日本語文章推敲支援ツール『推敲』における字面解析 — 助詞「は」に着目して —

3J-5

菅沼 明　松尾 朗*　牛島和夫
(九州大学工学部)

1. はじめに

日本語文章推敲支援ツール『推敲』は、機械可読な形で存在する日本語文章を解析して推敲に役立つ情報を提供することを目的としたツールである[1,2]。このツールの開発に当たって次の2つの方針を設けた。

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欲しい。

日本語の文章は分かち書きされていない。そのため、日本語の文章を解析するには、まず辞書に基づいて形態素解析を行い、構文解析を行うのが普通である。しかし、辞書を使った文法処理を行うと、実用規模の文章では解析時間がかかりすぎて上記②の方針を満たさない。このため、我々は辞書を使わず文法処理も行わず字面だけで解析する方法を採ることにした。このために、『推敲』が指摘するものの中には誤りが含まれる。開発方針①から、『推敲』が指摘したものは書き手が一度は目を通すので、第一種の誤り「指摘に漏れがある」は犯してはならないけれど、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」はある程度許容できる。字面だけの解析で第一種の誤りを犯さずに開発方針を満たす程度の精度が得られれば、『推敲』で採用する解析手法としては十分である。

すでに、我々は指示詞(「これ、それ、あれ」など)、受身などの助動詞「れる、られる」[3]、接続助詞「が」、否定表現の抽出[4]を上記の条件を満たす形で実現することに成功した。指示詞の抽出は単に文字列照合のみであるけれども、受身の抽出、接続助詞「が」、否定表現の抽出のアルゴリズムは機械可読な日本語文章約100万字を実際に調査して構築した。ここでは、推敲作業を支援する情報の一つとして助詞「は」に注目することにした。

2. 助詞(副助詞)「は」を指摘する意義[5]

副助詞「は」の主な用法には、題目と対照(限定)とがある。例えば次の文「象は鼻は長い」では鼻を強調した文になっているが、文のニュアンスによって、「は」(副助詞)ではなく「が」(格助詞)を用いて「象は鼻が長い」と、書き換えたほうがよくなる場合もある。

日本語文では、1文中に助詞の「は」、もしくは「は」の音(wa)が複数個含まれることが多くある。こうのような文は読み手にとって読みにくいし、また聞きづらいことがある。特に、文節の終わりに(wa)と発音する「は」が多く出現する次のような文、

「あるいは、この学校では、私は、〜」

は、読みにくい。また、対照の「は」は言葉の上ではいくらでも重ねることができるが、数が多くなると潜在する情報が膨大な量となり、わかりにくい文になり易い。

3. 助詞「は」

3. 1 助詞「は」の接続

助詞「は」の接続は、学校文法によれば、

(1)いろいろな語に接続する。
(2)用言(動詞、形容詞、形容動詞)には、連用形に接続する。

とある。受身や接続助詞「が」などこれまでに構築してきた抽出法では、検索対象文字列の1文字前に学校文法から設けた条件をつけるだけで、受身では約半分に、接続助詞「が」では約10分の1以下に候補の数を絞ることができた。しかし(1)から、助詞「は」の場合に、学校文法から1文字前の条件を設けることはできない。

3. 2 文字「は」の調査

前節で述べたように、助詞「は」の抽出には、接続の性質を用いた判定条件を設けることはできない。そこで、実際の日本語文章(総文字数683,867)中から文字「は」を検索し、その「は」が実際に助詞として使われているか否か、「は」がどんな文字に接続しているかを調査した。その結果を表1、2に示す。表に示したように、文字「は」は出現頻度が高く、また出現したもののうち95.3%が助詞であった。抽出の精度(指摘した候補の数に対する指摘すべきものの割合)は高いにこしたことはないが、悪くても95%以上くらいあれば十分だとしているので、文字列照合のみによる抽出でもよさそうである。しかし、抽出精度としては満足いく値が得られても指摘する絶対数が多いので、第二種の誤りを除去するいくつかの判定条件を設ける。

表1 68万字の文献中の「は」の数

「は」の1文字前	総数	助詞	割 合(%)
記号、英数字など	1,349	1,233	91.4
カタカナ	1,522	1,522	100.0
漢 字	3,110	3,051	98.1
ひらがな	4,536	4,219	93.0
合 計	10,517	10,025	95.3

※ 割合は総数に対する助詞の数の割合

表2 助詞でない「は」

項 目	数	割 合(%)
あるいは、或いは	158	32.1
または、又は	120	24.4
はじめに、はじめる はじめて、はじめは はじまる、はじまり	118	24.0
はっきり	19	3.9
「を」+「は〜」	13*	2.6
そ の 他	71	14.4
合 計	492	100.0

* 「を」+「は〜」は、13個あったがそのうちの7個は「はじめる」、「はっきり」と重複している。

3. 3 「は」で終わる接続詞

「は」で終わる接続詞「あるいは(或いは)」、「または(又は)」、「もしくは(若しくは)」の文字「は」

*現在 ㈱東芝

Textual Analysis Method in the Writing Tools for Japanese Documents - On extracting the particle "は" -
Akira SUGANUMA, Akira MATSUO and Kazuo USHIJIMA
Kyushu University

は(wa)と発音する。2節で述べたように(wa)と発音する「は」を重ねた場合も、助詞「は」を重ねた場合と同様に文が読みにくくなることがある。したがってここでは、助詞の「は」だけに限定しないで、「は」で終わる接続詞なども含めて、抽出の対象とする。

3．4　助詞でない「は」の除去

◎はじめ*、はじま*　　検索した「は」が助詞の「は」であれば、「は」に続く単語は自立語である。したがって、「じめ」、「じま」で始まる自立語を公用データベース日本語単語辞書[6]で調べた。その結果を表3、4に示す。このことから、以下の2つの判定条件を設けることができる。

判定条件1：「は」の後に「じめ」とつづいた場合、「め」の1文字後が「い、じ、つ、ん」のいずれかであれば、その「は」は助詞である。

判定条件2：「は」の後に「じま」とつづいた場合、「ま」の1文字後が「い、え、く、ま、わ、ん」のいずれかであれば、その「は」は助詞である。

表3　「じめ」で始まる語

じめい(自明)、じめじめ、じめつ(自滅)
じめん(字面)、じめん(地面)

表4　「じま」で始まる語

じまい(地米)、じまえ(自前)、じまく(字幕)
じまま、じまわり(地回り)、じまん(自慢)

◎はっきり　　上で述べたように、「は」を助詞と仮定すると「は」の次に来る文字は自立語の最初の文字である。ところが、促音で始まる自立語はないことから、「は」の1文字後が促音である場合、この「は」は助詞でないといえる。また同様な理由から、「は」の1文字後が撥音である場合も、その「は」は助詞ではない。そのため以下の条件を判定条件とする。

判定条件3：「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。

◎「を」＋（「は」で始まる語）　　格助詞の「を」は目的格を表す。この目的格の単語を強調するには「を」を「は」に変えるだけでよい。そのために格助詞の「を」に副助詞の「は」が続くことはない。また「を」で終わる単語もない。このため、以下の条件を判定条件に加える。

判定条件4：「は」の1文字前が「を」である場合、その「は」は助詞でない。

以上の1～4の判定条件によって、209個の第二種の誤りのうち、143個を除去できる。

3．5　抽出アルゴリズムの精度

前節で述べた4つの判定条件を用いると、(wa)と発音する「は」の候補として抽出する数が10,091個で、そのうち正しい候補（(wa)と発音する「は」）が10,025個で、その精度は約99.3%となる。これは、実用規模（約1万字）の文章で考えると、平均148.4個の指摘があり、そのうち誤って指摘してしまうものが、約0.97個しか含まれないことになる。

推敲作業を支援する情報としてわれわれが抽出したいものは、助詞「は」を重複して含んでいる文である。そこで、文字「は」を1文中に複数個含む文の数を調査した。その結果、68万字の文章で文の総数が21,447、そのうち「は」を複数含む文が1,829であった。これだけの作業で推敲の対象となる文の数をかなり減らすことができる。このことから、上で述べた(wa)と発音する「は」の候補の抽出法は、『推敲』の開発方針である「実際に推敲するのは書き手である」からすると、十分実用的であるとみなすことができる。

4．抽出アルゴリズムの評価

3．4節で述べたアルゴリズムは約68万字の日本語文章を調査して構築したものである。このアルゴリズムが他の日本語文章にも有効であることを確認するために、別の文章（総文字数 2,842,062文字）で評価を行った。

評価の方法は構築した判定条件にしたがって「は」を抽出し、(wa)と発音する「は」であるか否かを調査した。また、各判定条件によって除去される「は」について、第一種の誤りを犯していないことを確かめた。

上記の判定条件を満たす第二種の誤りを表5に示す。評価の結果、約280万字の文章中に「は」が33,783個含まれており、そのうち33,011個の「は」が4つの判定条件を満たす。目視によって調べた結果、助詞の「は」は31,550個であり、「は」で終わる接続詞は939個出現した。このことから、抽出の目的である(wa)と発音する「は」の数は32,489個となり、抽出の精度は98.4%となる。実用規模（約1万字）の文章で考えると、平均116.2個の指摘があり、その中に約1.8個の誤りを含むことになる。なお、「はあく、は握」がすべて「把握」と書かれていたと仮定すると、抽出精度は99.1%となる。

表5　指摘の誤り

項目	数	割合(%)
はあく、は握	243	46.6
名詞	59	11.3
あてはまる	31	5.9
はるかに	25	4.8
その他	164	31.4
合計	522	100.0

5．おわりに

文章中に存在する(wa)と発音する「は」を字面解析だけで抽出する方法を構築した。抽出の精度は約98%と悪くはないが、問題となるのは、指摘の数そのものが多いということである。そのため、「は」の抽出を『推敲』に組み込む場合には、それを基本コマンド[2]として実現し、通常のユーザに対しては他のコマンドと組み合わせた使いやすいマクロコマンドを考えなければならない。

参考文献

[1]牛島和夫他：日本語文章推敲支援ツール『推敲』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]倉田昌典他：日本語文章推敲支援ツール『推敲』のパソコン上での実現と使用，情報処理学会第29回プログラミングシンポジウム報告集，(1988), pp.45-54

[3]牛島和夫他：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, 1987, pp.894-897

[4] 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, (1988), 68-8

[5] 本多勝一：日本語の作文技術，朝日新聞社，(1976)

[6] 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, (1983), pp.335-361