自然言語処理 94－6
（1993. 3. 19）

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

下園 幸一　菅沼 明　牛島 和夫
九州大学工学部情報工学科

文章の推敲において、助詞「は」と「が」の使い分けや、あいまいな接続助詞「が」の指摘は有用である。日本語文章推敲支援ツール『推敲』で使用している抽出法は、実際の日本語文章を調査し、その結果を基に字面のみによる手法として構築している。今回、用言だけを要素として持つ辞書を構築し、それと字面解析手法を使って、新聞記事データから接続助詞「が」の抽出を試みた。また、文章中に現れる文字「が」から助詞「が」でない「が」と接続助詞「が」とをとり除くことによって、主格を表す格助詞「が」を抽出する方法を考案した。また、以前構築したとりたて詞「は」の抽出法と格助詞「が」の抽出法とを組み合わせることによって、「は」と「が」を複数含む文を高速に抽出して書き手に提示することができる。

Extraction Methods of Particles “は” (WA) and “が” (GA) in the Writing Tools for Japanese Documents.

Koichi SHIMOZONO, Akira SUGANUMA and Kazuo USHIJIMA
Department of Computer Science and Communication Engineering, Kyushu University

A system of writing tools SUIKOU analyses a machine-readable Japanese document only textually and provides writers with the useful information for polishing it. We compiled a small machine dictionary whose elements were verbs and adjectives alone, and linked it with the textual analysis method we constructed. The new analysis method is adopted to extract the conjunctive particle “GA”. We designed a method extracting the subjective particle “GA” using the extraction method of the conjunctive particle “GA”.

How to use the particle “GA” and “WA” is difficult. Thus, to point out those particles in a document is helpful for polishing Japanese documents. Using both the extraction method of the subjective particle “GA” and that of the particle “WA” which we constructed, we are able to search and display sentences containing more than one particle “GA” and/or “WA” very quickly.

1. はじめに

日本語ワードプロセッサの普及は著しい。この日本語ワードプロセッサの基本機能は、文章の入力、書式の設定、印刷、文章の保存である。文章が機械可読な日本語テキストになっているにももかかわらず、高度なテキスト処理は行われていない。我々は、文章の推敲作業に着目し、日本語文章推敲支援ツール『推敲』を開発してきた[1]。

文章を推敲する際、書き手は、文章を読み返して問題となる箇所を探し、その部分を検討して、必要であれば書き直すといったことを行う。この作業を全て計算機で行うのは困難である。そこで、「問題となる箇所を探す」という部分を計算機で支援し、書き手は計算機が指摘したものを吟味して必要があれば書き直すという方法をとれば、書き手の推敲作業をより質の高いものにできると考えた。

そこで、我々は『推敲』を開発するにあたり、

(1) 文章中に問題となりそうな箇所があればそれを指摘できればよい。(実際に推敲するのは書き手である)

(2) 実用規模(1万字程度: 図や表を含めると論文誌刷り上がり7～8ページの文字数)の文章を待ち遠しくない時間で処理して欲しい。

という方針を立てた。この方針に従い、現在『推敲』はパーソナルコンピュータ上に実現している。

『推敲』には指示詞、受身形、接続助詞「が」、否定表現などを抽出する機能がある。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。また、以前構築したとりたて詞「は」の抽出[3]と組み合わせた推敲支援についても述べる。

2. 助詞「が」について

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。以下にその例を示す。

順接: 今日は暖かかったが、明日も暖かいであろう。

逆接: 来いといったが、来ていない。

句をつなぐだけ: パッケージの仕様中にある関数IS_NULLであるが、これは抽象データ型listsで使うものである。

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[4]。

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」の使い分けは微妙であり、1文中に同じ助詞が繰り返されると耳障りであったり文が分かりにくくなったりするので推敲の対象として重要である。

3. 字面解析の精度

現在『推敲』で採用している日本語文章解析方法は、字面解析手法である。この字面解析手法は、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、抽出したいもの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。

文章を字面だけで解析する方法を採ったため、抽出精度は文法解析を行う場合よりも低いことが予想される。この抽出精度の指標として、情報検索の分野で使用されている再現率と適合率とを使用する。再現率、適合率は以下の式で定義される。

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

$$\text{適合率} = \frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{抽出法から得た候補の数}}$$

文章から問題となる箇所を抽出する場合に犯す誤りには2種類ある。第一種の誤り「指摘に洩れがある」と第二種の誤り「指摘すべきでないものまで指摘してしまう」である。第一種の誤りを犯せば、再現率が下がる。第二種の誤りを犯せば、適合率が下がる。現在までに構築してきた字面解析手法は、第一種の誤りを犯さないこととしているので、再現率は100%である。一方、『推敲』の開発方針(1)から、『推敲』が指摘したものは書き手が目を通すことになる。第二種の誤りを犯すことはある程度許容する。しかし、誤りは少なければ少ないほどよい。すなわち、再現率100%のもとで適合率をできるだけ高くできるような字面解析手法を構築してきた。

4. 格助詞「が」の抽出法の構築

4.1 格助詞「が」抽出の方針

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。そのため文字「が」の前の文字情報を利用してその「が」が格助詞であるかどうかを判定することはできない。また文字「が」の後の文字に注目してみると、「が」が格助詞であれば、次の文字は単語の先頭の文字、または句点である。そのため文字「が」の後の文字でその「が」が格助詞であるかどうかを判定することはできない。

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

文章中に現れる文字「が」で、格助詞以外の「が」を考えてみる。1つには、文字「が」を含む自立語がある。この自立語を公用データベース日本語辞書[5]で調べてみると約6,600個あった。さらに前述の接続助詞「が」がある。この接続助詞「が」の抽出法はすでに構築済である[2]。

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。

4.2 文字「が」の調査

まず、日本語文章中に出現する文字「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約1ヵ月分(総文字数3,494,993文字)

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

これらの文章を文字「が」をキーに KWIC 表示させ、目視でその「が」が何であるかを判定した。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。

4.3 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。

「が」を助詞であると仮定すると、「が」の後に続く文字は、自立語の最初の文字である。ところが、促音や撥音で始まる自立語はないことから、「が」の 1 文字後が促音か撥音の場合、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。これにより「～しながら」のような表現を候補から外すことができる。「ながら」という文字列を候補から落すことで、第一種の誤りを犯す可能性がある。

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。

4.4 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]。この抽出法を表 4に示す。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。すなわち、第一種の誤りは犯さないが、第二種の誤りを犯す。) 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。

以前に構築した抽出法を使用して 350 万文字文章から接続助詞「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、再現率は 100% であるが、適合率は 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 1 文字前の文字が「ん」である場合 (372 個)、「が」の 1 文字前の文字が「い」である場合 (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[6][7]。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

表 4: 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。

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。

5. 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。

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、その見出し語のほとんどが、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 19 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で十分解析が可能である。

用言辞書の構築の際に、複合語は、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。これにより、辞書の大きさを小さくすることができる (約 29KBytes) 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。

6. 格助詞「が」の抽出結果

前述のその他の「が」の除去と接続助詞「が」の抽出とを利用して格助詞「が」を抽出した結果を表 5に示す。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以外の「が」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。第一種の誤りは全て接続助詞「が」の抽出で接続助詞の候補となってし

表 5: 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 6: 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

まった格助詞である。第二種の誤りはその他の「が」である。

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。その結果を表 6に示す。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を得た。

7. とりたて詞「は」

とりたて詞「は」は、文中のある構成要素を取り出して提示する働きを持つ副助詞または係助詞である。この「は」の機能を分類すると題目 (提題) と対照との 2 つになる [8][9]。

題目の機能を果たす「は」は文のいろいろな成分に付いて、その成分が題目 (テーマ) であることを示す。

対照の「は」はある脈絡の中にあるものを指定して、それと関連するものを対比することで強調するものである。この対比されるものは文脈に明示されている場合もあれば、文脈に含まれている場合もある。

日本語文では、1 文中にとりたて詞「は」が複数含まれることが多くある。題目の「は」はその性質上 1 文に一つしか存在できないために、複数のとりたて詞「は」がある場合には、二つめ以降は対照の「は」と考えることができる。しかし、数が多くなると対比するものが多くなるために、分かりにくい文になりやすい。

以前、我々はとりたて詞「は」の抽出法を構築した [3]。その判定条件を表 7に示す。

判定条件 1 で、助詞「を」の後に単語の先頭が「は」である語を候補から外すことができる。「は」が助詞であることを仮定すると、「は」の後には自立語の最初の文字がこなければならない。促音、撥音で始まる自立語はないので、判定条件 2 を設けることができた。これにより、「はっきり、はん用、はんだ」などの文字「は」を助詞の候補から外すことができる。判定条件 3 および 4 により、「はじめに、はじまる」などを助詞の候補から外すことができる。この抽出法を新聞記事データに適用すると、再現率 100%, 適合率 97.5% で助詞「は」を抽出することができた。

8. 「は」と「が」について

今回構築した格助詞「が」の抽出法と、とりたて詞「は」の抽出法とを使って前述の 200 万文字朝日新聞記事データから「が」と「は」の候補を複数含む文の調査を行った。結果を表 8に示す。このうち、「が」または「は」、あるいは「が」および「は」が 1 文中に複数含まれる文は、10,459 文あり、全体の文の約 23.1% である。『推敲』が実用規模としている文章 (1 万字程度) に当てはめると、総文数は約 230 文に対して、「が」または「は」、あるいは「が」および「は」を複数含む文は、約 50 文となる。1 文中の「が」の数と「は」の数とを足したものが 4 個以上の文を抽出すれば、全体の文の約 5.0% (1 万字文章中約 12 文) となる。

『推敲』には、「1 文中の「は」の数と「が」の数とを足したものがいくら以上の場合のみ表

表 7: とりたて詞「は」の抽出法

判定条件 1:	「は」の 1 文字前が「を」である場合、その「は」は助詞でない。
判定条件 2:	「は」の 1 文字後が促音、撥音である場合、その「は」は助詞でない。
判定条件 3:	「は」の後に「じめ」と続いた場合、「め」の 1 文字後が「い、じ、つ、ん」のいずれかの場合だけ。その「は」を助詞の候補とする。
判定条件 4:	「は」の後に「じま」と続いた場合、「ま」の 1 文字後が「い、え、く、ま、わ、ん」のいずれかの場合だけ、その「は」を助詞の候補とする。

示させる」といった集合演算機能があるので、いくら以上を表示させるかをユーザが選べるようにしておくことも可能である。

以下に、新聞記事データから抽出したものの例を示す。下線つき文字は抽出法によって得られた候補である。各例の最後に実際の格助詞「が」、とりたて詞「は」の個数を示す。例 1、2 のような分かりにくいものもあれば、例 3 のような自然な文もある。

例文 1: 米国側には、しかし、その事実を承知のうえで、日本や欧州の輸出が伸びたのは、米国の強い個人消費と開放された市場があったためではないか、そちらこそ内需拡大、輸入増、市場開放をもっと進めるべきではないか、という考え方が根強い。(「は」4 個、「が」3 個)

例文 2: 高山病になりやすい人の条件には、さまざまな要素がからまり、高地順応性には個人差があって、いちがいには言えませんが、太った人が高山病になりやすいのはホントです。(「は」3 個、「が」3 個、第二種の誤り 1 個を含む)

例文 3: 民間企業の規模別では、従業員 1000 人以上の大企業の組織率が前年より 2.4 ポイント上がって 68% になったのに対し、100 人以上 1000 人未満は27.4%、30 人以上 100 人未満は6.2%、30 人未満は0.4% で、企業規模が大きいほど労組組織率が高い傾向がさらに進んだ。(「は」4 個、「が」4 個)

表 8: とりたて詞「は」と格助詞「が」の候補を複数含む文

1 文中の数	は		が	
	文の数	割合	文の数	割合
0	20,814	46.0	26,560	58.7
1	17,629	38.9	13,162	29.0
2	5,198	11.5	4,069	9.0
3	1,215	2.7	1,097	2.4
4	269	0.6	280	0.6
5 以上	75	0.2	107	0.2
合計	45,275	100.0	45,275	100.0

9. 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。『推敲』で用いられている他の抽出法と比較すると幾分(約 5% 程)抽出精度が悪い。これは、その他「が」を有効に候補から落

すことができなかったためである。また、開発方針から“第一種の誤りは犯してはならない”となっている。しかし、今回の抽出法では第一種の誤りをわずかながら犯してしまう。200万文字文章の結果を1万字の文章にあてはめてみると、格助詞「が」が122個含まれる、そのうち0.8個を見落とす、9.4個を誤って拾ってくる、ということになる。

用言辞書に関しては、大きさが十分小さいので、MS-DOSをOSとするパーソナルコンピュータでも主記憶に辞書を展開することができるので、これにより高速な検索ができると考えられる。パソコン版『推敲』に実装することが可能である。今回は、ワークステーション上でプロトタイプを作成し、それを用いて精度の評価を行った。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, 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, pp.55–67, 1989.

[2] 菅沼明, 石田朗子, 倉田昌典, 牛島和夫：“日本語文章推敲支援ツール『推敲』における字面解析手法とその評価”, 自然言語処理研究会報告, 68–8, 1988.

[3] 菅沼明, 牛島和夫：“日本語文章推敲支援ツール『推敲』におけるとりたて詞「は」の抽出法とその評価”, 情報処理学会論文誌, Vol.32, No.11, pp.1392–1400, 1991.

[4] 清水幾太郎：論文の書き方, 岩波新書, 1959.

[5] 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, pp.335–361, 1983.

[6] 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, pp.792–800, 1990.

[7] 下園幸一, 菅沼明, 牛島和夫：“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.

[8] 本多勝一：“日本語の作文技術”, 朝日新聞社, 1976.

[9] 中島文雄：“日本語の構造 — 英語との対比 —”, 岩波新書, 1987.