6B-1 日本語文章推敲支援ツール『推敲』における字面解析手法の評価

菅沼　明　　石田朗子*　　倉田昌典　　牛島和夫

(九州大学　工学部)

1. はじめに

日本語ワードプロセッサの普及は著しい。これに伴って、機械可読の日本語文章が増えつつある。しかし、ワードプロセッサは、文章の入力、書式の設定、文章の出力、保存を行うだけで、文章の推敲を支援する機能を積極的に備えているわけではない。入力した文章は機械可読の形であるにも拘らず、清書出力にしか使わないのは残念である。機械可読の文章から推敲に役立つ情報を計算機で抽出しようということで、我々の研究室では日本語文章推敲支援ツール『推敲』[1,2]の開発を行っている。

『推敲』の開発に当たって、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（推敲するのは書き手である）；　②実用規模（１万字程度）の文章を待ち遠くない時間で処理して欲しい；　という要求を課した。要求②を満たすために、『推敲』では辞書も使わず、文法解析もせずに、字面だけで文章を解析している。このために、『推敲』が指摘するものの中には誤りが含まれている。開発の要求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。

我々は、上記の要求を満たす形で、推敲に役立つ情報を抽出する字面解析手法の構築を行ってきた[3,4]。それらは、機械可読の日本語文章を実際に調査した結果を参考にして構築したものである。調査に使用した文章は我々の研究室で蓄えているもの（約100万字）であり、限定されたものといえる。そのために、これらの字面解析手法が一般の文章に対しても有効であることを確認するために、調査した文章とは別の文章で評価を行った。

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）、単語「より」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。そして、それらを抽出するために、個々に字面解析アルゴリズムを用意してきた。それらの中には文字列の照合だけしか行わないものもある。例えば、指示詞、とりたて詞、単語「より」の抽出は文字列照合だけで行っている。しかし、受身、接続助詞「が」、否定表現の抽出では、文字列照合をした後に、複数の判定条件の検査を行って第二種の誤りを少なくしている。例えば、受身と接続助詞「が」との抽出法は次のとおりである。

・受身の候補の抽出[3]

受身の助動詞「れる、られる」の候補を抽出する。文章中から「れ」を探しだし、以下の条件の検査を行う。

判定条件１：「れ」の１文字前は「か、さ、た、な、ま、ら、わ、が、ば」のいずれかである。

判定条件２：「れ」の１文字前が「な」の場合、「しなれ」、「死なれ」以外の「れ」は受身ではない。

判定条件３：ラ行下一段活用動詞のうち判定条件１を満たすものを適当に選び一覧表にして、これに一致したものを除去する。

判定条件４：「われわれ」は受身ではない。

・接続助詞「が」の候補の抽出[4]

まず、文字列照合で文章中の「が」を探しだし、「が」の前後の文字で、その文字が接続助詞であるかを判定する。

判定条件１：「が」の１文字前は「う、く、す、つ、ぬ、む、る、ぐ、ぶ、い、だ、た、ん」のいずれかである。

判定条件２：「が」の１文字後は促音「っ」、撥音「ん」ではない。

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」は接続助詞ではない。

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」は接続助詞ではない。

これらの判定条件を設けた根拠や、その他の表現の抽出法については紙数の関係で省略する。詳しくは参考文献[3,4]を参照されたい。

3. 評価対象データと評価方法

評価には、JICST科学技術文献ファイルの管理システム編（文献数14,380、総文字数2,842,062文字）の表題と抄録の部分を使用した。評価方法は、各字面解析手法に沿って候補を計算機で抽出し、その候補が正しいか否かを目視で確認する。さらに、各判定条件で取り除いたデータに対して、第一種の誤りを犯していないかを調べる。

4. 結果

それぞれの推敲情報（否定表現を除く）について、抽出したデータの調査結果を表1 に示す。この表から分かるように、精度はすべて90%を越えている。

---

*現在　三菱電機㈱

Evaluation of Textual Analysis Method in the Writing Tools for Japanese Documents

Akira Suganuma, Akiko Ishida, Masanori Kurata and Kazuo Ushijima

Kyushu University

表1 各推敲情報の調査結果

推敲情報	候補	正	精度（%）
指示詞	18,794	18,662	99.3
単語「より」	3,218	3,189	99.1
とりたて詞	8,937	8,831	98.8
受身	12,140	11,953	98.5
接続助詞「が」	3,419	3,213	94.0

$$精度 = \frac{正しい抽出データ数}{抽出データ数（候補の数）} \times 100\ [\%]$$

指示詞として11種類の文字列を取りあげている。その1つ1つが上記の精度で抽出されるわけではない。例えば、指示詞「あれ」の抽出精度は5.5%と非常に低い。しかし、文字列「あれ」の出現頻度は残りの指示詞の出現頻度に比べて極めて少ない（指示詞全体の0.4%）。そのために、指示詞全体の抽出精度にはほとんど影響していない。このような抽出文字列は「あれ」だけではなく、とりたて詞、否定表現などにも含まれている。

否定表現の調査結果を表2 に示す。否定表現の抽出精度は他に比べると低い。これは、否定の助動詞「ぬ」の終止形または連体形の「ん」を抽出しようとするからである。否定の助動詞「ん」は「ぬ」の発音上の言い替えであり、ほとんど話し言葉でしか使わない。そのため、文章中で否定の「ん」は「分かりません」のような会話文中で使用するのが主で、『推敲』で主な対象と考えている科学技術文章中にはほとんど出現しない。このことを考えて、否定表現を抽出するアルゴリズムを『推敲』に組み込む際には、すべての「ん」を候補から外している。

表2 否定表現の調査結果

推敲情報	候補	正	精度（%）
「ん」を含む	9,665	6,597	68.3
「ん」を除く	7,313	6,597	90.2

次に、『推敲』で基準としている1万字当りの誤りの数を求めた。その結果を表3 に示す。この数が大きければ、『推敲』を使用するのに苦痛を感じるであろう。

表3 1万字当りの数

推敲情報	候補の数	誤りの数
否定表現	34.0	2.5
接続助詞「が」	12.3	0.7
受身	42.7	0.6
指示詞	66.1	0.5
とりたて詞	31.4	0.3
単語「より」	11.3	0.1

否定表現を除いて、1万字当りの誤りの数は1未満である。また、誤りの数が多い否定表現（「ん」を除く）でも、2.5個しかない。『推敲』の立場（実際に推敲するのは書き手である）からすると、我々が構築してきた字面解析手法は、『推敲』に使用する抽出法として十分実用的であるとみなすことができる。

さらに、判定条件で除去したデータを目視で調査して、第一種の誤りを犯していないことを確かめた。『推敲』で否定の候補から外している否定の助動詞の終止形または連体形の「ん」は、今回調査した文章中には出現しなかった。

5. 考察

JICST科学技術文献ファイルを使用して、我々が構築してきた字面解析手法に従ってデータの抽出を行った。その結果、構築の際に調査した文章と同様の結果を得た。また、抽出の際に第一種の誤りを犯していないことも確認できた。ただ、否定表現では、否定の助動詞の終止形または連体形の「ん」を候補から外しているために、第一種の誤りを犯す可能性はある。このことは、『推敲』のヘルプ機能などでユーザに知らせる必要がある。

我々は、パソコン上で実現した『推敲』[5]に、これらの字面解析手法を既に組み込んでいる。それぞれの推敲情報を抽出するのに要する時間は、1万字の文章でおよそ1秒以下である[6]。この処理時間は『推敲』に課した要求の②を十分満たしている。

否定の「ぬ、まい」指示詞の「あれ」は、精度が50%程度やそれ未満であった。この調査結果から、これら精度が悪い抽出法の改善を行うことが今後の課題である。また、今回の調査に使用した文章も科学技術文章の抄録であり、日本語文章全体からみれば偏ったものである。そのために、さらに広範な文章で評価を進めていきたい。

分かち書きをしていない日本語文章を解析するためには、形態素解析を行い、構文解析をするのが普通である。最近、各社が製品化し発表した文書校正支援システムではほとんどが文法解析を行っている[7]。しかし、文法解析には時間がかかるし、解析結果には曖昧さが残ることもあるだろう。我々が構築してきた字面解析手法の反応時間と精度とを考えると、字面解析手法を使っても十分実用に耐えうるツールを作ることができる。

謝辞

本研究を進めるにあたり、JICSTデータの使用について便宜をはかっていただいた姫路短大の田中康仁先生と、御討論を頂いた牛島研究室の皆様に感謝いたします。

参考文献

[1] 牛島和夫他：日本語文章推敲支援ツール『推敲』の使用について，九州大学大型計算機センター広報 Vol.18, No.1, 1985

[2] 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

[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

[4] 菅沼明他：日本語文章推敲支援ツールにおける推敲情報抽出アルゴリズムの構築，日本ソフトウェア科学会第4回大会，1987

[5] 倉田昌典他：日本語文章推敲支援ツール『推敲』のパーソナルコンピュータでの実用化，情報処理学会第35回全国大会，1987

[6] 倉田昌典他：日本語文章推敲支援ツール『推敲』における応答時間，情報処理学会第37回全国大会，1988

[7] 大用昌之：次世代ワープロの決め手となるか校正支援／可読性評価ツール，日経バイト，No.43, pp.96-104, (1988.3)