様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月１５日現在

機関番号：６２６１５
研究種目：研究活動スタート支援
研究期間：平成２２年度～平成２３年度
課題番号：２２８０００７８
研究課題名（和文）日本語文の意味構造解析技術確立に向けた意味構造タグ付きコーパスの作成と応用
研究課題名（英文）Building semantic structure annotation corpus to establish semantic structure analysis technology for Japanese
研究代表者
松林　優一郎（MATSUBAYASHI YUICHIROH）
国立情報学研究所・コンテンツ科学研究系・特任研究員
研究者番号：20582901

研究成果の概要（和文）：
日本語文章の意味構造を自動解析する技術を確立する目的で、解析の規範となり、広く応用可能となりうる意味タグ付き文書を作成した。本研究では、特に動詞周辺の意味的構造に着目したが、言葉の意味という明快でない事象を扱うため、まず予め、意味タグ付き文書作成の基盤とする見通しのよい意味構造の理論を設計し、これまで曖昧だった意味タグを明確に定義した。その上で、新聞記事に出現する主要な動詞周辺の意味構造を対象に、この理論を用いた意味タグ付き文書のプロトタイプを作成した。

研究成果の概要（英文）：
To automatically analyze the semantic structures of Japanese texts, we created semantically annotated texts that can be a guideline of the analysis for computers. To analyze texts with computers we need a systematic representation of the meaning of language, so we first developed a broadly-applicable theory of meaning to act as a foundation, defining a set of guidelines for representing the semantic structures of verbs. Then, focusing on verbs frequently appearing in news articles, we annotated a collection of texts using these guidelines. The concrete results of our research are a theory for representing the verb meaning of Japanese and a collection of text annotated with these guidelines.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	740,000	222,000	962,000
2011 年度	1,090,000	327,000	1,417,000
年度			
年度			
年度			
総　計	1,830,000	549,000	2,379,000

研究分野：自然言語処理
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　意味役割　述語項構造　語彙概念構造　語彙意味論　コーパス

１．研究開始当初の背景

近年、組織レベルから個人レベルまで、電子媒体を用いた情報発信が爆発的に進み、大

量の知的資源が Web やデータベースに蓄積されている。その多くは文書により発信されており、これらの知的情報を適切に収集整理し可視化するための技術を構築することは、言語処理分野に課された課題の一つである。しかし同時に、高密度な情報を持った利用価値の高いテキストほど、ユーザーに提示する情報の正確性が求められており、大規模な情報集約を達成するためには、文の内容を構造的に捉える高度な意味解析技術が必要とされる。

このような文の構造解析技術のうち、構文構造解析に関しては、今日までに既に実用レベルに到達しつつある。特筆すべきなのは、９０年代より言語処理の基盤技術の全般が、正解タグ付きコーパスを利用した統計的モデル獲得の手法により飛躍的な発展を遂げた点である。正解タグ付きコーパスの効能は、統計手法の導入だけに留まらず、各要素技術において開発の具体的な指針とより正確な評価をもたらすことで、開発と評価の効率的かつ潤滑なサイクルを形成する鍵となってきた。文の構文構造を解析する技術においても、正解タグ付きコーパス主導の手法により、ニュースドメイン、バイオドメインなどで高精度の解析が行えるようになった。

意味構造解析については、近年、英語圏において、FrameNet、PropBank 等の意味タグ付きコーパスが登場した事によって同様の飛躍的発展を見せつつあり、また統語タグ付きコーパスと意味タグ付きコーパスが統合的に整備されたことによって、統語解析と意味解析の融合の兆しも見せつつある。一方、日本語においては、これまで本格的な意味タグを持ったコーパスが存在せず、効率的な技術開発のための正解コーパス整備が急務の課題となっている。

一方、研究代表者はこれまでに文の意味構造、特に意味役割と呼ばれる情報で扱われる構造に関心を持ち、英文の意味構造解析に利用されている主だった二つのコーパスにおいて、統計モデルの改善や、異なる言語資源間の多様な意味タグ情報を統合的に処理するための方法について議論してきた。英語圏におけるこれらの基盤資源と応用技術の発展を受け、日本語における今後の意味構造解析研究の基盤とすることのできる意味タグ付きコーパスを早期に整備する必要があるとの考えに至った。また、意味タグ付きコーパスにおいて今日問題とされている意味タグの定義について、これまでの研究により、述語の概念構造式を基準とすることで明示的な定義を与える手法の着想に至り、日本語意味構造付きコーパス整備の過程で、実証的な検証を行うべく、本提案を申請したものである。

２．研究の目的

本研究課題では中規模のコーパスに統一的な理論で意味タグを与えることで、日本語意味解析技術の土台となるデータを作成することを目的とする。具体的には、日本語統語解析のための主たるコーパスである京都大学テキストコーパス（京大コーパス）と同一の文書範囲に、上位のレイヤーとして、出現する全述語（動詞、事態性名詞）の、同一文中の項との意味的関係を付与する。

また、英語圏の既存のコーパスにおいて明示的な定義の無かった意味ラベルセットに関して、述語の概念構造式と各意味ラベルとの構造的な関係を根拠に、明示的な定義を与えることも研究の対象とする。関連する研究として、近年「ガ」「ヲ」「ニ」などの文中の表層格関係のタグ付けが行われたが、これらの統語機能的なタグは述語間で意味の揺れが大きく、異なる述語間での体系的な意味の取り扱いが出来ない。本研究で作成するコーパスの目的は、異なる述語間で統一的な扱いのできる意味タグの付与を実現することであり、これは、複数文あるいは文書間での体系的な意味構造解析を行うための有益な情報をもたらす。この統一的な意味を持つタグセットにより、文の構造的な意味解析を必要とする「関係抽出」「機械翻訳」等の技術への貢献のみならず、述語間の意味的関係情報を必要とする「言い換え」、「含意関係抽出」等への応用も期待される。

本研究の長期的な目標は京大コーパスの全域に対して意味構造を付与することであるが、研究期間を考慮に入れて、初期段階のプロトタイプデータを作成することを目標とする。また、句の省略にあたるゼロ代名詞については、本研究では取り扱わない。

また、作成したデータが実際に解析技術の構築に機能することを確認する目的で、機械学習による意味構造解析器のプロトタイプを作成する。

タグ付け作業者間の合意度や意味構造解析器の学習曲線により、コーパスの定量的評価を行うところまでを本研究の目的とする。

３．研究の方法

(1)動詞意味論の拡張

本研究では、各動詞が項として持つ意味役割に体系的な説明を与えることが目的の一つであるため、動詞内部の意味的構造を構成的に表現することの出来る理論である語彙概念構造理論（Lexical Conceptual Structure: LCS）を用いて、コーパスに出現する述語に適切な意味構造を与える。この際、

大量の実データに出現する多種多様な述語を頑健に表現できる理論とする必要があるため、コーパス中の実際の文章例を根拠に、既存の理論の拡張を行う。理論の拡張過程では、コーパスに高頻出の動詞から段階的に概念構造と呼ばれる意味構造を実際に記述し、実例文の説明が可能であることを検証しながら行う。語彙概念構造の設計時は、動詞項構造シソーラスや京大格フレームなどの既存の有用なデータを適時参照しながら作成する。

(2)述語項構造アノテーションのための意味タグ整備

タグ付与対象の京都大学テキストコーパス中の述語について、高頻度のものから１００個程度について、実際の文中の述語項構造を分析し、述語毎に語彙概念構造を決定する。この段階では文へのタグ付け作業は参考適度に行うに止め、項構造と意味タグセット全体の設計方針を固めることに注力する。意味タグの設計については、言語学で広く使われており、英語圏の主要なコーパスでも利用されている、述語に独立な意味概念を持った閉じたラベル集合である主題役割と比較をしながら行う。ただし、主題役割は研究者によって多くのバリエーションがあり、ラベルの数も６～３０個など様々であるため、本研究では、語彙概念構造の構成性を利用して、既存の主題役割の意味を整理し、厳密な形で再定義する。この方法で作成された意味タグは、動詞の概念構造から明確に意味付けされ、文の構文構造と結び付けられるため、文の意味に深く取り入る必要のある本研究のタグ付けに際しても、作業者間の揺れが減少することが期待できる。

(3)反復的小規模アノテーションとタグ付け方針の決定

タグ付け方針を決める目的で、反復的な小規模のアノテーションを行う。具体的には、まず、小規模な文書群を観察しながら簡単なタグ付け方針を設計した上で、タグ付け対象のコーパスから、１０単語程度の対象述語を選び、その語が含まれる文をランダムに１００文選択する。次に、二人のタグ付け作業者がこれらの１００文に互いに独立にタグ付けを行う。作業の後、タグ付け結果を比較し、結果の異なった部分や作業中に疑問を感じた部分について、合議の上タグ付け方針を取り決め、その方針に従い、結果を修正する。タグ付け方針更新の後、再び新たな動詞１０単語程度に対し１００文をタグ付けすることを適切回繰り返す。この過程において、作業者間で９０%以上の合意を得られるタグ付け方針を作ることを目標とする。

(4)コーパスの定量的評価・言語学的分析

作成したコーパスに対して、同一文に対する作業者間のタグの一致度を測ることで、タグ付けの質を定量的に評価する。また、共通の意味役割を取る動詞クラスの解析、本研究の理論で定義された意味ラベルと、従来研究の意味ラベルの比較等、理論的な考察を行い、研究成果として国際会議での報告を行う。

４．研究成果

２２年度の研究では、述語項構造タグ付きコーパス作成時に必要となる、解析の規範・基準となる意味論を構築した。具体的には、その基本的な枠組みとして語彙概念構造を採用し、これを実際のコーパスに出現する多様な述語表現を説明出来る形に拡張した。コーパス主導型の分析では、様々な言語現象が多種多様に現れるため、その基軸となる理論には高い一般性と頑健性が求められる。そのような背景から、語彙概念構造の表現形式を厳密性、一般性、柔軟性の観点で拡張し、日本語動詞の比較的広い範囲を説明することの出来る理論を構築した。また、拡張した理論を用いて、実際に対象コーパス中に頻出する述語に対して語彙概念構造を定義し、理論の頑健性を確認した。

さらに、これらの過程において、第一に、これまで既存の述語項構造解析の枠組みでは不明瞭であった意味役割ラベルの定義を、拡張された語彙概念構造を用いて基本となる意味機能に分解することにより、より明示的に説明できることを示した。第二に、上記で各述語に対して定義した概念構造を利用すれば、述語構造間の汎化された意味的関係を自動的に計算できることを新たに示した。この枠組みを利用することで、述語項構造を解析したテキスト同士の、簡単な言い換え関係を特定することが出来るため、近年盛んに研究されている言い換え・含意関係認識といった応用技術において、語彙概念構造による解析の枠組みが利用できる可能性を示したと言える。

２３年度には、２２年度に作成した日本語動詞に対する語彙概念構造辞書と、述語項構造解析の枠組みに基づいて、新聞記事データを主とする日本語コーパスに対して、対象とする述語表現に意味的に関わりのある句をほぼ網羅的に解析する述語項構造アノテーションを行った。本研究で対象とする意味タグは、各述語の語彙概念構造から直接的に説明される種類のもの（主要項）と、あらゆる述語に対して横断的に定義される種類のもの（周辺項）を分けることで、意味定義の面で見通しの良い設計を実現している。周辺項に関しては、現在までに特定の標準的なタグ

セットが存在しないため、実際のアノテーションに先立ち、コーパス中の文書を予め分析する事により、必要な周辺項の種類を定義した。実際のコーパスアノテーションを前提として、日本語の周辺項のタグセットを網羅的に設計するのは本研究が初の試みである。

述語項構造アノテーションに関しては、構築した概念構造辞書に含まれる各述語に対して、それぞれコーパスから抽出した１００事例に主要項と周辺項のタグ付けを行った。本研究で作成した述語項構造アノテーションは、既存の統語構造解析済みコーパスの上に、明確な意味機能をもった意味タグを用いてタグ付けされているため、自動的な述語項構造解析器の開発に不可欠な統語構造と意味構造の間の関係を紐解く資源として利用できる。また、コーパス作成過程において作成された述語の概念構造辞書は、言い換えや含意関係認識等の言語処理分野における応用技術のための有用な資源となるだけでなく、コーパス主導の方法で構築された現実的な根拠を有する資源として、言語学分野においても、動詞意味論等の進展の原動力として寄与することを期待できるものである。

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７件）

① Yuichiroh Matsubayashi, Yusuke Miyao and Akiko Aizawa. Framework of Semantic Role Assignment based on Extended Lexical Conceptual Structure: Comparison with VerbNet and FrameNet. The 13th conference of the European chapter of the Association for Computational Linguistics. 査読有. Aplil 2012. Avignon, France.

② 松林優一郎、宮尾祐介、相澤彰子、語彙概念構造を用いた日本語述語項構造コーパスの設計、第１８回言語処理学会年次大会、査読無、2012 年 3 月、広島

③ Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara and Koji Murakami. Safety Information Mining – What can NLP do in a disaster –. The 5th international joint conference on natural language processing. 査読有. November 2011. Chiang Mai, Thailand.

④ 松林優一郎、宮尾祐介、相澤彰子、日本語動詞概念構造コーパスの設計、NLP 若手の会 第 5 回シンポジウム、査読無、2011 年 6 月、東京都

⑤ Yuichiroh Matsubayashi, Yusuke Miyao and Akiko Aizawa. Building Japanese Predicate-argument Structure Corpus using Lexical Conceptual Structure. The eighth international conference on Language Resources and Evaluation. 査読有. May 2012. Istanbul, Turkey.

⑥ 松林優一郎、宮尾祐介、相澤彰子、語彙概念構造を用いた語彙間意味関係グラフの自動構築、人工知能学会第２５回全国大会、査読無、2011 年 5 月、岩手県

⑦ 松林優一郎、宮尾祐介、相澤彰子、語彙概念構造による意味役割の形式化と複数役割の割り当て、第１７回言語処理学会年次大会、査読無、2011 年 3 月、愛知県

６．研究組織

(1)研究代表者

松林　優一郎
（MATSUBAYASHI YUICHIROH）
国立情報学研究所・コンテンツ科学研究系・特任研究員
研究者番号：20582901

(2)研究分担者
なし

(3)連携研究者
なし