字面解析による用言の活用形推定について

菅沼, 明
九州大学システム情報科学研究院知能システム学部門

笠原, 晋
九州大学大学院工学研究科情報工学専攻:九州松下電器株式会社

牛島, 和夫
九州大学大学院システム情報科学研究科情報工学専攻

出版情報：情報処理学会論文誌. 37 (6), pp.1007-1016, 1996-06. 情報処理学会
バージョン：

字面解析による用言の活用形推定について

菅沼 明[†] 笠原 晋[††,☆] 牛島和夫[†††]

日本語文章は単語単位に分割して書かれないために，これを機械処理するには，まず形態素解析などの文法処理を行うのが普通である．しかし，文法処理を行っても，解が一意に定まらなかったり，解析時間がかかったりする問題がある．そのため，我々は，字面だけの情報で接続助詞「が」や否定表現，受身形などの候補を抽出する方法を構築し，それを応用して日本語文章推敲支援ツール『推敲』を開発している．本論文では，簡単な表を利用した字面解析手法を用いて日本語文章中にある活用語の活用形を推定する方法（「活用チェック法」と呼ぶ）について述べる．活用語を語幹と語尾に分け，それぞれについて表を設ける．文を後ろから前に遡る方向にスキャンしながら，まず語尾の表と比較を行う．その後，語幹の表との比較を行う．語幹に関しては最後の1文字しか評価しない．活用チェック法で使用する表の大きさは32 KB程度で，これを主記憶に載せたとしても，記憶域を圧迫するものではない．そのため，主記憶容量が小さな計算機であっても主記憶に載せることが可能である．従来の字面抽出法と比較するために，接続助詞「が」と否定表現の抽出に活用チェック法を適用し，抽出精度を比較した．その結果，従来の字面抽出法に比べて適合率が約10%向上した．再現率は100%を保っている．さらに，活用チェック法を使用した抽出法は，従来の抽出法と比較して，抽出に要する時間はほとんど変わらない．

Estimation of the Conjugation of Japanese Verbs and Adjectives with a Textual Analysis

AKIRA SUGANUMA,[†] SUSUMU KASAHARA[††,☆] and KAZUO USHIJIMA[†††]

Japanese documents are usually analyzed with the grammatical analysis like the morphological analysis. Since this analysis requires much time, we constructed an extraction method with only textual information for the passive voice, the negative expressions, the conjunctive particle "GA," and so on. And we are developing a system of writing tools SUIKOU with the textual analysis method. This paper describes a method to estimate the conjugation of Japanese verbs and adjectives with a textual analysis. We prepare two kinds of tables for a conjugational part and a non-conjugational part of Japanese verbs and adjectives. This method analyzes a Japanese document from tail to head of a sentence with these tables. Using the method, we construct a new extraction method for the conjunctive particle "GA" and the negative expressions. This extraction method guarantees to extract all requested items, but it may extract some unrequested items. The precision of the new extracting method for the conjunctive particle "GA" is above 98%, and is improved by 7% in comparison with that of old one, although the new method is able to extract the conjunctive particle "GA" as fast as old one.

1. はじめに

日本語文章は単語単位に分割して書かれないために，それを計算機で解析する場合には，形態素解析を行い単語単位に分割するのが普通である．しかし，形態素解析を行う際に用いる辞書の項目は有限であり，言語現象で出現する単語は無限にあるといってよい．そのため，文中に出現する単語がすべて辞書に登録されているとは限らない．この傾向は，名詞において強く現れる．この未知語の問題などがあるために，形態素解析を行っても必ずしも正解が含まれているとは限らない．

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

† 九州大学大学院システム情報科学研究科知能システム学専攻
Department of Intelligent Systems, Kyushu University
†† 九州大学大学院工学研究科情報工学専攻
Department of Computer Science and Communication Engineering, Kyushu University
††† 九州大学大学院システム情報科学研究科情報工学専攻
Department of Computer Science and Communication Engineering, Kyushu University
☆ 現在，九州松下電器株式会社
Presently with Kyushu Matsushita Electric Co., LTD

情報を書き手に提供するものである．このツールは，次の2つの方針のもとで開発を進めてきた．

(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）の文章を待ち遠しくない時間で処理してほしい．ユーザは，コマンドを入力して10秒も待たされると待ち遠しいと感じる．

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)~5)]．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，処理能力が小さい計算機であっても高速な解析が可能である．また，処理対象の文章が大規模になっても，大幅に待たされることなく処理が終了することが期待できる．

『推敲』に採用している字面解析手法では，活用語の後に続く語を抽出するものが多い．字面解析手法の構築を通して，文全体にわたる解析を行わなくても，文中に存在する用言の位置と活用形の情報が分かれば，推敲に役立つ情報の抽出が可能であることが分かった．また，用言の活用形を推定することができれば，今までに構築してきた字面解析手法の抽出精度が向上する．

本稿では，語幹の文字に関する表と活用語尾表とを利用した字面解析手法を用いて，日本語文章中にある用言の活用形を推定する方法について述べる．さらに，この方法を接続助詞「が」と否定表現の抽出に適用した結果について報告する．本稿で述べる手法は，『推敲』に組み込んだ従来の字面解析手法に比べ抽出精度が向上し，抽出に要する時間もほとんど変わらないという特徴を持つ．

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）のような文法的な意味を持つ単語を指摘する機能がある．それらを文章中から取り出すために，個々に抽出法を構築してきた．それらは，文字列照合を基本としているが，照合の後に前後の文字に関するいくつかの判定条件を付加しているものもある．この論文では，判定条件によるこれまでの抽出法を従来法と呼ぶ．たとえば，従来法による接続助詞「が」の抽出では，文章中に出現する文字「が」を捜し出し，**表1**に示す判定条件を満たすものだけを抽出してくる[4),6)]．

表1にある判定条件1は，接続助詞「が」が活用語の終止形に接続する性質を持つことから設けている．判定条件1にあげた文字は活用形の終止形の最後の文字である可能性があるので，判定条件1を満たす「が」を接続助詞の第一次近似の候補として取り出す．その際に接続助詞であるものを取りこぼすことはない．また，表1にある判定条件2~5は第一次近似の候補のうち接続助詞でないものをふるい落とすための条件である．これらの条件は実際の文章を調査して設けている．

表1 接続助詞「が」の抽出法
Table 1 The old method to extract the conjunctive particle "GA".

判定条件1	「が」が接続助詞であるためには，その1文字前が「う，く，す，つ，ぬ，む，る，ぐ，ぶ，い，だ，た，ん」のいずれでなければならない．
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「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．
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「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．

『推敲』では辞書を使わずに字面だけの情報で抽出を行っているため，抽出精度は文法処理を行った場合よりも悪くなることが予想できる．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を使用する．再現率と適合率は以下の式で定義される．

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

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

文章から問題となる箇所を抽出する際に犯す誤りには2種類ある．1つは第一種の誤り「指摘すべきものを取りこぼしてしまう」で，もう1つは第二種の誤り「指摘すべきでないものまで指摘してしまう」である．これら2つの誤りは再現率，適合率と密接な関係がある．抽出の際に第一種の誤りを犯せば，再現率が下がり，第二種の誤りを犯せば，適合率が下がる．『推敲』に組み込んでいる字面解析手法は，第一種の誤りを犯さないので，再現率は100%である．また，『推敲』の開発方針(1)から，第二種の誤りはある程度許容している．そのため，抽出するものにもよるが，適合率は約90%である．しかし，第二種の誤りも少なければ少ないほどよいので，再現率100%のもとで適合率をで

きるだけ高くするような抽出法の構築を行ってきた．

3. 表を用いた字面解析

この章では，表を用いた字面解析手法（活用チェック法と呼ぶ）に関して述べる．この方法は，字面の情報から用言の活用形を推定するものである．

3.1 構築の方針

表1の5つの条件のうち判定条件2を除く4つの条件は，文字「が」の直前にある文字列が活用語の終止形であるか否かを判定するためのものである．しかも，文字「が」の直前の文字列が活用語の終止形である場合は，すべてこの判定条件を満たす．また，『推敲』に組み込んでいる他の字面抽出法の判定条件も，抽出すべきものの前側にある単語の活用形を文字から判断するものが多く含まれている．

第一種の誤りを犯さないという『推敲』の開発方針に従うと，活用形を推定する方法に要求されることは，たとえば，終止形と答えてほしい場合，推定対象の文字列が終止形ならば必ず終止形と答えることである．つまり，各活用形に関して十分条件を設定しなければならない．そのため，推定対象の文字列によっては，複数の活用形を答えることはある程度容認する．

表1にあげた接続助詞「が」の字面抽出法のうち判定条件1は活用語の終止形を抽出するために設けている．この条件は文字「が」の直前の1文字だけで，候補を絞り込もうとしているのである．しかし，1文字の情報だけで絞り込むのではなく，複数の文字で判定を行うようにすれば推定の精度を上げることができる．この複数の文字に関する条件を表の形で保持し，文章中の文字を調べる際にその表を参照する．

3.2 語幹に関する表

まず，活用語のうち用言について考える．用言は語幹と活用語尾に分かれるものが大部分を占めているので，表に登録する際にも語幹と活用語尾に分けて登録する．

昭和56年に出された内閣告示第三号「送り仮名の付け方」によると，送り仮名は活用語尾を送るのが通則とされている．そのため，用言を漢字仮名混じりで書く場合，語幹の最後の文字は漢字であることが多いと考えられる（たとえば，動詞「見分ける」の場合，「見分」が語幹であるので文字「分」が語幹の最後の文字となる）．

漢字によって動詞として使われるもの，形容詞として使われるものなど，使われ方に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[7]を用いて調査した．JIS第一水準漢字2,965字に関して用言のどの品詞として使われるかを調べると，**表2**のようになる．

表2 漢字で終わる語幹の分類
Table 2 The classification of Kanji characters.

語幹の品詞	漢字数	割合（%）
動詞のみ	1,204	40.6
形容詞のみ	11	0.3
形容動詞のみ	142	4.8
動詞，形容詞	18	0.6
動詞，形容動詞	417	14.1
形容詞，形容動詞	15	0.5
動詞，形容詞，形容動詞	52	1.8
語幹の最後にはならない	1,106	37.3
合　計	2,965	100.0

表中の「動詞のみ」とは，ある漢字が動詞の語幹の最後の文字にはなるが，形容詞，形容動詞の語幹の最後の文字にはならないことを意味している（たとえば，文字「逢」が語幹の最後に現れるならば，ワ行五段動詞「逢う」にしかならない）．また，「動詞，形容詞」は，ある漢字が動詞と形容詞の語幹の最後の文字にはなるが，形容動詞の語幹の最後の文字にはならないことを意味している（たとえば，「煙る」(ラ行五段動詞)，「煙い」(形容詞)）．

表2から，1,357種類の漢字（45.7%）は，動詞か形容詞，形容動詞のいずれかの語幹の最後であることが一意に定まることが分かる．また，1,106種類の漢字（37.3%）は用言の語幹の最後の文字にはならない．ただ，502種類の漢字（17%）だけは，2つまたは3つの品詞の語幹の最後の文字となりうるので，文字だけでは動詞，形容詞，形容動詞のいずれの語幹の最後の文字になりうるかの判断がつかない．このことから，83%の漢字は，漢字だけの情報で語幹の品詞（動詞，形容詞，形容動詞または用言にはならない）を推定できることになる．この調査結果をもとにして，漢字と品詞を組にして表を作成した（大きさ約12KB）．この表を漢字表と呼ぶ．品詞は，動詞，形容詞，形容動詞を登録し，動詞はさらに五段活用（活用の行も区別），上一段活用，下一段活用，カ変，サ変，ザ変の活用型も区別して登録した．

語幹の最後の文字が平仮名である用言と，用言が平仮名書きされた場合とを処理するために，平仮名をキーとして引く平仮名表を漢字表と同様にして作成した（大きさ約0.4KB）．品詞を登録する際に，平仮名書きされた用言も登録しているので，1つの平仮名文字に対して登録してある品詞の数が多くなっている．そのため，平仮名表は，品詞の候補を絞るのに漢字表

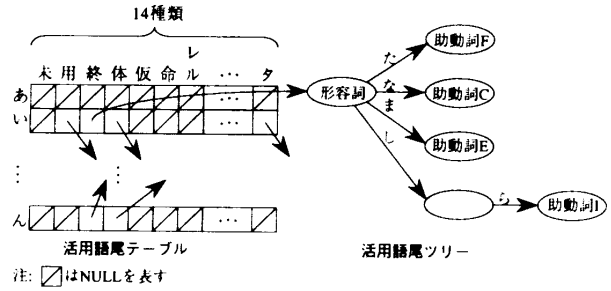

図1 活用語尾表

Fig. 1 A table for conjugational part of Japanese verbs and adjectives.

ほどは有効ではないと考える．しかし，用言の語幹の最後の文字になりえない平仮名も明かに存在するので，平仮名表を用意して語幹の品詞の判定に用いる．

3.3 活用語尾に関する表

活用語尾の処理では，文字列から品詞と活用形を求めたい．また，助動詞の活用形も用言の活用語尾と同様の処理が必要である．そのため，用言の活用語尾と助動詞の活用形とを登録した表を作成する（この節では，助動詞の活用形と用言の活用語尾の両方をあわせて活用語尾と記す）．この表を活用語尾表と呼ぶ（全体で大きさ約19.6KB）．活用語尾表は**図1**の構造を持っており，活用語尾テーブルと活用語尾ツリーからなる．

3.3.1 活用語尾テーブル

図1にあるように，活用語尾テーブルは文字と活用形とでインデックスをつけた表で，活用語尾の最後の文字からその文字列がどのような活用形になることが可能かが引けるようになっている．もし，ある文字がある活用形となることが可能であれば，そのノードには活用語尾ツリーへのリンクが登録されている．このテーブルを参照してNULLが登録されていれば，その文字で終わるその活用形はないことを表している．たとえば，図1では，文字「あ」の終止形の欄はNULLとなっている．これは，文字「あ」で終わる終止形がないことを意味している．また，文字「い」の終止形の欄には活用語尾ツリーへのリンクが張ってある．これは，文字「い」で終わる終止形があり，その情報が活用語尾ツリーに保存されていることを意味している．

3.3.2 活用語尾ツリー

活用語尾ツリーはTRIE構造をしており，活用語尾の最後の文字以外を登録している．このツリーのノードには，そのノードに到達した場合にどんな品詞の活用語尾になりうるかが登録してある．また，文字列を後ろから前に遡る方向でツリーを作成している．そのため，ある文字から文頭の方向に遡った文字列が活用語尾ツリーに登録してあれば，その文字列の品詞と活用形を判断できる．

文字「い」で終わる終止形の活用語尾には形容詞「い」，助動詞「たい」，助動詞「ない」，助動詞「まい」，助動詞「らしい」の5種類がある．そのため，図1にあるように，活用語尾テーブルの文字「い」の終止形の項目が指すノードは「た，な，ま，し」の遷移がある．文字「い」だけで形容詞の活用語尾になりうるので，そのノードには形容詞が登録されている．「し」の遷移をたどると，「しい」だけでは活用語尾とはならないので，そのノードには何も登録されていない．しかし，そのノードからは「ら」の遷移があり，その遷移先のノードには助動詞が登録されている．このような形式をとることで，文章中の文字列を遡ってあるノードまで到達した場合，それまでに辿ったノードに登録してある品詞の可能性があることが判断できる．

上に述べたような情報を活用語尾表に登録して活用語の活用形の推定を行う．その際，用言のうち語幹と語尾の区別がないものは，他の用言とは異なる処理が必要となる．語幹と語尾の区別がない用言には上一段動詞または下一段動詞「居る，射る，着る，似る，煮る，干る，見る，得る，寝る，経る，出る」がある．これらは，語幹の処理をせずに活用語尾の処理だけで活用形を推定する必要がある．そのため，これらを活用語尾ツリーに登録する際に，他の上一段動詞，下一段動詞とは区別して登録する．

3.4 表を使用した品詞の推定法

3.2節，3.3節で述べた3つの表（漢字表，平仮名表，活用語尾表）を使用して，以下の手順で品詞と活用形の推定を行う．この処理を始めるにあたって推定を始める文字の位置と要求する活用形をパラメータとして与える．

(1) 推定開始位置の文字と要求された活用形とで活用語尾テーブルを参照し，活用語尾ツリーへのリンクがあるか否かを調べる．もし，活用語尾テーブルにNULLが登録されていれば，推定開始位置より前側の文字列は要求された活用形ではないとして推定を終了する．

(2) 推定開始位置からテキストを遡ってスキャンし，文章中の文字列が活用語尾ツリーに登録されているか否かを調べる．文字列が登録されていれば，活用語尾ツリーに登録してある品詞を一時的に保持しておく．登録されてい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) 手順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，平仮名

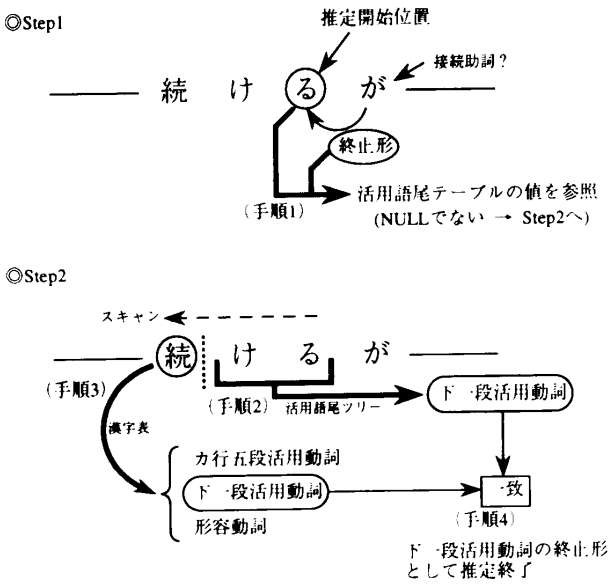

図2 活用チェック法の動作

Fig. 2 To estimate the conjugation of Japanese verbs and adjectives.

である場合は平仮名表を引く.

(4) 漢字表もしくは平仮名表に登録されている品詞と，手順2で一時的に保持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.

図2は，接続助詞「が」の抽出を例として活用チェック法の動作を表したものである．文章中にある文字「が」が接続助詞であるか否かを調べるために，文字「が」の1文字前の文字「る」を推定開始位置として活用チェック法を始める．その際，接続助詞「が」は活用語の終止形に接続するので，要求する活用形は終止形となる．手順1で文字「る」と終止形の対で活用語尾テーブルを引くと活用語尾ツリーへのリンクがあるので手順2へ進む．手順2において，活用語尾ツリーに登録されている文字列と照合すると，文字列「ける」が下一段動詞として登録されている．そのため，活用語尾の推定結果として，文字列「ける」を下一段動詞の終止形と一時的に確定する．手順3で，文字列「ける」の1文字前の文字「続」を用いて語幹の品詞の推定を行う．漢字表を調べると,「続」にはカ行五段動詞，下一段動詞，形容動詞が登録してある．最後に手順4によって，活用語尾と語幹との推定結果の一致を調べる．この場合，活用語尾，語幹の推定結果のどちらにも下一段動詞が存在するので，下一段動詞の終止形として確定して終了する.

活用チェック法では，文章中の数文字だけを参照し，その情報によって品詞と活用形を推定している．活用語尾表に登録されている文字列の長さは最大で4文字である．そのため語幹の処理を含めても5文字程度の文字の参照で品詞と活用形を推定していることになる．このような推定法をとっているので，活用チェック法は文全体にわたって単語解析を行うわけではない．一方，形態素解析は文全体を解析して品詞などを決定する．形態素解析では文中に存在するすべての単語の候補が判明する．この点において活用チェック法と形態素解析は異なる.

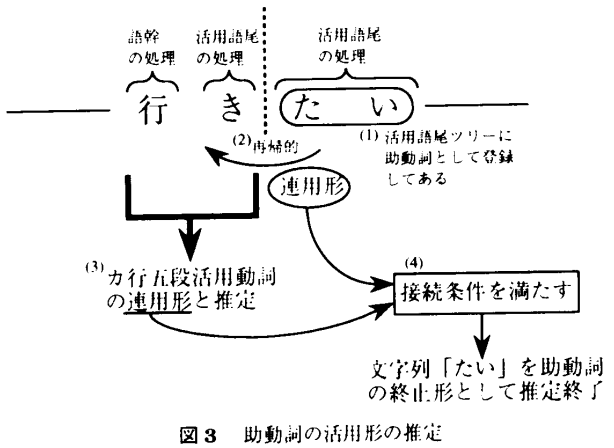

図3 助動詞の活用形の推定

Fig. 3 To estimate the conjugation of Japanese auxiliary verbs.

3.5 助動詞の取扱い

助動詞はすべて平仮名で記述され，1文字あるいは2文字のものが多い．また，助動詞の活用形は活用語尾表のみに登録されているため，活用語尾の判定と語幹の判定との照合（3.4節の手順4）を行わない．これらが原因で，助動詞の推定においては第二種の誤りを犯すことが多い．そのため，活用チェック法で助動詞を処理する際には，助動詞の候補とする文字列の前側の文字列をさらに検査して，活用形を推定する.

助動詞の多くは，その直前に活用語をともない特定の活用形を要求する（接続条件）．そのため，**図3**に示すように，助動詞の直前の活用語の活用形の推定を活用チェック法を用いて行う．再帰的に呼び出した推定の結果，接続条件を満たせば助動詞の推定が成功したとし，そうでなければ助動詞の推定は失敗したとして推定を終了する．図3にあるように，推定開始位置から文字列を遡って，活用語尾表の終止形の項にある「たい（助動詞）」にマッチした場合,「たい」の前側の文字列に対して活用形の推定を行う．助動詞「たい」は連用形接続であるので，要求する活用形を連用形，推定開始位置を文字「き」として活用チェック法を再帰的に呼び出す.「行き」はカ行五段動詞の連用形と推定できるので，助動詞の接続条件を満たす．そのため，

文字列「たい」を助動詞の終止形と推定する．

上記のように処理するため，活用語尾ツリーには直前の活用語に要求する活用形で分類して助動詞を登録している．助動詞は，接続の観点から未然形接続，連用形接続，終止形接続，連体形接続，体言接続に大きく分類できる．また，接続の表現形の違いから，「れる，せる」，「られる，させる」，「う」，「よう」，「ぬ」，「まい」，「ます」，「た」の8グループを特別扱いし，合計13通りに助動詞を分類している．図1の例では，「たい」を「助動詞F」，「ない」を「助動詞C」のように活用語尾ツリーに登録する情報で区別できるようにしている．さらに，活用語尾テーブルに登録する活用形のインデックスに上記8グループを付け加えている．

4. 活用チェック法の応用

前節で構築した活用チェック法を接続助詞「が」および否定表現の抽出に適用する．これにより，活用形を推定する際に要求した設計方針を活用チェック法が満たしているか否かの評価を行う．

4.1 接続助詞「が」の抽出への適用

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つの用法を持っており，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ため，文章を作成する際に接続助詞「が」の使用を吟味したほうがよい[8]．

4.1.1 判定条件

従来法では，表1に示した5つの判定条件によって接続助詞「が」の抽出を行っている．これらの条件のうち判定条件2以外の4つは文字「が」の前にある文字列が活用語の終止形になりうるかを判定する条件である．そのため，この4つの条件は，活用チェック法による活用語の終止形の推定に置き換えることができる．その結果，接続助詞「が」を抽出する際の条件は**表3**に置きかわる．

4.1.2 抽出精度の比較

活用チェック法による接続助詞「が」の抽出法を使用して，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： 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．1988年版，1～6月から抜粋（総文字数1,981,950文字）．

調査結果を**表4**に示す．67万字文章における調査結果によると，活用チェック法を利用した抽出法で抽出した場合，従来法で抽出した場合よりも第二種の誤りが30個減っている．また，再現率は100%のままである．また，従来法でふるい落とした接続助詞でない「が」は，活用チェック法を使用した抽出法でもすべて候補からふるい落としている．抽出の結果は，適合率は91.3%から98.0%になり，約7%向上している．従来法で抽出した候補の中には，ワ行五段動詞の転生名詞に格助詞「が」が接続したもの（たとえば「違いが」）が17個含まれていた．この第二種の誤りを活用チェック法ですべて取り除くことができる．このことが適合率向上の大きな要因としてあげられる．

しかし，**表5**に示すものについては，取り除くこ

表3 活用チェック法を使用した接続助詞「が」抽出法
Table 3 The new method to extract the conjunctive particle "GA".

判定条件1	活用チェック法により「が」の前の文字列に対して品詞の推定を行い，活用語の終止形でなければ，その「が」は接続助詞でない．
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．

表4 接続助詞「が」の抽出結果
Table 4 A result of extracting the conjunctive particle "GA".

調査対象文章：67万字文章

項　目	数	適合率	再現率
「が」の総数	6,987	—	—
従来法で抽出する候補	437	91.3%	100%
活用チェック法で抽出する候補	407	98.0%	100%
接続助詞「が」	399	—	—

調査対象文章：200万字文章

項　目	数	適合率	再現率
「が」の総数	30,570	—	—
従来法で抽出する候補	3,799	85.2%	100%
活用チェック法で抽出する候補	3,377	95.8%	100%
接続助詞「が」	3,236	—	—

表5 接続助詞「が」抽出における第二種の誤り（67万字文章）
Table 5 Errors of the second kind appearing in the evaluation of the new method to extract the conjunctive particle "GA" (about 670 thousand Japanese characters).

誤り	個数
いいがたい	3
ふるいがある	2
避けたがる	1
むだが多い	1
おばさんが	1
合計	8

とができなかった.「いいがたい」と「ふるいがある」はどちらも形容詞に接続助詞「が」が接続したものと誤って抽出している.「が」の直前の文字「い」を形容詞の活用語尾，その前の文字列を語幹としている．語幹については，最後の文字しか調べないので,「い」「る」をそれぞれ平仮名表で引き，語幹の推定を行っている．しかし，形容詞は平仮名表のほとんどの文字に語幹の品詞の候補として登録してあるので,「が」の前側の文字列を形容詞の終止形と推定する．このように，語幹が平仮名で終わる活用語の品詞の推定は，語幹が漢字で終わるものよりあいまいになる．表5のうち「避けたがる」以外の誤りは，語幹が平仮名で終わる活用語の品詞の推定を誤ったことが原因である．

調査対象文章を200万字文章とした場合でも，表4に示すように，従来法と比べて適合率が10%向上している．また，再現率は100%を保っている．

【推敲】で実用規模の文章としている1万字の科学技術文献において，接続助詞「が」の抽出を行った場合に得られる候補の数を表4から求めると，6.1個となる．科学技術文献における接続助詞「が」の抽出の適合率は98%であるので，この値から考えると接続助詞「が」の候補6.1個のうち第二種の誤りは0.12個となる．1個に満たない数である．同様に，1万字の新聞記事における接続助詞「が」の候補の数を求めると16.8個となり，第二種の誤りは0.71個となる．この場合でも，1個に満たない数である．

4.2 否定表現の抽出への適用

日本語の文章中に出現する否定は推敲の対象となる[5]．たとえば，二重否定を使うことで文章の意味が分かりにくくなったり，まわりくどくなったりすることがある．さらに，別の表現として,「ように」＋否定表現の形のものがある．この表現もあいまいな表現となりがちである．科学技術文章では，明確な表現をする必要がある[9]．そのため，二重否定のようなまわりくどい文や,「ように」＋否定表現のようなあいまいさが残る文はできるだけ控えた方がよい．このような理由から，文章中から二重否定や「ように」＋否定表現を指摘する機能を**【推敲】**に設けている．

上で述べた表現を抽出するには，まず，否定表現を抽出することが必要である．この否定表現の抽出に活用チェック法を応用する．

4.2.1 判定条件

否定を表す単語には「ない」「ぬ（ん）」「まい」の3つがある．これらのうち,「ない」は形容詞の場合と助動詞の場合とがある．形容詞の「ない」は活用語以外の語の後ろに続くので,「ない」の抽出に活用チェック法を適用することはしない．また,「ない」の抽出においては従来法による抽出でも十分な抽出精度を得ている[5].「ない」以外において，従来法で抽出精度が良くないものとして，助動詞「ぬ（ん）」の連用形「ず」と終止形「ん」がある．これらの抽出に活用チェック法を適用してみる．

従来法による否定の助動詞「ず」の抽出では，文字「ず」の前側の文字列に関する判定条件が2つ，後側の文字列に関する判定条件が2つある．これらのうち，前側の判定条件を活用チェック法に置き換える．否定の助動詞「ず」は活用語の未然形に接続するので，**表6**の判定条件1を設ける．また，従来法にある文字「ず」の後側の文字列に関する判定条件はそのまま使用する．そのため，活用チェック法を使用した否定の助動詞「ず」の抽出法は表6のようになる．また，否定の助動詞「ん」についても,「ず」と同じ判定条件が適用できる．

4.2.2 抽出精度の比較

従来法と活用チェック法とで否定の助動詞「ず」の抽出を行った結果を**表7**に示す．調査対象文章は200万字文章である．従来法では取り除くことができなかった第二種の誤り108個を活用チェック法を使用することにより取り除けるようになった．また，従来法で取り除いていたものは，活用チェック法を使用しても取り除くことができている．このため，適合率が71.2%から78.3%になり，約7%向上している．また，第一種の誤りは犯しておらず，再現率は100%に保っている．

表6 活用チェック法を使用した否定の助動詞「ず」の抽出法
Table 6 The new method to extract the negative auxiliary verb "ZU".

判定条件1	活用チェック法により「ず」の前側の文字列に対して品詞の推定を行い，活用語の未然形でなければ，その「ず」は否定の候補ではない．
判定条件2	「ず」の1文字後が促音・撥音である場合，その「ず」は否定の候補ではない．
判定条件3	「ず」の1文字後が「れ」である場合,「れ」の1文字後が「い，き，こ，つ，っ，て，ん」のいずれでもなければ，その「ず」は否定の候補ではない．

表7 否定の助動詞「ず」の調査結果（200万字文章）
Table 7 A result of extracting the negative auxiliary verb "ZU" (about two million Japanese characters).

項　目	数	適合率	再現率
「ず」の総数	2,161	—	—
従来法で抽出する候補	1,188	71.2%	100%
活用チェック法で抽出する候補	1,080	78.3%	100%
否定の助動詞「ず」	846	—	—

表8 否定の助動詞「ん」の調査結果（200万字文章）
Table 8 A result of extracting the negative auxiliary verb "N" (about two million Japanese characters).

項目	数	適合率	再現率
「ん」の総数	6,451	—	—
従来法で抽出する候補	3,245	8.6%	100%
活用チェック法で抽出する候補	1,242	22.6%	100%
否定の助動詞「ん」	281	—	—

表9 否定表現の抽出結果（200万字文章）
Table 9 A result of extracting the negative expressions (about two million Japanese characters).

項目	数	適合率	再現率
従来法で抽出する候補	12,969	73.1%	100%
活用チェック法で抽出する候補	10,858	87.3%	100%
否定表現	9,474	—	—

従来法と活用チェック法とで否定の助動詞「ん」の抽出を行った結果を**表8**に示す．従来法では敬称の「～さん」1,523個を第二種の誤りに含んでいたが，活用チェック法によりこのうち1,410個を取り除けるようになった．しかし，依然として「たくさん」「きちんと」「みんな」など取り除くことができない第二種の誤りが残っており，適合率は22.6%と低い．しかし，再現率は100%に保っている．

否定表現全体での抽出に関して，従来法と活用チェック法とで比較したものが**表9**である．この表を見ても分かるように再現率を100%に保ったまま，適合率を約14%向上させている．**『推敲』**で実用規模の文章としている1万字の文章では，否定表現の候補の数が約54.8個になる．否定表現の抽出における適合率が87.3%であることから，54.8個のうち約7.0個が第二種の誤りとなる．

一方，形態素解析を援用した字面解析手法を構築し，否定表現の抽出を行ってみた[10]．その抽出法を新聞記事データに適用して否定表現を抽出した結果，適合率が91.0%で再現率が99.3%であった．この結果と活用チェック法を用いた抽出法での抽出結果とを比べると，適合率が約3.7%向上するにとどまっている．また，形態素解析を援用した抽出法では，再現率が100%でないため，第一種の誤りを犯していることが分かる．このことから，活用チェック法を用いた否定表現の抽出法は，再現率100%を考慮すればかなりの抽出精度を上げていると見なすことができる．

文章を推敲する上で注意すべきものは，否定表現そのものというよりは，二重否定や「ように」＋否定表現のような表現である．**『推敲』**では否定表現の候補を使用してそれらの候補を抽出する．その際，第一種の誤りを犯していないので，二重否定や「ように」＋否定表現の候補の抽出においても指摘洩れはない．実際に書き手が吟味する項目は，この抽出法で抽出する否定表現の候補より少なくなる．誤りの数は単純に少なくなるとはいえないが，項目数が少なくなれば，第二種の誤りも少なくなる傾向にあると考えられる．現に文献2)（文字数17,359文字）で二重否定の候補と「ように」＋否定表現の候補の数を調べると9個と6個であった．この程度の数に指摘を絞り込むことができるので，活用チェック法を適用した否定表現の抽出は，**『推敲』**に適用する上で十分な抽出精度といえる．

5. 応答時間

『推敲』を開発する際の方針の1つとして「実用規模の文章を待ち遠しくない時間で処理してほしい」をあげている．ここでは，活用チェック法による接続助詞「が」の抽出法がこの方針を満たしているかを評価する．

『推敲』は現在パソコン上に実現している．ユーザがキーボードからコマンドを発すると，**『推敲』**はすべての候補を検索し，結果の先頭部分を画面に表示する．画面からあふれた結果に対して，ユーザは画面をスクロールさせて候補を1つ1つ吟味していく．このように，**『推敲』**を使用する際には，コマンドを発してから先頭部分の画面が表示されるまでの時間（応答時間）が問題となる．

応答時間の測定にあたって，従来法による接続助詞「が」の抽出法と活用チェック法による抽出法とをパソコン版**『推敲』**Ver.1.9のコマンド“接続助詞「が」の抽出”に実装した．このコマンドを実行してから，結果が画面に出力されるまでの時間を測定した．測定にはPC-9801 RA（CPU 80386 CLOCK 16 MHz）を使用し，調査対象文章は，研究室で蓄えている科学技術文献である．測定結果を**図4**に示す．

応答時間は1万字の文章では0.45秒，3万字の文章でも0.6秒程度で，従来法も活用チェック法もほとんど変わらない．図4から誤差最小二乗法により，処理文字数が0字の時の応答時間を算出すると，0.38秒であった．この値から考えると，応答時間の大半は抽出にかかる時間ではなく，抽出結果を画面へ出力するために費やされる時間である．「が」の抽出にかかる時間は非常に短いため，抽出時間のみの比較は困難である．しかし，1万字の文章を抽出対象文章として，活用チェック法で接続助詞「が」の抽出を行っても，応答時間は0.45秒であり，これは**『推敲』**の開発方針である「実用規模（1万字程度）の文章を待ち遠しく

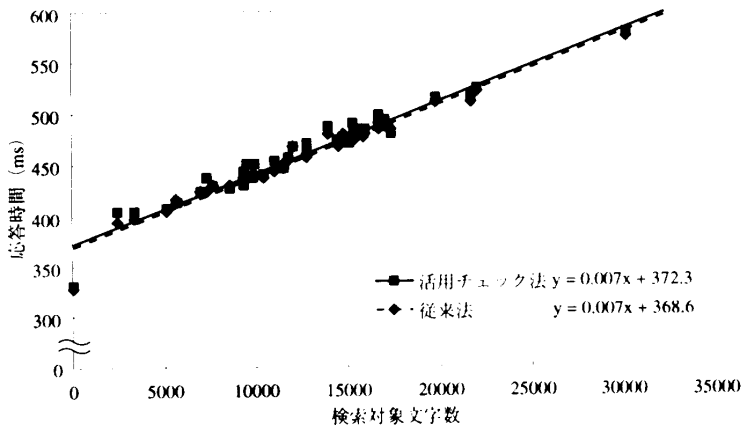

図4 接続助詞「が」抽出の応答時間の比較
Fig. 4 The response time of the new method to extract the conjunctive particle "GA".

ない時間で処理してほしい」を十分満たしているといえる.

6. 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推定する方法を構築し，それを使用して接続助詞「が」および否定表現の抽出を行った．その結果，接続助詞「が」の抽出において，従来の判定条件だけによる抽出に比べて約10%の適合率の向上が見られた．一方，再現率は100%に保ったままであった．また，否定表現の抽出においても，従来法では精度があまり向上しなかった否定の助動詞のうち「ず（連用形）」「ん（終止形）」の抽出において，活用チェック法で精度を向上できることが分かった．

この活用チェック法で使用する活用語尾表・漢字表・平仮名表の大きさは合わせて32 KB程度である．これらの表を主記憶に載せたとしても，記憶域を圧迫するものではない．そのため，パソコン上でも余裕を持って実現できる．

また，活用チェック法を使用した抽出法をパソコン上の日本語文章推敲支援ツール『推敲』に組み込んで応答時間の測定を行った．その結果，従来法で抽出する場合と同程度の処理速度で抽出できることが分かった．

活用チェック法は処理の開始点と要求する活用形を決めてから推定を開始する．今後の課題は，文の最後の文字から活用チェック法を適用し，文を逆向きに解析していく方法を構築することである．その方法によって，平仮名部分の誤り（変換忘れ，文字の挿入・削除ミス，助動詞の接続誤りなど）を抽出し，書き手に指摘する方法の構築を行う．

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．また，応答時間の測定には製品科学研究所の森川浩氏から提供していただいた打鍵データ収集システムを使用した．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, pp.35–46 (1986).
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, pp.55–67 (1989).
3) 牛島和夫，石田真美，尹 志熙，高木利久：日本語文章推敲支援ツールにおける受身形の抽出法，情報処理学会論文誌，Vol.28, No.8, pp.894–897 (1987).
4) 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，情報処理学会自然言語処理研究会，No.68, pp.1–8 (1988).
5) 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).
6) 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「が」の抽出について，情報処理学会論文誌，Vol.35, No.8, pp.1652–1660 (1994).
7) 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).
8) 清水幾太郎：論文の書き方，岩波新書 (1959).

9) 木下是雄：理科系の作文技術，中公新書 (1981).
10) 畑中隆洋，菅沼　明，牛島和夫：形態素解析を援用した字面解析による否定表現の抽出と評価，情報処理学会自然言語処理研究会，No.100, pp.113–120 (1994).

(平成 7 年 10 月 19 日受付)
(平成 8 年 4 月 12 日採録)

菅沼　明（正会員）

1961 年生．1986 年九州大学工学部情報工学科卒業．1988 年同大学院工学研究科情報工学専攻修士課程修了．1991 年同博士後期課程修了．同年九州大学工学部情報工学科助手勤務．1993 年同大学工学部情報工学科講師，1996 年同大学院システム情報科学研究科助教授，現在に至る．工学博士．日本語情報処理，ユーザインタフェース，ニューラルネットワークの応用などに興味を持つ．1994 年情報処理学会奨励賞受賞．日本ソフトウェア科学会会員．

笠原　晋（正会員）

1971 年生．1994 年九州大学工学部情報工学科卒業．1996 年同大学院工学研究科情報工学専攻修士課程修了．同年九州松下電器に入社．日本語情報処理に興味を持つ．

牛島　和夫（正会員）

1937 年生．1961 年東京大学工学部応用物理学科（数理工学コース）卒業．1963 年同大学院修士課程修了．同年九州大学中央計数施設勤務．1977 年九州大学工学部情報工学科教授（計算機ソフトウェア講座担当），1996 年同大学院システム情報科学研究科長，現在に至る．1990 年 4 月から 1994 年 3 月まで九州大学大型計算機センター長を兼務．1991 年度情報処理学会九州支部長．1995 年 5 月から本学会監事．工学博士．電子情報通信学会，ソフトウェア科学会，ACM 各会員．