様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　5 月 31 日現在

研究種目：若手研究（スタートアップ）
研究期間：2008～2009
課題番号：20800029
研究課題名（和文）　類義述語句同定のための語彙的知識の体系化と集積

研究課題名（英文）　Research on systematization and collection of lexical knowledge for recognizing synonymous predicate phrases

研究代表者
松吉　俊（MATSUYOSHI SUGURU）
奈良先端科学技術大学院大学・情報科学研究科・特任助教
研究者番号：10512163

研究成果の概要（和文）：本研究の目的は、日本語動詞文や形容詞文が 2 つ与えられた時に、それらが類似した意味を持つかどうかを自動的に認識することである。本研究では、この自動認識に必要、かつ、それぞれの語に対して記述すべき知識を体系化し、集積した。主な研究成果は、抽象的なレベルで 16,157 対、実際に文に出現するレベルで 351,264 対の述語項構造間の関係知識と、モダリティ動詞 3,145 語とモダリティ形容詞 517 語を収録する辞書である。

研究成果の概要（英文）：I have constructed a system of lexical knowledge for recognizing synonymous predicate phrases, and compiled a knowledge base including the following two main components: (i) A list of semantic relations between two predicate-argument structures, whose size is 16,157 at abstract level and 351,264 at surface-form level; (ii) A dictionary of 3,145 modality verbs and 517 modality adjectives.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,330,000	399,000	1,729,000
2009 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	2,530,000	759,000	3,289,000

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：言語資源、言い換え、述語句、自然言語処理、言語学

１．研究開始当初の背景

日本語や英語など、人間が使う言語には、類似した意味を表す類義表現が数多く存在する。例えば、次の 2 つの文はほぼ同じことを述べており、互いにもう 1 つの文の類義表現となっている。

A)　この制度が変わるかもしれないそうだ。
B)　この制度が変更されるらしいとのことだ。

人間は、これらの文が同じ意味を表す類義表現であると容易に判断することができる。これは高度な推論機構によるものであり、現在の自然言語処理の技術では、計算機上で、任

意の２つの表現が類義表現であるかどうか判断することを十分な精度で実行することは困難である。

今日、情報通信網が高度に発達し、誰もがインターネットを通して大量のテキストにアクセスすることが可能になった。しかしながら、そこに存在する情報は冗長性が高いため、同じことを述べたテキストを複数の異なるウェブ文書において目にすることが少なくない。現在は、ウェブ閲覧者である人間が実際に多数のテキストの内容を確認し、これらテキスト間の類義性を判断している状況にある。限られた時間内で効率良く情報を得るためにも、類義表現を適切にまとめ、質を保ったまま情報の量を圧縮し簡潔化するシステムが望まれる。

このような要約システムの基礎となる技術は、冒頭に示したような２つの文が類義関係にあるかどうかを判断する類義表現同定技術である。自然言語処理において、この類義表現同定に関する研究は発展途上にあり、まだ十分な技術は確立されていない。日本語に関する類義表現同定技術の進展に最も重要であるのは、その基盤となる語彙的知識のデータベースの整備であると考える。なぜならば、例えば、英語においては、WordNet(名詞や動詞に対する同義語や上位語・下位語のデータベース)や FrameNet(動詞項構造に対するフレームや、2 つ以上の動詞項構造間の項の対応と関係に関するデータベース)など、大規模な語彙的知識のデータベースが存在するが、一方、日本語においてはこのような知識がほとんど利用可能ではないからである。

このような理由により、本研究では、日本語における類義表現を同定するために必要な語彙的知識を整理して体系化するとともに、それらの知識を集積し知識データベースを編纂する。ここでは、対象とする表現を、一つの事象(event)について述べる基本的な表現単位である述語句に定める。冒頭の２つの表現(A)、(B)は、述語句の典型例である。これまで、研究代表者は、「について」や「かもしれない」などに代表される日本語機能表現の体系化を行い、約 17,000 の機能表現の出現形を収録する辞書(知識データベース)を編纂した。この辞書は、任意の２つの機能表現(例えば、「そうだ」と「とのことだ」)が類義であるかどうかを判断するのに必要な情報を提供することができる。この辞書とともに、新たに述語項構造に関する語彙的知識を整備し、類義述語句同定技術の基礎を確立することを目指す。

２．研究の目的

本研究では、次の３点を中心に研究を行い、類義述語句同定のための知識データベースを編纂することを目指す。

(1) 類義述語句同定に必要な語彙的知識の体系化
　類義述語句同定に際して、あらかじめ語彙的知識として記述しておくべき情報の種類を明らかにする。

(2) 同定の基盤となる知識データベースの編纂
　上の体系に基づいて、類義述語句同定の基盤となる知識データベースの仕様を定める。この仕様に基づき、実際に大規模な語彙的知識のデータベースを構築する方法を確立する。

(3) 類義述語句同定システムの実現
　編纂した知識データベースを用いて類義述語句を同定するシステムを実装する。これにより、作成した語彙的知識の妥当性と有効性を検証する。

３．研究の方法

本研究では、２つの述語句間の類義性を、計算機を用いて自動的に判断する技術の基盤を確立することを目指す。これを実現するためには、述語句に関する大量の語彙的知識が必要であり、それゆえに、本研究では、語彙的知識の集積に重点を置く。

本研究の対象言語は日本語であり、対象表現は、一つの事象について述べる基本的な表現単位である述語句である。この述語句は、統語論の観点から、大きく次の４つの構成要素に分解することができる。

　述語、項構造、格要素の名詞、末尾の機能表現

各々の構成要素に対して、その類義表現同定に必要となる情報の種類を整理し、これらをまとめ、類義述語句の同定に必要となる語彙的知識を体系化する。

具体的には、前節で述べた３つのサブ目的に従って、以下のように研究を進めた。

(1) 類義述語句同定に必要な語彙的知識の体系化

① 現在利用可能な言語資源の調査
　現在利用可能な言語資源のみを用いて、類義述語句の同定がどれほど可能であるのかを調査した。この調査により、類義述語句同定に必要となる情報の種類を明らかにする。

② 類義述語句の同定に必要な語彙的知識の体系化
　述語句の構成要素ごとに、類義表現同定に必要な語彙的知識を整理し、その体系化

を行った。上の調査に基づき、意味の近さに基づく木構造を持つシソーラスや、補足情報を自由に記述することができるリスト形式の辞書などを選択することにより、妥当な知識体系を設計する。

(2) 同定の基盤となる知識データベースの編纂

① 述語

述語として動詞、形容詞、形容動詞を対象とし、既存のシソーラスの拡張、および、国語辞典からの関係知識の抽出を行った。「類義」と裏表の関係にある「反義」を考慮した既存のシソーラスを利用することにより、類義述語句同定システムが高精度で利用することができる知識の構築を目指す。言語学や日本語教育学の辞書や文献などを参考にし、人手により正確な語彙的知識を集積する。

② 項構造

通常の格交替現象や、「能動態から受動態へ」など、態が変化した時の格交替に関する知識を人手で記述した。

③ 末尾の機能表現

述語句において述語の後に続く語列(機能表現を含む)を適切に解析し解釈する「拡張モダリティ解析」のための言語資源を構築した。述語句の末尾に複数の機能表現が存在する場合においても、その部分の全体の意味を解析する拡張モダリティ解析というアイディアは、類義述語句同定においても非常に有用であることから、機能表現列を抽象化した体系である「拡張モダリティ体系」を独自に定義し、その解析に必要な言語資源を整備する。

(3) 類義述語句同定システムの実現

① 類義述語句同定システムの試作

編纂した知識データベースを用いて、ヒューリスティックスに基づいて類義述語句を同定するシステムを試作した。システム評価のための大規模なコーパスは利用可能でないため、システム動作の定性的な評価を行う。

② 知識データベースの改善

上の評価結果をフィードバックさせ、知識データベースを改善する作業を行った。知識データベース利用時に、単位の不一致の問題(例えば、「持っていく」という一語か、「持っ」「て」「いく」という複数語か)と表記のゆれの問題(例えば、「受け付ける」、「受けつける」、「受付ける」、「うけつける」など)が発生した。これらの問題を解決するため、国立国語研究所が配布している形態素解析用辞書 UniDic の単位と語彙素を用いて、知識データベース内のエントリーを標準化する。

４．研究成果

本研究の遂行により、類義述語句を同定するために必要となる語彙的知識に関する基盤体系と基礎データベース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、(2)の③と(3)の②で整備した知識データベースは、自然言語処理での実際の利用を考慮した知識表現体系を採用しており、類義述語句同定に役立つだけではなく、この種の知識の標準規格としての性格を有していると思われる。

昨年より、英語におけるように、日本語においてもいくつかの大規模な語彙的知識のデータベースが利用可能になった。これらに対して、本研究の知識データベースは、述語句内の構成要素を対象として、精度を重視して人手で構築した言語資源であるという特長を持つ。

今後の展望としては、知識の精度ではなく被覆率を重視して、機械学習手法を用いて、大量の文書群から知識を自動的に獲得することが考えられる。本研究の延長として、どのような手法を用いれば、どのような種類の知識が効率良く大量に獲得できるのかといったこと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) 類義述語句同定に必要な語彙的知識の体系化

① 現在利用可能な言語資源の調査

日本語 WordNet が利用可能となる前に調査を行い、分類語彙表等の大きなシソーラスにおいては、名詞や副詞に関する知識は有用であるが、その一方で、動詞間および形容詞間の類義・反義に関する知識は相対的に分類が粗いことが分かった。一方、機能表現に関しては、述語句の末尾に複数の機能表現が存在する場合、既存の言語資源のみに頼る手法では、後処理のための人手規則を整備しなければならないことが分かった。

② 類義述語句の同定に必要な語彙的知識の体系化

上の調査に基づき、次のような体系を定めた。

述語: 複数のシソーラスと補助的な関係知識リストからなる知識体系

項構造: 各述語に対して格交替のリスト

末尾の機能表現: 機能表現列を抽象化した、7 項目からなる拡張モダリティ体系

なお、格要素の名詞と述語に係る副詞に関しては、既存のシソーラスの体系と知識

を用いることにした。

(2) 同定の基盤となる知識データベースの編纂

① 述語

反義関係を考慮した既存のシソーラスの意味クラスを人手で拡張し、新たに語義単位で約1万の動詞と約3,500の形容詞を追加することにより、19,456 エントリーからなる述語シソーラスを構築した。

さらに、補助的な知識として、国語辞典の見出し語と、その語釈文に存在する述語との間に人手で関係を付与することにより、10,115 の類義関係知識、1,432 の反義関係知識、4,755 の上位関係知識を整備・集積した。

② 項構造

出現頻度の高い 1,376 の動詞(の語義)に対して、それがとりえる複数の項構造間の項の対応関係を人手で付与した。

③ 末尾の機能表現

拡張モダリティ解析に必要な言語資源として、機能語や複合辞のように述語句のモダリティに直接関与する動詞や形容詞についての知識を人手で整備した。具体的には、辞書情報として記述すべき項目について調査・検討し、辞書の仕様を定めた。そして、内省に基づき、動詞 3,145 語、形容詞・形容動詞 517 語に対して、下位の述語に関与する、態度、真偽判断、価値判断のモダリティに関する情報を人手で記述した。

(3) 類義述語句同定システムの実現

① 類義述語句同定システムの試作

構築した知識データベースを用いて類義述語句を同定するシステムの性能を定性的に評価した結果、述語に関して、単位の不一致の問題と表記のゆれの問題があることが分かった。

② 知識データベースの改善

単位の不一致の問題と表記のゆれの問題を解決するため、国立国語研究所が配布している形態素解析用辞書 UniDic の単位と語彙素を用いて、知識データベース内のエントリーを標準化し、関係知識の適用範囲を拡大させた。具体的には、すでに構築していた知識データベースから、出現頻度の高い動詞を前件に持つ約 19,000 の関係知識と、形容詞を前件に持つ約 2,000 の関係知識を抽出し、これらのエントリーを、プログラムと人手により、UniDic の単位で語彙素と対応付けた。最終的に、語彙素単位で 16,157 エントリー、実際に文に出現するレベルの単位である書字形基本形単位で 351,264 エントリーの関係知識を得た。

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）

〔雑誌論文〕（計 2 件）

① 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,” 電子情報通信学会情報・システムソサイエティ論文誌, vol.J93-D, no.6, 2010（掲載確定）, 査読有

② 乾 健太郎, 松吉 俊, “言語情報編集のための広義モダリティ解析に向けて,” JAPIO 2009 YEAR BOOK, pp.128-133, 2009, 査読無

〔学会発表〕（計 7 件）

① Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto, “Annotating Event Mentions in Text with Modality, Focus, and Source Information,” Proceedings of the seventh international conference on Language Resources and Evaluation (LREC), pp.1456-1463, 2010.5.20, Valletta Malta

② 江口 萌, 松吉 俊, 佐尾 ちとせ, 乾 健太郎, 松本 裕治, “日本語文章の事象に対する判断情報アノテーション,” 情報処理学会研究報告, 自然言語処理研究会, 2009-NL-193, no.5, pp.1-8, 2009.9.28, 京都

③ Suguru Matsuyoshi, Koji Murakami, Yuji Matsumoto, and Kentaro Inui, “A database of relations between predicate-argument structures for recognizing textual entailment and contradiction,” Proceedings of the Second International Symposium on Universal Communication (ISUC), pp.366-373, 2008.12.15, Osaka

④ 松吉 俊, 村上 浩司, 増田 祥子, 松本 裕治, 乾 健太郎, “事象間関係知識の整備と類似・対立認識への応用,” 情報処理学会研究報告, 自然言語処理研究会, 2008-NL-187, pp.15-22, 2008.9.24, 静岡

６．研究組織

(1)研究代表者

松吉 俊 (MATSUYOSHI SUGURU)

奈良先端科学技術大学院大学・情報科学研究科・特任助教

研究者番号：10512163