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

織田 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』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、科学技術文書を調査して、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」によると、連用中止法を文章中に用いることはあるが、その場合の文法上の問題として次の二つの問題がある。

i) 中止法をもつ前件と後件で、主語がずれかわること。

ii) 長文になって、文の組み合わせが変える。

これらの問題から、連用中止法抽出のための判定条件を構築することにした。

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、が、せ、そ、な、へ、べ、め、れ、く、す、て、で」である。連用形が連続で終わるもの（例：そうだ）は連用中止法に用いられないものな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接点またはコマンダが来るものと接点もコマンダも来ないものがある。後者の調査（表1）では前者は2,239個、後者は115個出現した。しかし、前者の出現数に対して、後者は前者の出現数に対して少ないため、抽出の対象から除外することにした。よって、次の判定条件を設ける。

判定条件：連用形活用语尾の次の文字は、読点（。）またはコマンダ（、）である。

機械可読の日本語文章（卒論、修士論、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35.7%と低い。「推敲」では第一種の誤り（指摘に溜めがある）が発生してしまい、第二種の誤り（指摘すべきでも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、点読またはコマンダ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」で終わる動詞は「解く、読く、読む、読む」であるので、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。そこで、連用データベース日本語辞書辞書[9]によると、行文五段動詞、行文上一段動詞のうち、「き」の1文字前にくる平仮名は「あ、で、お、か、が、こ、そ、す、せ、ざ、た、だ、つ、づ、と、と、な、ぬ、ね、は、び、ぴ、ふ、ま、め、や、ら、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「手続き」の1文字前に手がつくることはない。これらのことか

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>項目</td>
<td>個数</td>
<td></td>
</tr>
<tr>
<td>総文字数</td>
<td>669,842</td>
<td></td>
</tr>
<tr>
<td>判定条件を満たす活用語尾</td>
<td>5,932</td>
<td></td>
</tr>
<tr>
<td>実際の活用語尾</td>
<td>2,091</td>
<td></td>
</tr>
</tbody>
</table>

表2: 連用形活用語尾における「き」の抽出

<table>
<thead>
<tr>
<th>種類</th>
<th>出現数</th>
<th>調学 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>運用形活用語尾</td>
<td>75</td>
<td>56.8</td>
</tr>
<tr>
<td>名詞「とき」</td>
<td>44</td>
<td>33.3</td>
</tr>
<tr>
<td>名詞「手続き」</td>
<td>12</td>
<td>9.1</td>
</tr>
<tr>
<td>その他</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>計</td>
<td>132</td>
<td>100</td>
</tr>
</tbody>
</table>

Extraction Method of "ren'you-tyuusui" in the Writing Tools for Japanese Documents,
Akihiro NAWATA, Akira SUGANUMA and Kazuo USHIJIMA
Department of Computer Science and Communication Engineering, Kyushu University
ら、判定条件をまとめて次のようになる。

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をとします。ただし、「き」の1文字前の「と」であるときは「と」の1文字前の「の」のときには抽出の候補から外す。
ii) 「き」の1文字前に漢字であればその「き」を連用形活用語尾の「き」とみなす。ただし、「き」の1文字前に「続」である場合その1文字前に「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文字前に「も」「但」である場合、その1文字前に漢字、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文字前に「か」「だ」である場合はその2文字前に漢字、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1文字前に「た」であるとき、「た」の1文字前に「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[2]。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文字前に「死、し」であるすればその「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」がなければ連用形とする。

3.7 「て」の抽出

表3の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文字前に「の」であるもののうち「の」の1文字前に「も」でないものは、連用形ではない。

<table>
<thead>
<tr>
<th>平仮名</th>
<th>そだてる、はだてる、しょって、</th>
<th>ばてる、もてる、</th>
</tr>
</thead>
<tbody>
<tr>
<td>漢字</td>
<td>立てる、煽る、育てる、果てる、企てる、</td>
<td>隔てる、遮れる、覆える、持てる、拾てる、打てる、凍てる、当てる、</td>
</tr>
</tbody>
</table>

ii) 「で」の1文字前に「こ」であるものは、「こ」の1文字前に文頭になれば、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。また第二種の端として格助詞の「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、せ、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。

それぞれの項目には判定条件を設けなくても抽出精度が良いので、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るとき、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検索：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、これは連用中止法は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