様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年５月３１日現在

機関番号：１３５０１
研究種目：若手研究（Ｂ）
研究期間：2011～2012
課題番号：２３７００１７６

研究課題名（和文）　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

研究課題名（英文）　Research on creation of language resources for high-precision modality analysis in Japanese

研究代表者
松吉　俊（MATSUYOSHI SUGURU）
山梨大学・医学工学総合研究部・助教
研究者番号：１０５１２１６３

研究成果の概要（和文）：本研究の目的は、文章の書き手が表明している心的態度や真偽判断、価値判断などの情報（事象のモダリティ）を解析するために必要となる言語資源を構築することである。本研究では、モダリティタグ付与コーパスの体系を洗練し、特に否定表現に関して言語資源を構築した。主な研究成果は、約 4 万事象のモダリティコーパスの公開、約 1,800 の否定文に対する否定の焦点コーパスの構築、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 Japanese modality analysis. They include the following three main components: (1) a corpus of 41,799 events with extended modality tags, which are extracted from the latest version of “Balanced Corpus of Contemporary Written Japanese”, (2) a corpus of 1,785 pairs of negation triggers and their foci, and, (3) a detector of the focus of a negation trigger in Japanese, which achieved accuracy of 80 % in the domain of posted reviews about hotel service.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理、モダリティ解析、言語学、言語資源、機能表現

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。

テキスト解析の流れにおいて、この述語項構造解析の後には、そこで特定された事象に対して文章の書き手が表明している心的態度や真偽判断、価値判断などの情報（事象のモダリティ）を解析するのが自然である。例えば、次の例文の下線の事象に対して、その直下に示すような解析結果を出力することができれば、テキスト理解や高度な情報検索・情報抽出などの応用に直接役立てることができる。

(1) 彼が薦めていたら商品 A を購入したのに。
解析結果: “叙述,不成立,後悔”
(2) 商品 A を買うつもりだ。
解析結果: “意志,高確率で成立,肯定的価値”
(3) お金無いけど、商品 A を買いたいなあ
解析結果: “欲求,成立可能性不明,肯定的価値”

しかしながら、現在のところ、この事象のモダリティを高い精度で解析するシステムは英語においても日本語においても提案されていない。先行研究としては、このようなモダリティをマークアップする手法や、その一部の情報を解析する手法が提案されているのみである。

ブログに代表される個人型情報発信メディアの爆発的な普及に伴い、膨大なテキスト情報が Web 上に加速度的に蓄積されつつある。これらの情報を目的に合わせてうまく整理し、そこから有益なデータを得るためには、記述される個々の事象に対して、「誰がいつどこで何をするのか」のみを認識するだけではなく、前述のモダリティの情報も解析し、総合的に分析を行う必要がある。このような情報は、対象としている事象が実際に成立した事実であるのか、それとも、成立しなかったことであるのか、もしくは、書き手がその成立を望んでいるだけであるのか等を情報の受け手に伝える重要な役目を担っている。

２．研究の目的

本研究の目的は、文章の書き手が表明している心的態度や真偽判断、価値判断などの情報(事象のモダリティ)を解析するために必要となる言語資源を構築し、実際にこの解析システムを実装することである。前章の例にて示した解析結果を出力することにより、テキスト理解や高度な情報検索・情報抽出に役立てることを目指す。具体的には、以下の 3 点を研究の中心とする。

(1) モダリティ解析のための言語資源構築。
言語学的観点から事象のモダリティ解析に必要な言語資源について検討し、実際にこれらの言語資源を構築する。
(2) コーパス構築。
テキストに存在する事象に対してモダリティのラベルを付与するとともに、上記の言語資源を利用し、これらのラベル推定に有用な情報をマークアップする。
(3) 解析システム実装。
上記の言語資源とコーパスを利用し、モダリティ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、文章の書き手が表明している心的態度や真偽判断、価値判断などの情報を高い精度で解析するシステムを実現することを目指した。

(1) モダリティ解析のための言語資源構築
(2) コーパス構築
(3) 解析システム実装

(1) モダリティ解析のための言語資源構築
① 言語資源に関する調査

文に存在する述語のモダリティを決定する語句にどのようなものがあるか、言語学の先行研究を調査した。具体的には、モダリティや談話に関する言語学の文献を収集し、これらの語句について整理した。

(2) コーパス構築
① 既存のコーパスの洗練

これまでに構築した拡張モダリティタグ付与コーパスのタグを見直した。具体的には、先行研究を参考にし、モダリティタグを付与する事象の細分類を行った。加えて、事象性に関して新しいラベルを導入した。

既存のコーパスは、2009 年版の「現代日本語書き言葉均衡コーパス」(BCCWJ)を元に設計されていたため、最新版の BCCWJ とは互換性がなかった。この対応を取るため、最新版のデータ形式を解析し、ここにモダリティタグを追加することができるツールを実装した。

(1) モダリティ解析のための言語資源構築
② 否定要素に関する調査

上記の調査により、文に存在する述語のモダリティを決定する語句のうち、真偽判断と価値判断の決定に最も重要であるのは、否定要素であることが分かった。それゆえ、否定要素に関して、言語学の先行研究を詳しく調査した。具体的には、否定辞の種類、否定のスコープ、否定の焦点に関する言語学の文献を収集し、これらの言語現象について整理した。同時に、自然言語処理の分野において、否定とそのスコープ・焦点がどのように自動検出されているか調査した。

(2) コーパス構築
② 否定の焦点コーパスの構築

日本語における否定の焦点のコーパスを設計し、実際にコーパス構築作業を行った。具体的には、構築した言語資源に基づき、検出する否定要素を定義した。そして、その否定要素が文内に持つ焦点を人手でマークアップした。マークアップするにあたり、最適な XML フォーマットを設計した。コーパスには、否定の焦点の情報だけでなく、焦点とな

る項の種類や意味分類などの有益な情報も同時に付与した。

(3) 解析システム実装

① 否定の焦点解析システムの実装

上記の言語資源とコーパスを利用し、日本語の否定の焦点を解析するシステムを構築した。具体的には、コーパス内の事例を観察することにより、否定の焦点に関する特徴的なパターンを特定し、これらをデータベースに集積した。解析時には、手がかり語句に関する言語資源と、このパターンデータベースを利用して、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、否定の焦点解析システムを評価した。具体的には、宿泊施設のレビューデータと新聞データの２種類を利用して、それぞれシステムの正解率を測定した。

４．研究成果

本研究の遂行により、日本語モダリティを解析するために必要となる基盤言語資源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(2)の①と②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、モダリティ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、モダリティや否定のスコープ・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。

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。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。1 つのモダリティに複数の語句が関連している場合、その相互作用についてはまだうまく体系化できていない。本研究の延長として、コーパス内の事例を観察することにより、この相互作用に関する言語資源を構築することが考えられる。これに関す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) モダリティ解析のための言語資源構築

① 言語資源に関する調査

文に存在する述語のモダリティを決定する語句にどのようなものがあるか、言語学および日本語教育学における関連研究の文献を広く収集することにより調査した。この調査の結果、副詞や機能語、複合辞に代表される明らかな手がかり語に加え、動詞や形容詞などの用言が語彙的な知識としてモダリティの決定に大きく寄与することが再認識された。モダリティ解析のための言語資源を構築する際には、動詞や形容詞などの用言の意味分類を考慮し、これらの体系をうまく取り込む必要があるという重要な知見が得られた。

② 否定要素に関する調査

「ない」や「ず」、「不」に代表される否定要素に関して、言語学の先行研究を詳しく調査した。次のように、否定辞や、内容語であっても否定要素として働く語句などを体系的に集積した。

助動詞: 「ない」、「ず」
接尾辞: 「ない」
接頭辞: 「非」、「不」、「無」、「未」、「反」、「異」
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

さらに、否定のスコープと否定の焦点に関する言語学の文献を収集し、これらの言語現象について整理した。自然言語処理の分野における先行研究も調査し、英語においては否定の言語現象に関する言語資源や解析ツールの整備が進んでいるが、日本語においてはそのようなものがほとんど利用可能でないことが分かった。

(2) コーパス構築

① 既存のコーパスの洗練

これまでに構築した拡張モダリティタグ付与コーパスを見直した。含意認識のタスクにおいては、文章に存在するすべての事象を認識することが重要となる。既存の体系では、ある語句が事象の述語であるかどうかは、確定的に定めていた。「事象可能」というクラスを用意することにより、柔軟に事象認識できるよう、ラベル体系を改めた。この体系に従い、上記のコーパスの一部に、新しい事象ラベルを人手で付与した。

既存のコーパスは、2009 年版の「現代日本語書き言葉均衡コーパス」(BCCWJ)を元に設計されていたため、最新版の BCCWJ とは互換性がなかった。この対応を取るため、最新版のデータ形式を解析し、ここにモダリティタグを追加することができるツールを実装し

た。そして、モダリティタグの情報を差分データとして、ツールとともにWeb上で一般公開した。公開したコーパスには、約4万の事象が収録されている。

② 否定の焦点コーパスの構築

日本語における否定の焦点のコーパスを設計した。以下のような、ラベル付与の体系を独自に定義した。

[否定要素]
- 表層文字列: 文に出現した否定要素の表層文字列
- 形態素ID: 否定要素の形態素のID
- 品詞: 助動詞、接尾辞、接頭辞、形容詞、名詞、否定複合辞のいずれか
- 最終更新日: 更新日の情報

[否定の焦点]
- 代表表層文字列: 焦点の表層文字列。ただし、代表形態素のみを記述
- 代表形態素ID: 焦点の代表形態素のID
- 項・節の種類: ガ格、ヲ格、デ格、副詞、ノの項、ナの項、テ節、ト節など、焦点の統語的分類
- とりたて詞の有無: 「しか」や、数量語に付く「も」が存在するかの情報
- 意味分類: 制限-時間、制限-場所、制限-対象、付加-連用修飾、付加-連体修飾、付加-アスペクトなど、意味解釈に基づいた、否定されている語句の分類
- 判断の根拠: その箇所を焦点であると判断するに至った根拠。自由記述
- 手がかり語句: 文章中に存在する、焦点判断の手がかりとなった語句

「楽天トラベル: レビューデータ」の5,178文と、BCCWJの新聞5,582文に対して、上記の体系に従い、XMLフォーマットを用いて否定要素とその焦点となる形態素にラベルを付けた。このコーパスも差分データとして、関連ツールとともにWeb上で一般公開する予定である。

(3) 解析システム実装

① 否定の焦点解析システムの実装

構築した言語資源とコーパスを利用し、日本語の否定の焦点を解析するシステムを構築した。コーパス内の事例を観察することにより、否定の焦点に関する特徴的なパターンを特定し、これらをデータベースに集積した。これらのパターンを16種類の規則集合に整理し、それらの間に優先順位を設定した。作成した焦点解析システムは、文の構文解析結果を入力として受け取り、上記の優先順位に従って、文中の手がかり語句とこれらの規則集合を用いて、文章に存在する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、否定の焦点解析システムを評価した。クローズドテストのデータを用いて、パラメーターを調節した。クローズドテストでは、提案システムは、約86%の正解率を達成した。オープンテストでは、「楽天トラベル: レビューデータ」とBCCWJの新聞において、それぞれ約73%と約80%の正解率を達成した。日本語における焦点解析システムの先駆的研究として、これは高い精度であると思われる。今後は、広い範囲の文脈情報を参照することにより、焦点解析の精度を向上させることが課題である。コーパスにラベル付けした、項・節の種類や意味分類の情報などをうまく活用することも、さらなる精度向上につながると思われる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１２件）

①大槻諒、松吉俊、福本 文代、否定の焦点コーパスの構築と自動検出器の試作、言語処理学会第19回年次大会、2013年3月15日、名古屋。

②松吉俊、大槻諒、福本文代、日本語における否定の焦点アノテーション、第3回コーパス日本語学ワークショップ予稿集、2013年3月1日、東京。

③鈴木敬文、阿部佑亮、宇津呂武仁、松吉俊、土屋雅稔、代表・派生関係を利用した日本語機能表現の解析方式の評価、言語処理学会第18回年次大会、2012年3月15日、広島。

④Yusuke Abe, Takafumi Suzuki, Bing Liang, Takehito Utsuro, Mikio Yamamoto, Suguru Matsuyoshi and Yasuhide Kawada, Example-based Translation of Japanese Functional Expressions utilizing Semantic Equivalence Classes, In Proceedings of MT Summit XIII 4th Workshop on Patent Translation, 2011.09.23, China.

他８件

〔その他〕

拡張モダリティタグ付与コーパス 公開ページ:
http://cl.cs.yamanashi.ac.jp/nldata/modality/

６．研究組織

(1) 研究代表者

松吉　俊（MATSUYOSHI SUGURU）
山梨大学・医学工学総合研究部・助教
研究者番号：１０５１２１６３

(2) 研究分担者

なし

(3) 連携研究者

なし