否定の焦点コーパスの構築と自動検出器の試作

大槻 諒 † 松吉 俊 ‡ 福本 文代 ‡
† 山梨大学工学部 ‡ 山梨大学大学院医学工学総合研究部
{t09kg005, sugurum, fukumoto}@yamanashi.ac.jp

1 はじめに

自然言語処理の分野において、言語解析に関する研究は精力的に行われており、現在、誰もが自由に利用することができる形態素解析器や構文解析器・述語項構造解析器が存在する。文章に述語項構造解析を行うことにより、「誰がいつどこで何をするのか」という事象を自動的に認識することが可能になった。

事象の末尾に、「ない」や「ん」、「ず」などの語が付くと、いわゆる否定文となり、一般には、その事象が成立しないことが表現される。これらの語による否定の働きが及ぶ範囲をスコープと呼び、その中で特に否定される部分を焦点 (フォーカス) と呼ぶ [8]。例えば、次の否定文 (1) において、「ん」のスコープは「今日は車では来ませ」で表現される事象であり、焦点は「車で」である。

(1) 今日は車では来ませ ん でした。
(2) 別に入りたくて入った わけではない。
(3) 忙しかったので、その本を読ま なかっ た。

否定文 (2) において、否定の複合辞「わけではない」のスコープは「入りたくて入った」であり、否定の焦点は「入りたくて」である。否定文 (3) の「なかっ」のスコープは、「その本を読ま」で表現される事象である。「なかっ」の焦点は、無しとせず、便宜上、スコープ全体であると考える。

否定の焦点がスコープ全体でない場合、スコープの事象が成立しないことだけでなく、類似の別の事象は成立することが推測できる。例えば、上の否定文 (1) において、「今日車で来た」という事象は成立しないが、「今日来た」ことは成立することが推測できる。同様に、否定文 (2) では、「入りたくて入った」ことは否定されるが、「入った」ことは成立することが推測できる。既存の述語項構造解析の技術を用いれば、否定のスコープの事象を高い精度で認識可能であると思われる。その一方で、現在のところ、日本語において、否定の焦点を検出するツールは利用可能ではない。

本研究の目的は、日本語における否定の焦点を自動検出することである。本論文では、その基盤データとして構築したコーパスと、試作した焦点検出システムについて報告する。

2 関連研究

言語学の分野では、英語においても日本語においても、否定とその焦点に関する研究 [6, 8] がある。

自然言語処理の分野では、近年、否定のスコープを対象とした研究が増えている。英語では、生医学分野の文章を対象として否定のスコープをアノテーションした BioScope[3] を用いて、否定のスコープを解析する研究が主である。日本語では、新聞を対象として否定のスコープをアノテーションする研究 [4] がある。

否定の焦点を対象とする研究はほとんどない。英語においては、PropBank を対象として否定の焦点をアノテーションする研究 [1] や、否定の焦点を検出するためのヒューリスティックに関する研究 [2] がある。日本語においては、否定の焦点をアノテーションする研究 [5] はあるが、この研究で実際にアノテーションした事例は少ない。

3 否定の焦点アノテーション

この章では、実際に文章に否定の焦点をアノテーションする方法と判断基準について述べる。

3.1 否定要素

本論文では、文中において否定を表す表現のことを否定要素と呼ぶ。本研究では、次の 3 種類の語群をまとめたものを否定要素と定める。

否定辞 助動詞「ない」と「ず」、接尾辞「ない」、接頭辞「非」、「不」、「無」、「未」、「反」、「異」
非存在の内容語 形容詞「無い」、名詞「無し」
否定を表す複合辞 「のではない」、「わけではない」、「わけにはいかない」など

否定辞のみでなく、非存在の内容語まで含める理由は、「無い」は、存在の内容語「ある」の丁寧な否定「ありません」と同等であると思われるからである。否定辞「ん」が使用されている「ありません」は対象と

し、「無い」は内容語なので対象としないのは、不合理であると思ったからである。

言語学の文献 ([9] など) において、否定を表す複合辞とされる表現は、1 形態素の否定辞と異なる性質を持つと思われるので、区別して扱う。

以下のような、否定辞を含む 2 形態素以上の慣用表現は、全体を 1 語とし、否定要素の候補から除外する。

複合語 「物足り ない」、「仕方が ない」、「思わ ず」など

否定以外の意味を持つ複合辞 「なけれ ばなら ない」、「かもしれませ ん」、「だけで なく」など

助動詞「ない」か形容詞「無い」を使った単純な否定表現に言い換えられない否定の接頭辞も否定要素の候補から除外する。例えば、「不 十分」は、「十分でない」ことであるので、対象とする。一方、「不 気味」は、「気味が悪い」ことであり、「気味がない」や「気味でない」に言い換えられないので、候補から除外する。

3.2 否定のスコープ

本来ならば、文章中に明示的にアノテーションすべきではあるが、人的コストがかかるので、本研究ではアノテーションしない。人間が否定の焦点を判断する時には、文章全体を読み、その否定のスコープを目で確認することとする。

3.3 否定の焦点

1 章で述べたように、否定要素によって特に否定される部分が否定の焦点である。これを安定して判断するために、我々は次のような判断基準を定めた。

> 対象とする文から、一部の表現と否定要素を除外した事象を生成する。その事象が成立することが推測できれば、除外した表現の部分を否定の焦点と判断する。

例えば、1 章の例文 (1) の場合、一部の表現「車では」と否定要素「ん」を除外して、「今日来た」という事象を生成する。この事象は成立すると推測されるので、除外した「車 (で)」を否定の焦点と判断する。次の例文 (4) の場合、「あまり」と否定要素「なかっ」を除外して、「僕が朝食を食べた」という事象を生成する。

(4) 僕は朝食をあまり食べ なかっ た。

全く食べなかったわけではなく、少しは食べたことが推測できるので、この事象は成立すると判断し、除外した「あまり」を否定の焦点とする。

対象とする文だけでなく、周りの文脈を広く参照し、焦点を判断する必要がある。焦点をアノテーションする際には、次の情報も記述する[1]。

- ガ格やデ格、副詞などの項の情報
- ノデ節やテ節、連体節などの節の情報
- 「は」や「しか」などのとりたて詞の有無

便宜上、コーパスにおいて否定の焦点は代表 1 形態素にラベル付けする。これは、自動検出システムの出力結果との比較を容易にするためである。代表 1 形態素は、次のように定める。

- 内容語
- 複合語の場合、接尾辞を除く末尾の語
- 修飾語が存在する場合、それが係る末尾の語

4 コーパス構築

前章で説明したアノテーション法に基づき構築した 2 つのコーパスについて述べる。

4.1 楽天トラベル: レビューデータ

我々は、まず、楽天データ[2]の楽天トラベル: レビューデータを用いて、否定の焦点コーパスを構築した。

対象としたレビュー集合は、小池ら [7] が使用したものと同じである。90%以上の宿泊施設はレビュー数が 1 から 58 の範囲にあるという調査結果に基づき、レビュー数が 10 から 58 の範囲の宿泊施設の全体から、無作為に 40 の宿泊施設を抽出し、ラベル付与の対象とした。独自の文分割規則により半自動的に文分割を行い、5,178 文のデータを得た。形態素情報のみに基づいて抽出した否定要素の候補は、1,246 個であった。

4.2 現代日本語書き言葉均衡コーパス

我々は、次に、現代日本語書き言葉均衡コーパス (BCCWJ)[3]におけるコアデータ内の新聞 (PN) を用いて、否定の焦点コーパスを構築した。

BCCWJ 全体の約 1/100 のデータがコアデータに指定されており、このデータは、その他の部分と比較して高い精度で解析が施されている。コアデータの一部に言語学的情報を付与する場合、国立国語研究所が定めたファイル優先順位に従うことが推奨される。我々は、コアデータ内の新聞 340 ファイルのうち、優先順位が 1 から 54 までの “A” グループを対象とした。このデータにおいて、1 文もしくは文の断片を表す、XML の “sentence” 要素の数は 2,708 であり、形態素情報のみに基づいて抽出した否定要素の候補は、406 個であった。

4.3 コーパスの現状

コーパスは独自の XML 形式で表現されており、人間が見やすい HTML 形式に自動変換可能である。

[1] 詳細を記述した基準書は次のサイトで公開予定である。
http://cl.cs.yamanashi.ac.jp/nldata/negation/

[2] http://travel.rakuten.co.jp/

[3] http://www.ninjal.ac.jp/corpus_center/bccwj/

表 1: 否定要素候補の分布

	レビュー	新聞	計
助動詞	637	173	810
接尾辞	116	33	149
接頭辞	19	34	53
形容詞	211	53	264
名詞	28	6	34
否定複合辞	12	5	17
(上記小計)	(1,023)	(304)	(1,327)
複合語	94	30	124
その他複合辞	121	72	193
解析誤り	8	0	8
(上記小計)	(223)	(102)	(325)
計	1,246	406	1,652

表 2: 焦点の分布

	レビュー	新聞	計
副詞	140	18	158
ガ格	30	5	35
ヲ格	7	5	12
ニ格	49	11	60
デ格	17	6	23
マデ格	5	4	9
カラ格	3	2	5
ト格	3	1	4
その他の格	1	2	3
ノの項	20	7	27
連体の述語	8	8	16
接頭辞「全」	1	0	1
テ節	1	2	3
ト節	1	0	1
アスペクト	14	0	14
計	300	71	371

2 つのコーパス (「レビュー」と「新聞」) における、否定要素候補の分布を表 1 に示す。否定要素は、この表の上半分に示されている。2 つのコーパスにおいて、否定要素はそれぞれ 1,023 個と 304 個であり、いずれも、助動詞「ない」と「ず」が過半数を占めることが分かる。

2 つのコーパスにおいて、否定の焦点がスコープ全体でないものは、それぞれ 300 個と 71 個であった。「レビュー」では、29%(300/1,023) の否定要素が、「新聞」では、23%(71/304) の否定要素が、スコープ全体でない焦点を持つことが分かる。これらの焦点の分布を表 2 に示す。「レビュー」には、焦点が副詞である否定要素が多いことが分かる。「新聞」のデータ数が少ないので、確定的なことは言えないが、どの格が焦点になりやすいかも、2 つのコーパスで異なる傾向があるようである。

焦点である部分にどのようなとりたて詞が付いていたかを表 3 に示す。2 つのコーパスを合わせ、35%(128/371) の焦点に何らかのとりたて詞が付いていたことが分かる。焦点の候補に「しか」が付いていた場合、それは焦点である可能性がかなり高い。その一方で、「は」や「も」は、焦点がスコープ全体の事例にも多く出現するので、注意が必要である。

表 3: 焦点に付いていたとりたて詞

	レビュー	新聞	計
「は」	65	13	78
「しか」	34	7	41
「も」	7	1	8
「だけ」	0	1	1
計	106	22	128

5 焦点の自動検出

入力された文に存在する否定要素に対して、その焦点を検出するシステムを試作した。否定の焦点を正確に検出するためには、省略解析や照応解析、さらに、対象とする文だけでなく、その前後の文脈の情報も利用する必要がある。本研究では、焦点検出の第 1 歩として、文脈情報を用いず、構文解析によって得られる情報のみで焦点を検出するシステムを試作した。

本システムは、構文解析され、すでに否定要素が検出された 1 文を入力とし、その否定要素の焦点となる形態素に「焦点」ラベルを付けて出力する。焦点がスコープ全体であるとシステムが解析した場合は、便宜上、否定要素の形態素に「焦点」ラベルを付けて出力する。システムは、次のような、優先順位が付いた語の辞書を利用する。

1. 副助詞「しか」、「だけ」、「まで」、「ほど」
2. 程度の副詞「さほど」、「あまり」、「なかなか」、「そう」、「ちょうど」、「ほぼ」
3. 時間の副詞「とき」、「今回」、「次回」、「前回」
4. 様態の副詞「うまく」、「きちんと」、「ゆっくり」など
5. アスペクト「(し) きれ (ない)」
6. 副詞「ほとんど」

システムは、優先順位に従い、これらの語を探索する。

まず、システムは、否定要素の品詞が接頭辞である場合、焦点をスコープ全体であると解析する。そうでない場合、システムは、否定要素から文頭に向かって優先順位 1 の副助詞 4 語を探索する。この探索は、文頭の形態素に到達するか、もしくは、接続助詞か読点を発見するまで続ける。優先順位 1 の語が発見された場合、その前に存在する、接尾辞でない内容語を焦点であると解析する。優先順位 1 の語がいずれも発見されない場合、優先順位 2 の副詞 6 語を同様に探索する。これらの語が発見された場合、これを焦点であると解析する。優先順位 2 の語がいずれも発見されない場合、上と同様に、優先順位 3, 4, 5, 6 の語を順に探索し、発見されれば、それを焦点であると解析する。

表 4: システムの全体正解率と FOC 正解率

	レビュー	新聞
全体正解率	80%(351/437)	78%(237/304)
FOC 正解率	52%(64/124)	24%(17/71)

表 5: 「レビュー」に対する 2 値分類結果

↓正解 \ システム →	一部	全体	計
スコープの一部	72	52	124
スコープ全体	25	288	313
計	97	340	437

辞書の語がいずれも発見されない場合、焦点をスコープ全体であると解析する。

接続助詞か読点が発見された際に探索を打ち切るのは、否定が作用する領域を越えて探索を行わないようにするためである。

入力された文に複数の否定要素が存在する場合、それぞれに対して独立に上記の処理を行う。

6 実験

2 つのコーパスを用いてシステムの評価実験を行った。

6.1 実験方法

「レビュー」の半分のデータ (2,209 文、437 の否定要素) と「新聞」の全体を用いる。この実験はクローズドテストである。次の 3 つの尺度により、システムを評価する。

全体正解率 すべての否定要素に対する、スコープ全体を含む焦点の正解率

FOC 正解率 焦点がスコープ全体でない否定要素に対する、焦点の正解率

2 値分類正解率 焦点がスコープ全体か一部かの 2 値分類の正解率

6.2 実験結果と考察

システムの全体正解率と FOC 正解率を表 4 に示す。全体正解率はおよそ 80%であったが、FOC 正解率はかなり低いことが確認できた。この主な原因は、以下に例示する、述語の格が焦点である場合、および、「は」が付く項が焦点である場合を解析するのは難しく、まだ対処できていないからである。

(5) 風呂は 洗い場$_{焦点}$ に蛇口が 無かっ たので少し不便でした。

(6) また家庭と学校の不干渉が徹底している 欧米$_{焦点}$ では、家庭訪問は基本的に「ない」という。

頻度の副詞や数詞が焦点である場合と合わせ、これらの問題に取り組む必要がある。

「レビュー」と「新聞」に対する 2 値分類結果を、それぞれ表 5 と表 6 に示す。正解がスコープの一部であるのに、スコープ全体であると解析した誤りが多いことが見て取れる。

表 6: 「新聞」に対する 2 値分類結果

↓正解 \ システム →	一部	全体	計
スコープの一部	21	50	71
スコープ全体	11	222	233
計	32	272	304

7 おわりに

本研究では、日本語における否定の焦点を自動検出するための基盤データとして、否定の焦点コーパスを構築した。そして、試作した焦点検出システムについて報告した。今後は、焦点自動検出の問題を検討し、検出システムを改善していく予定である。

構築したコーパスは、楽天データおよび BCCWJ との差分形式で、前記のサイトにて無償で一般公開する予定である。

謝辞: 本研究の一部は、科研費若手研究 (B)「高精度モダリティ解析のための言語資源構築に関する研究」(課題番号: 23700176、代表: 松吉俊) の支援を受けている。

参考文献

[1] Eduardo Blanco and Dan Moldovan. Semantic Representation of Negation Using Focus Detection. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 581–589, 2011.

[2] Eduardo Blanco and Dan Moldovan. Some Issues on Detecting Negation from Text. In *Proc. of the 24th International Florida Artificial Intelligence Research Society Conference*, pp. 228–233, 2011.

[3] Veronika Vincze, György Szarvas, Richárd Farkas, Gy*ddotor*gy Móra, and János Csirik. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. In *BMC Bioinformatics*, pp. 1–9, 2008.

[4] 川添愛, 齊藤学, 片岡喜代子, 崔栄殊, 戸次大介. 言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4. Technical Report of Department of Information Science, Ochanomizu University, 2011.

[5] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 705–713, 2010.

[6] 加藤泰彦, 吉村あき子, 今仁生美（編）. 否定と言語理論. 開拓社, 2010.

[7] 小池惇爾, 松吉俊, 福本文代. 評価視点別レビュー要約のための重要文候補抽出. 言語処理学会第 18 回年次大会論文集, pp. 1188–1191, 2012.

[8] 日本語記述文法研究会（編）. 現代日本語文法 3. くろしお出版, 2007.

[9] 森田良行, 松木正恵. 日本語表現文型 用例中心・複合辞の意味と用法. アルク, 1989.