様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１１　日現在

機関番号：　１３５０１
研究種目：　若手研究(B)
研究期間：　2013～2014
課題番号：　２５８７０２７８

研究課題名（和文）否定焦点コーパス構築と焦点自動解析に関する研究

研究課題名（英文）Research on Corpus Construction and Text Analyzer for Negation Focus in Japanese

研究代表者
松吉　俊（MATSUYOSHI, Suguru）

山梨大学・総合研究部・助教

研究者番号：１０５１２１６３

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究の目的は、現代日本語における否定焦点の統計的分布を明らかにし、否定の焦点を自動的に特定する解析システムを実装することである。本研究では、独自にアノテーション体系を定め、ラベル付けをするとともに、統計的分布を活用し、解析システムを実装した。主な研究成果は、2,147文を対象とした否定焦点コーパスの構築、データ共有のためのデータフォーマット仕様の策定、そして、精度80%の焦点解析システムの実装である。

研究成果の概要（英文）：I have constructed language resources and tools for negation focus analysis in Japanese text. They include the following three main components: (1) a corpus of 2,147 pairs of negation triggers and their foci, (2) a new data format to support annotation layers for data sharing, and, (3) a detector of the focus of a negation trigger in Japanese, which achieved accuracy of 80 % in the domain of hotel reviews and news text.

研究分野：　自然言語処理

キーワード：　自然言語処理　コーパス言語学　否定表現　モダリティ　アノテーション　事実性解析

１．研究開始当初の背景

自然言語処理の分野において、文章を解析するための技術は古くから研究されており、これまでに様々な解析ツールが開発されてきた。例えば、形態素解析器や構文解析器は、その最も基礎的なものであり、現在、誰もが自由に利用することができるこれらの解析器が存在する。近年、テキストに存在する動詞や形容詞などの述語に対してその項構造を特定する技術、すなわち、「誰がいつどこで何をするのか」という事象を認識する技術が盛んに研究されており、英語においても日本語においてもその解析ツールが公開され、その利用を前提とした研究を進めることが可能になってきた。事象認識に続く研究として、事象の肯定・否定を自動認識するための研究を実施することはごく自然である。

事象の末尾に、「ない」や「ん」、「ず」などの否定辞が付くと、いわゆる否定文となり、一般には、その事象は事実でないことが表現される。否定の働きが及ぶ範囲の事象をスコープと呼び、その中で特に否定される部分を焦点と呼ぶ。例えば、次の否定文において、否定辞「ん」に対して、「今日は車で来た」がそのスコープであり、「今日は」がその否定辞の焦点にあたる。

今日は車で来ませんでした。

この文では、「今日、車で来た」という事実は否定されているが、「今日来た」ことは事実であると解釈できる。つまり、焦点のみが否定されており、スコープから焦点を除去した事象は事実である。計算機が的確に否定の焦点を認識することができれば、上の例のように、より正確に情報の事実性を自動的に認識することが可能となる。

言語学の分野においては、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先行研究が多数存在する。しかしながら、現在のところ、日本語を対象として、実際に否定の焦点をラベル付けした大規模なテキスト言語資源(以下、「コーパス」と呼ぶ)や、否定の焦点を自動的に特定する解析システムは、利用可能ではない。

２．研究の目的

本研究では、現代日本語における否定焦点の統計的分布を明らかにするとともに、否定の焦点を自動的に特定する解析システムを実装する。具体的には、以下の 3 点を研究の目的とする。

(1) 否定の焦点をラベル付けしたコーパスの構築

『現代日本語書き言葉均衡コーパス』等を対象として、否定辞、その焦点、そう判断した根拠をラベル付けし、否定焦点コーパスを構築する。

(2) 否定焦点の分析

構築したコーパスを観察し、現代日本語における＜否定辞, 焦点＞の対の分布を明らかにする。焦点の種類ごとに、その特定に有用な語句・文脈などについて詳細に分析する。

(3) 解析システム実装

上記のコーパスと言語学的知見を利用し、否定の焦点を特定する解析システムを構築する。

３．研究の方法

本研究では、次の 3 点を中心に研究を行い、現代日本語における否定焦点の統計的分布を明らかにするとともに、否定の焦点を自動的に特定する解析システムを実現することを目指した。

(1) 否定の焦点をラベル付けしたコーパスの構築
(2) 否定焦点の分析
(3) 解析システム実装

(1) 否定の焦点をラベル付けしたコーパスの構築

①　否定の焦点に関する文献調査

否定辞、および、否定の焦点に関する先行研究と文献を調査し、その調査結果をコーパスのラベル付けに応用した。

②　コーパスの整備

楽天データの「楽天トラベル: レビューデータ」と、現代日本語書き言葉均衡コーパスの新聞データをテキストデータとして利用した。前年度から我々のグループで構築を進めていた否定焦点コーパスにおけるラベルを見直すとともに、新しい情報を付与し、コーパスを整備した。

③　データフォーマット仕様の策定

自然言語処理コミュニティにおいて、複数のグループが同じテキストデータに対して付与した多くの情報をうまく共有するための新しいデータフォーマット仕様を策定した。そして、上記のコーパスを、XML 形式からこのデータフォーマット仕様の形式に自動変換するプログラムを実装した。

(2) 否定焦点の分析

構築したコーパスを観察し、現代日本語における＜否定辞, 焦点＞の対の分布を明らかにした。

(3) 解析システム実装

①　システム実装

否定辞の種類、周辺のとりたて詞、否定辞とよく共起する語句等を自動的に検出し、それらの情報を利用して否定の焦点を解析するシステムを実装した。

②　システムの評価

構築したコーパスを用いて、提案する解析システムを評価した。

４．研究成果

本研究の遂行により、現代日本語に関して否定の焦点を解析するために必要となる基盤言語資源とツールを構築できたのではないかと思われる。特に、本研究の主な成果である、前章の(1)の②と③で整備したコーパスは、自然言語処理での利用を考慮した設計を採用しており、解析システム構築時の学習データとして役立つだけではなく、この種の情報ラベルを文章に付与する時の標準規格としての性格を有していると思われる。

近年、英語においては、否定のスコープ・焦点に関する大規模なコーパスが構築され、公開・利用されるようになった。一方、日本語においては、このようなコーパスは存在しなかった。本研究で構築したコーパスや解析システムは、日本語を対象とした意味処理技術の発展のために利用することができ、これらの言語資源をコミュニティで共有できることは、大きな意義があると考える。加えて、データを蓄積・共有するための新しいデータフォーマット仕様を策定し、公開したことは、コーパスを利用する言語研究分野の発展に深く貢献できたのではないかと思われる。実装した解析システムは、事実性解析や情報検索・情報抽出の発展を促す意義のある成果であると考える。

今後の展望としては、実装した解析システムを実際の意味処理タスクに応用することが考えられる。事実性解析や含意関係認識などの応用の観点から、解析システムの評価をすることが必要であると思われる。否定の焦点を解析する際に前文までの文脈情報を必要とする事例に関してはまだうまく分析できていない。本研究の延長として、新たなテキストを対象としてコーパスの規模を大きくし、多くの事例を観察することにより、文脈の適切な扱いについて研究することが考えられる。このような談話構造を捉える研究は今後の課題としたい。

本研究の具体的な研究成果を以下に列挙する。

(1) 否定の焦点をラベル付けしたコーパスの構築

① 否定の焦点に関する文献調査

否定辞、および、否定の焦点に関する先行研究と文献を調査した。この調査の結果、否定要素と呼応する程度・頻度の副詞が重要であり、特に、完全否定と弱否定の分類が焦点の有無に深く関わることが分かった。日本語においては、とりたて詞の種類が否定焦点の位置に影響し、とりたて詞のスコープと否定辞のスコープの包含関係により、文の解釈が決定することを考慮する必要があるという知見も得られた。

② コーパスの整備

楽天データの「楽天トラベル: レビューデータ」の 5,178 文と、現代日本語書き言葉均衡コーパスの新聞データの 5,582 文を対象としてコーパスを構築した。否定辞を含む文 2,147 文に対して、否定とその焦点に関するラベル情報を精練し、コーパスを整備した。

③ データフォーマット仕様の策定

否定焦点コーパスをコミュニティで共有するためのデータフォーマット仕様を提案した。さらに、この仕様に従った否定焦点コーパスも自動生成した。

(2) 否定焦点の分析

否定のスコープの一部に焦点がある場合は全体の約 23%であることや、テキストのジャンルによって焦点部分の品詞列に異なる傾向があること等、統計的分布に関する知見が得られた。また、否定焦点をとりまく文脈について分析を行った結果、対象文内の内容語とその前文脈の内容語との関連性を単純に計測するだけでは、否定焦点の特定に十分ではなく、もう少し深い意味解析が必要であることが分かった。

(3) 解析システム実装

① システム実装

分析した否定焦点の統計的分布に基づいて、ルールベースにより否定の焦点を解析するシステムを実装した。このシステムは、大分類で 14 個のルールを利用する。手がかり語句としては、主に、副詞、とりたて詞、構文パターンを利用する。否定辞を含む 1 文が入力されると、システムは、その文に存在する手がかり語句を検出した後に、優先度に応じてルールを適用し、否定の焦点を解析する。

② システムの評価

構築したコーパスを用いて評価した結果、提案システムは、レビューデータにおいて 78%、新聞データにおいて 80%の正解率を達成した。単純な規則に基づくベースラインシステムの正解率は、それぞれ 68%と 73%であり、その性能を 10%ほど上回る成果を達成することができた。

５．主な発表論文等

〔雑誌論文〕（計１件）

松吉 俊、否定の焦点情報アノテーション、自然言語処理、査読有、Vol.21、No.2、2014、pp.249-270
DOI:10.5715/jnlp.21.249

〔学会発表〕（計５件）

①松吉 俊、否定の焦点情報アノテーション、言語処理学会第 21 回年次大会 招待論文セッション、2015.3.18、京都大学(京都府・京都市)

② Suguru Matsuyoshi, Ryo Otsuki, and

Fumiyo Fukumoto, Annotating the Focus of Negation in Japanese Text, Proceedings of the 9th edition of the Language Resources and Evaluation Conference, pp. 1743-1750, 2014.5.29, Reykjavik (Iceland)

③名取 芙美香、松吉 俊、福本 文代、含意認識タスクに関するかき混ぜ文対データの構築、言語処理学会第 20 回年次大会、pp.745-748、2014.3.20、北海道大学（北海道・札幌市）

④松吉 俊、浅原 正幸、飯田 龍、森田 敏生、拡張 CaboCha フォーマットの仕様拡張、第 5 回コーパス日本語学ワークショップ予稿集、pp.223-232、2014.3.7、国立国語研究所（東京都・立川市）

⑤磯野 史弥、松吉 俊、福本 文代、Web 掲示板における皮肉の分類および自動検出、情報処理学会 第 213 回自然言語処理研究会、Vol.2013-NL-213、No.7、pp.1-8、2013.9.12、山梨大学(山梨県・甲府市)

〔その他〕
①日本語否定の焦点情報アノテーションコーパス　公開ページ:
http://cl.cs.yamanashi.ac.jp/nldata/negation/

②本研究成果に関して論文誌『自然言語処理』に投稿した雑誌論文「否定の焦点情報アノテーション」が、2014 年度論文賞を受賞したことを報道する言語処理学会のページ:
http://www.anlp.jp/award/ronbun.html

６．研究組織
(1)研究代表者
松吉　俊（MATSUYOSHI, Suguru）
山梨大学・総合研究部・助教
研究者番号：１０５１２１６３