が」の指摘漏れを起こさないため、再現率(候補中に含まれる接続助詞「が」の数÷文章中の接続助詞「が」の数)は100%になる。また、適合率(候補中に含まれる接続助詞「が」の数÷接続助詞「が」の候補の数)は84.2%となった。この抽出法をそのまま格助詞の抽出法に適用すると、883個の格助詞「が」を格助詞の候補から落してしまう。

883個の誤りの内で多かったものは、「が」の前に「ん」がきた場合(372個)、「が」の前に「い」がきた場合(356個)であった。「が」の前に「ん」がきた場合、その「が」が接続助詞であるためには、「ん」は否定を表す助動詞でなければならない。しかし、現状の接続助詞「が」の抽出法のままでは、敬称の「さん」に格助詞「が」が接続したものを接続助詞「が」の候補として抽出してくる。これまでに我々は、否定表現の抽出法を構築した[4]。この抽出法を「が」の前の「ん」に適用した。

また、「が」の前に「い」がきた場合、「が」が接続助詞であるためには「い」は形容詞変化をする用言の終止形でなければならない。現状の抽出法をそのまま利用した場合、ワ行五段動詞が名詞化したもの(例えば、「違い」)を抽出してしまう。このため我々は用言のみを要素として持つ辞書を作成し、「が」の前の「い」が形容詞変化をする用言の終止形と判定できる場合だけ、その「が」を接続助詞の候補とした。

この結果、接続助詞「が」の抽出法によって誤って抽出する格助詞「が」の数を883個から332個にまで減らすことができた。また、その他の「が」も193個から174個になった。これにより接続助詞「が」の抽出法の適合率は91.9%となった。

4. 用言辞書の構築

本稿の抽出法に使用する辞書は、文全体を形態素解析するために使用するのではなく、ある一部だけを解析するために使用するものである。また、『推敲』は、その開発方針に、"実用規模の文章を待ち遠しくない時間で処理して欲しい。"ということを掲げている[1]。

一般に、形態素解析や仮名漢字変換に用いられる機械可読辞書は、その内容のほとんどが、名詞(サ変名詞も含む)である。我々の研究室で利用できる辞書[5](見出し語約9万語)について言えば、名詞が80.0%、動詞が9.9%、副詞が4.5%、形容動詞が3.5%、形容詞が1.0%であった。しかし、名詞は、実世界でどんどん増えていく。そのため、辞書に含まれない単語が多く存在する。一方、用言はあまり増えないと考えられる。これより、ある抽出法を構築する際に、"名詞に接続する"といった条件ではなく、"用言に接続する"という条件を作ることができれば、用言のみを要素として持つ辞書を作成すればよい。また、複合語の場合は、その基本となる部分だけ(「食い違う」ならば「違う」)を要素とし、基本となる部分が同じ複合語は省いた。これにより、辞書の大きさを小さくする(約1/20)ことができるので、2次記憶を使わずに高速に辞書を検索することができると考えられる。この方針をもとに、研究室で利用できる辞書から用言辞書を構築した。この用言辞書の見出し語数は、約4,100語である。

本稿の場合にでは、接続助詞「が」は、用言、および、助動詞の終止形につく。そのため、上記用言辞書を用いて十分な精度で解析できると考えた。

表4: 格助詞「が」の抽出結果(350万文字文章)

分類	個数
格助詞「が」の候補	44,572
候補中の格助詞「が」	42,247
第一種の誤り	332
第二種の誤り	2,325
文章中の格助詞「が」	42,579

表5: 格助詞「が」の抽出結果(200万文字文章)

分類	個数
格助詞「が」の候補	26,287
候補中の格助詞「が」	24,406
第一種の誤り	168
第二種の誤り	1,881
文章中の格助詞「が」	24,574

5. 格助詞「が」の抽出結果

前述の接続助詞「が」の抽出とその他の「が」の抽出を利用して格助詞「が」を抽出した結果を表4, 5に挙げる。ここで、第一種の誤りとは、抽出してこなかった格助詞「が」の数であり、第二種の誤りとは、候補に含まれてしまった格助詞「が」以外の数である。この数をもとにして再現率を計算すると99.2%になる。また適合率は93.9%である。

また別の朝日新聞記事データ(総文字数1,981,950文字)で格助詞「が」の抽出法の評価を行なった。再現率は99.3%であり、適合率は92.8%であった。抽出法構築に使用した文章の場合と同等の結果が出た。

6. まとめ

実際の文章に現れる文字「が」を調査し、その「が」の8〜9割が格助詞であることがわかった。このことから、格助詞以外の文字「が」を候補から外すことにより格助詞「が」の抽出法を構築した。今後、今回構築した用言辞書を用いてさらに他の抽出法の構築を考えていく。

参考文献

[1] 倉田昌典他:"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, pp.55-67, (1989)

[2] 菅沼明他:"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", 自然言語処理研究会報告, No.68, 68-8, (1988)

[3] 下園幸一他:"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", 自然言語処理研究会報告, 94-7, (1993)

[4] 下園幸一他:"字面解析手法を用いた否定表現抽出方法の評価 — 朝日新聞記事データへの適用 —", 第44回情報処理学会全国大会論文集, 5C-4, (1992)

[5] 吉田将他:"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, pp.335-361, (1983).