5C−4

字面解析手法を用いた否定表現抽出法の評価
── 朝日新聞記事データへの適用 ──

下園 幸一　菅沼 明　牛島 和夫
(九州大学工学部情報工学科)

1. はじめに

我々の研究室で研究開発している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作業に役立つ情報を書き手に提供するツールである.[1]『推敲』で使用している字面解析手法は,実際の日本語文章を調査し,その結果を基に構築している.[2] 構築の際に使用した文章は科学技術文章であった.この字面解析手法が一般の文章に対しても有効であるかどうかを確認するために,今回は朝日新聞の半年分の記事データを使用して評価した.

2. 否定表現の抽出法

否定を表す単語には,形容詞及び助動詞「ない」,助動詞「ぬ」「まい」の3つがある.文献[2]による否定表現の抽出法をまとめると,2.1 ～ 2.3 のとおりになる.

2.1 形容詞及び助動詞の「ない」の抽出

「ない」の活用文字列(「なかろ」,「なかっ」,「なく」,「ない」,「なけれ」)を検索し,以下の判定条件をあてはめる.

判定条件: 「な」の1文字前は「少,危,切,行,損」ではない.但し,「少」の1文字前が「少,多」の場合,「切」の1文字前が「一」の場合,「行」の1文字前が数字または漢数字または「数」の場合は除く.

漢字表記の場合は,活用形の文字列を抽出するだけである.

2.2 否定の助動詞「ぬ」の抽出

2.2.1 連用形の「ず」

判定条件1: 「ず」の1文字前が表1に示した文字のいずれかであれば,その「ず」は否定の候補である.

判定条件2: 「ず」の1文字後が促音,撥音ならば,その「ず」は否定の候補でない.

判定条件3: 「ず」の1文字前が「か,さ,た,な,ま,ら,わ,が,ば,ち,り,ぎ,じ,び,け,て,め,れ,げ,ぜ,べ」のいずれかの場合,「ず」の2文字前の文字が平仮名または漢字のときに限って,その「ず」を否定の候補とする.

判定条件4: 「ず」の1文字後が「れ」の場合,「れ」の1文字後が「い,き,と,っ,つ,て,ん」のいずれかに限って,その「ず」を否定の候補とする.

2.2.2 終止形,連体形の「ぬ」,「ん」

「ず」と同じ判定条件を用いる.

2.2.3 仮定形の「ね」

判定条件: 「ね」の1文字後は「ば」である.

2.3 推量否定の助動詞「まい」の抽出

文字列「まい」を抽出する.

表1: 否定の助動詞「ず」の1文字前にくる可能性がある文字

平仮名	か,さ,た,な,ま,ら,わ,が,ば, い,き,ち,に,ひ,み,り,ぎ,じ,び, え,け,せ,て,ね,へ,め,れ,げ,ぜ,で,べ,と
漢字	干,居,見,似,射,煮,煎,卑, 宛,化,気,経,昏,出,消,寝,耽,貞,得,悉,来

3. 評価用データ

今回の評価には,新聞記事を使用した.これは,新聞記事の文章が科学技術文章より日常書かれる文章に近いと考えられるからである.さらに,大量の文章を機械可読な形で入手することができたためでもある.使用した新聞記事は朝日新聞の半年分(1988年前半)の記事データで,総文字数は20,969,926文字である.

4. 結果

否定候補の抽出法に沿って抽出した結果を表2に示す.この表で,候補とは抽出法により抽出してきた文字列の数であり,否定とは実際に目視で確認した否定文字列の数である.精度は否定の数を候補数で割り,百分率で表したものである.科学技術文献抄録での精度とは,文献[2]で評価した際に用いた科学技術文献抄録(総文字数 2,842,062文字)での抽出法の精度である.

否定表現抽出法の構築の際に設定した条件は,指摘すべきでないものまで指摘してしまう(第二種の誤り)のはある程度よいが,指摘洩れ(第一種の誤り)は犯してはならないということであった.

形容詞,助動詞「ない」については,科学技術文献抄録とほぼ同様な精度が得られた.「なく」の精度が科学技術文献抄録の場合と比べて落ちている.候補に含まれる第二種の誤り(1,249)のほとんど(84.8%)は,「なくなる」,「なくす」とその活用形であった.また,漢字表記の場合は,「無く」以外の精度は100%であった.「無く」の精度は27.3%であり,第二種の誤りは「無くす」,「無くなる」の活用形だけであった.

助動詞「ぬ」の場合,全体的に精度が低い.「ず」の抽出に関しては,第二種の誤り(3,353)の中で「まず」(31.6%),

Evaluation of a Textual Analysis Method to Extract Negative Expressions in the Writing Tools for Japanese Documents
Koichi SIMOZONO, Akira SUGANUMA and Kazuo USHIJIMA
Dept. of Comp. Sci. and Comm. Eng., Kyushu University

表 2: 否定候補の抽出法の評価

否定を表す文字列	候補	否定	精度	科学技術文献抄録での精度
なかろ	167	167	100%	100%
なかっ	8,008	8,005	99.9%	100%
なく	11,452	10,203	89.1%	93.7%
ない	62,380	61,935	99.3%	95.4%
なけれ	3,415	3,415	100%	100%
漢字表記	223	159	71.3%	73.9%
小計	85,645	83,884	97.9%	95.7%
ず	12,231	8,878	72.6%	68.7%
ぬ	2,255	1,981	87.8%	96.4%
ね	837	837	100%	100%
ん	38,221	3,480	9.1%	0.0%
小計	53,544	15,176	28.3%	49.4%
(「ん」を除く)	(15,323)	(11,696)	(76.3%)	(81.7%)
まい	2,045	654	32.0%	13.5%
計	141,234	99,714	70.6%	80.3%
(「ん」を除く)	(103,013)	(96,234)	(93.4%)	(92.0%)

科学技術文献抄録の「ん」は否定であるものが存在しなかった.

「わずか」(17.0%),「とりあえず」(7.5%) が多かった.「ぬ」の抽出の場合,第二種の誤りに,特に傾向はなかった.

否定の「ん」は助動詞「ぬ」の発音上の言い替えである.文章中で用いられる場合は,助動詞「ます」に続いて「ません」という形で用いられる場合が多い.実際に今回の場合,否定であった文字列 (3,480) の大部分は,「ません」(88.3%) であった.

推量否定の助動詞「まい」の場合,第二種の誤りとして抽出したもの (1,391) は,「〜てしまい」(25.0%),「あいまい」(20.6%), が多かった.

5. 考察

否定候補抽出法の精度は,「ん」の抽出を除く場合,新聞記事に対しても,科学技術文献抄録の場合と同様に有効であるということが分かった.また第一種の誤りはどの場合においても犯していなかった.

否定の「ん」は否定表現全体の 3.5% しかないが候補として抽出した「ん」は候補全体の 27% を占めている.そのため全体の抽出精度をかなり落している.

否定表現抽出法を構築する際,否定の「ん」に関しては,出現頻度が少なかったために,以下のような場合を考えた.

i) 「ん」を抽出しない
ii) 「ません」だけを抽出する
iii) 「ず」と同じ判定条件を使用する

しかし,場合 i, ii では,抽出法構築の際に設定した条件「第一種の誤りは犯してならない」に反する.ここでは「ず」の抽出法を改良して「ん」の抽出に適用することを考える.

今回の抽出法で「ん」を抽出したとき,第二種の誤りである「ん」の 1 文字前の文字を表 3に示す.

表3から「さん」という抽出候補に誤りが多いことが分かる.「さん」の 95%(18,183) は敬称の「さん」であった.この敬称の「さん」を抽出候補から取り除くために今回の記事データで敬称の「さん」の一文字前の文字を調べてみた.この結果最も多く出現した 26 文字 (表 4) を考慮することにより,敬称の「さん」を 50% 程候補から外すことができる.この方法を使用しても第一種の誤りは犯さない.

表 3: 第二種の誤りである「ん」の前に来る文字

文字	出現数	文字	出現数
さ	19,163	い	952
こ	2,607	み	842
な	2,006	ら	813
た	1,773	か	811
が	1,098	その他	4,634

表 4: 敬称の「さん」の前に来る文字

子, 田, 郎, 母, 藤, 父, 一, 野, 夫, 川, 奥, 雄, 木, 屋, 美, 本, 村, 皆, 井, 男, 原, 山, あ, 沢, 客, 谷

また,「ん」は用言の未然形に接続する. 表 1で示した文字は,未然形の活用語尾となる文字である.ここで,文字「な」はナ行五段活用動詞の活用語尾である.ナ行五段活用動詞は「死ぬ」しかないので, 文字「な」の前が「し,死」の場合のみ「な」の後の「ん」を否定として抽出すればよい. これより表 3に示した文字列「なん」を否定の候補から外すことができる.

判定条件 2 については「ん」の場合,「ん」の後に促音,撥音が来る単語がないので考慮する必要がない.判定条件 3 は用言の活用語尾についての条件である.公用データベース日本語辞書 [3] で調べた結果,活用語尾の前の文字が促音である場合はなかった.これを考慮すると「〜ったんです」のような表現を候補から外すことができる.また,「ん」の後に「れ」が来る単語自体がないので判定条件 4 も考慮する必要がない.

このように抽出法を改良すると,今回の記事データの「ん」の抽出候補は 38,221 個から 26,750 個となり,11,471 個減らすことができる.

また場合 ii について考察してみる. 文字列「ません」を今回の記事データから抽出してきた場合, その「ません」全てが否定表現であった.否定の「ん」の 90% 近くが「ません」であったことを考えると,場合 ii はある程度有効であることがわかる.

実際に否定表現抽出法を『推敲』などのツールに組み込む場合には,推敲の対象となる文章を考えて,ユーザが抽出法を選択できるようにするのがよいだろう.

参考文献

[1] 倉田, 菅沼, 牛島 : “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.

[2] 菅沼, 倉田, 牛島 :“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, June 1990

[3] 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).