1W－9

日本語文章推敲支援ツール『推敲』における連用中止法の抽出について

縄田 晶裕　菅沼 明　牛島 和夫
九州大学工学部情報工学科

1. はじめに

日本語文章推敲支援ツール『推敲』[1] は日本語文章を字面だけで解析して、推敲に役立つ情報を書き手に提供することを目的として開発しているツールである。本論文では、科学技術文献を調査して、字面解析による連用中止法抽出のための判定条件を構築する。

2. 連用中止法とは

連用中止法とは、「学校へ行き、先生に会う。」のように連用形で文を区切って、次につなげていく用法である。『論文の書き方』[2] によると、連用中止法を文章中に用いると次のような危険を生じる。

i) 中止法をはさむ前件と後件とで、主語がすりかわることがある。
ii) 長文になって、文の照応がねじれる。
iii) 一文に、多くの項目をつめこんで、文意がとおりにくくなる。

よって、連用中止法を抽出することは文章を推敲する上で意義あることといえる。

3. 字面解析による抽出

3.1 連用中止法の抽出

連用形活用語尾の平仮名の候補は25種類「き、し、り、ち、み、い、ぎ、び、に、じ、ひ、え、け、げ、せ、ぜ、ね、へ、べ、め、れ、く、ず、て、で」である。連用形が促音で終わるもの（例：そうだっ）は連用中止法に用いられることはないので抽出の対象から外してよい。また連用中止法には、連用形活用語尾の後に読点またはコンマが来るものと読点もコンマも来ないものがある。後述の調査（表1）では前者は2,239個、後者は、115個出現した。しかし、後者は前者の出現数に対して少ないため、抽出の対象から外すことにする。よって、次の判定条件を設ける。

判定条件: 連用形活用語尾の次の文字は、読点（、）またはコンマ（，）である。

Extraction Method of "ren'you-tyuusi" in the Writing Tools for Japanese Documents.
Akihiro NAWATA, Akira SUGANUMA and Kazuo USHIJIMA
Department of Computer Science and Communication Engineering, Kyushu University

表 1: 連用形の抽出

項目	個数
総文字数	669,842
判定条件を満たす活用語尾	5,932
実際の活用語尾	2,091

機械可読の日本語文章（卒論、修論、翻訳文、レポート、その他）669,842文字について、この判定条件を満たすものを抽出したものが表1である。これより、連用形の抽出精度（候補中に含まれる連用形の数 ÷ 文章中の連用形の数）は35.2%と低い。『推敲』では第一種の誤り（指摘に洩れがある）は犯してはならないが、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、ある程度許容している。ここでは、第二種の誤りを減らして抽出精度を上げるため、更に各文字について、実際の連用形を抽出するための判定条件を構築する。

3.2 「き」の抽出

文字「き」の後に、読点またはコンマが来るものを抽出したものが、表2である。第二種の誤りとして一番多いのは、名詞「とき」である。連用形が「とき」で終る動詞は「解く、溶く、説く、釈く」があるので，単に文字列「とき」を外してしまうと第一種の誤りを犯してしまう。ここで公用データベース日本語単語辞書[3] によると、カ行五段動詞、カ行上一段動詞のうち、「き」の1文字前にくる平仮名は「あ、で、お、か、が、こ、ざ、す、ず、せ、ぞ、た、だ、つ、づ、と、ど、な、ぬ、ね、は、ば、ひ、び、ふ、ぶ、ま、め、や、ろ、わ 」である。「き」の1文字前にこれらの平仮名が来れば、その「き」を連用形活用語尾の候補とみなしてよい。次に第二種の誤りが多い「手続き」については、動詞「続き」の1文字前が手で終ることはない。これらのことか

表 2: 67万文字による「き」の抽出

種類	出現数	割合（%）
連用形活用語尾	75	56.8
名詞「とき」	44	33.3
名詞「手続き」	12	9.1
その他	1	0.8
計	132	100

ら、判定条件をまとめると次のようになる。

判定条件:

i) 「き」の1文字前の平仮名が上記のものである場合のみ連用形活用語尾とみなす。ただし、「き」の1文字前が「と」であるときは「と」の1文字前が「の」のときには抽出の候補から外す。

ii) 「き」の1文字前が漢字であればその「き」を連用形活用語尾の「き」とみなす。ただし、「き」の1文字前が「続」である場合その1文字前が「手」であるときは抽出の候補から外す。

この判定条件を用いると、連用形の「き」の抽出精度は、56.8% から 100% に上げることができる。

3.3 「し」の抽出

判定条件:

i) 「し」の1文字前が「も」「但」である場合、その1文字前が漢字、平仮名またはカタカナでなければ、その「し」は候補から外す。

ii) 「し」の1文字前が「か」「だ」である場合はその2文字前が漢字、平仮名またはカタカナでなければ、その「し」は候補から外す。

この判定条件を用いると「し」の抽出精度は 48.5% から 88.6% に上げることができる。

3.4 「め」の抽出

判定条件：

i) 「め」の一文字前が「た」であるとき、「た」の1文字前が「く、ぐ、す、つ、ぶ、む、る、ぬ、う、な、の」であれば、その「め」は連用形の「め」ではない。

この、判定条件を用いると連用形の「め」の抽出精度は 12.6% から 100% になる。

3.5 「ず」の抽出

我々の研究室では以前に否定表現の抽出のための判定条件を構築した[4]。それを用いると、「ず」の抽出精度は 91.8% になる。

3.6 「に」の抽出

判定条件：

i) 「に」の1文字前が「死、し」であればその「に」を連用形活用語尾の「に」とする。

ii) 読点またはコンマの前に漢字「似、煮」がくれば連用形とする。

3.7 「て」の抽出

表3のタ行下一段動詞を抽出すれば良い。

3.8 「で」の抽出

判定条件：

i) 「で」の1文字前が「の」であるもののうち「の」の1文字前が「も」でないものは、連用形ではない。

表 3: タ行下一段動詞

平仮名	そばだてる、ほだてる、しょってる、ばてる、もてる、
漢字	立てる、煽てる、育てる、果てる、企てる、隔てる、建てる、慌てる、持てる、捨てる、打てる、凍てる、当てる、

ii) 「で」の1文字前が「こ」であるものは、「こ」の1文字前が文頭にくれば、それは連用形活用語尾ではない。

この判定条件を用いると抽出精度は 6.1% から 14.0% になる。まだ第二種の誤りとして格助詞の「で」を含んでいるため抽出精度は低い。しかし、格助詞の「で」と断定の助動詞の連用形「で」はどちらとも名詞に接続するため、字面では判別が困難である。

3.9 その他のもの

「ぎ、ひ、ぜ、ね」については、今回の調査では、出現しなかったので、特に判定条件を設けることはしていない。また、それ以外のものには判定条件を設けなくても抽出精度が高いので、特に判定条件を付け加えない。更に連用形が漢字であるものもあるが、そのような漢字は 21 種類しかないのでその漢字を抽出してくる。

4. まとめ

これまでに構築した判定条件を用いると、判定条件を満たす候補の数は 2,538、そのうち実際の連用形は 2,091 で、抽出精度は 82.4% であった。すべての判定条件を用いても第一種の誤りを犯していないことを確認した。さらに別の文章（JICST 科学技術文献ファイルの抄録: 総文字数 709,492）を用いると、抽出精度は 84.1% となり、同等の抽出精度を得ることができた。また、読点もコンマも来ない連用中止法については、抽出の対象から外しているが、これは連用中止法抽出という立場から考えれば、第一種の誤りを犯しているといえる。今後の課題として、ここまで含めた抽出法の構築があげられる。

参考文献

[1] 倉田昌典他：日本語文書推敲支援ツール『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] 尾川正二：原稿の書き方，講談社現代新書，1976.

[3] 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.

[4] 菅沼明他：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, 792-800, 1990.