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　 ８ 日現在

機関番号：６２６１５
研究種目：挑戦的萌芽研究
研究期間：2012～2014
課題番号：２４６５２０８９

研究課題名（和文）含意関係コーパスの分析に基づく自然言語の統一的形式意味論の研究

研究課題名（英文）Research on unified formal semantics based on the analysis of textual entailment corpora

研究代表者

宮尾　祐介（MIYAO, Yusuke）

国立情報学研究所・コンテンツ科学研究系・准教授

研究者番号：００３４３０９６

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：本研究では、自然言語の意味を形式的に記述するための統一的枠組みについて研究を行った。自然言語では、論理的な意味関係、単語・フレーズ間の意味関係、時間・アスペクト、談話関係など、様々な意味的関係が複雑にからみあっている。しかし、これまでの言語学的研究は、それぞれの分野の特定の現象にフォーカスしたものがほとんどで、これらを統一的に記述し、推論を行う枠組みは存在しない。そこで、意味推論の実世界のデータとして含意関係コーパスを利用し、これを分析することで、統一的な形式的意味表示を構築した。また、これに基づき含意関係認識システムを構築し、高速かつ高精度な推論が可能であることを示した。

研究成果の概要（英文）：This research aimed to study a unified framework for a formal representation of natural language semantics. Natural language semantics involves logical relations, semantic relations among words and phrases, tense and aspect, and discourse relations, while they have complicated interactions with each other. However, linguistic studies mainly focused on a specific issue in each research field, and no framework for representing these features in a unified way. Therefore, we developed a unified formal representation of natural language semantics based on the analysis of textual entailment corpora. In addition, we developed a system for textual entailment recognition, and demonstrated that our framework allows for fast and high accuracy semantic inference.

研究分野：自然言語処理

キーワード：形式意味論　含意関係認識

１．研究開始当初の背景

自然言語が表す「意味」とは何か、どのように形式化するのか、という問題は言語研究において大きな目標であるが、未だ確立した解は無い。例えば、論理学を出発点とする形式意味論では論理結合子や量化子などの意味的振る舞いについて厳密な理論を提供するが、それぞれの単語が持つ意味的性質については何も言わない。一方、オントロジーによる概念の体系化は単語間の意味的関係の形式的記述方法を与えてくれるが、単語列がどのように文の意味を構成するかについては何も言わない。他にも時制・アスペクトの理論や、モダリティに関する研究など、言語のある意味的側面に関する理論・研究分野は世界中で多岐に渡る。しかし、それらはほぼ独立しており、意味の様々な側面を統一的に記述する枠組みは全く不明である。一方で、自然言語処理技術の急速な発展により、テキストから大規模な知識（オントロジーやスクリプトなど）を自動獲得する手法は確立しつつある。しかし、これらの大規模知識を意味解析に適用するための理論が存在しないため、現状ではこれらのリソースは個別のアプリケーションにおいて場当たり的に利用されているにすぎない。つまり、意味を記述するための統一理論が存在しないことが高度な自然言語処理のボトルネックとなっている。

２．研究の目的

本研究は、様々な意味論の成果を一つの枠組みに統合した統一的意味表現を構築することを目的とする。これは言語研究創始以来の古い問題ではあるが、現在のように研究分野が細分化して各論が独自の発展を遂げている状況では極めてチャレンジングな課題である。この問題に取り組むためには、形式論理、様相論理、生成語彙論、語彙概念理論、知識表現、イディオム、文法理論、その他意味論や語用論に関する言語理論など、多岐に渡る論理学・言語学について深い理解が必要であり、さらにこれらを統合するための枠組みとして形式言語理論や型理論の理解も必要である。残念ながら、現状ではほとんどの研究者はそれぞれの研究分野に閉じて研究活動を行っており、統一的に意味を記述する枠組みが提案される可能性はほとんどない。研究代表者は自然言語処理研究者としての立場からこのようなボトルネックを認識しており、また今までの研究経歴において動的論理、語彙概念理論、文法理論などについてある程度把握している。これらの理論を一つの形式的枠組みに載せることを出発点とし、様相表現や知識表現の理論を取り入れることで意味の統一的記述のための意味表現を構築する。この過程で、研究代表者が現在までに開発した含意関係コーパスを実例データとして利用し、実例データに頻出する現象を中心に研究対象とすることで、意味に関する膨大な問題に優先順位を付け、限られた研究期間内に実例データを近似的ながら広くカバーする意味表現を構築するというアプローチを採る。

ここで目標とする意味表現は、自然言語のテキストが直接的に表すもののみを対象とする。生成的な比喩や換喩、会話における含意や言語行為などは、本研究の成果の先にあるものと考え、研究対象としない。ただし、本研究の成果はこれらの研究に必要不可欠な前提であり、自然言語の意味に関する様々な研究に大きなインパクトを与えると期待される。

３．研究の方法

(1) 自然言語の意味に関する理論の調査

本研究の目標は現在乱立している様々な意味理論を統一的に記述することであり、まずそれらの意味理論について網羅的に調査することが必要である。現在の言語学においては意味に関する研究分野は多岐に渡りそれぞれ独立しているが、それらは自然言語の意味にまつわる諸問題をある程度カバーしていると考えられる。そこで、現在の関連研究を網羅的に調査し、統一的記述枠組みに組み入れるべき対象を決定することができる。形式論理（動的論理など）、語彙的意味論（オントロジー的知識の記述や含意・前提などの知識）、文法理論（統語構造と意味構造のインターフェース）、様相表現の理論、時間・アスペクトに関する理論、語用論等について調査を行う。

(2) 含意関係コーパスの分析

研究代表者はこれまでに含意関係コーパスの開発を行っており、これを意味的現象の分析のための実例データとして利用する。含意関係コーパスとは、２つのテキスト間の含意関係（論理的な含意関係ではなく、一方のテキストを人間が読んだ時、もう一方のテキストが真と言えるかどうかという関係）について人手で正解を付与した言語リソースである。

含意関係：あり
t1: スレイマン１世率いるオスマン帝国は絶頂期を迎えていた。
t2: オスマン帝国は、スレイマン１世の時代が最盛期であった。

含意関係：なし
t1: オスマン帝国南部であるアラブ圏ではスンナ派が中心を成していたが、イラク南部ではシーア派が多数存在していた。
t2: オスマン帝国の国教はシーア派のイスラーム教であった。

上図は、実際に開発した含意関係コーパスのサンプルである。含意関係コーパスは自然言語処理において含意関係認識技術のための学習・評価コーパスとして利用されているが、言語学的にはテキストの意味的同値性・差異を集めた実例データとして見ることが

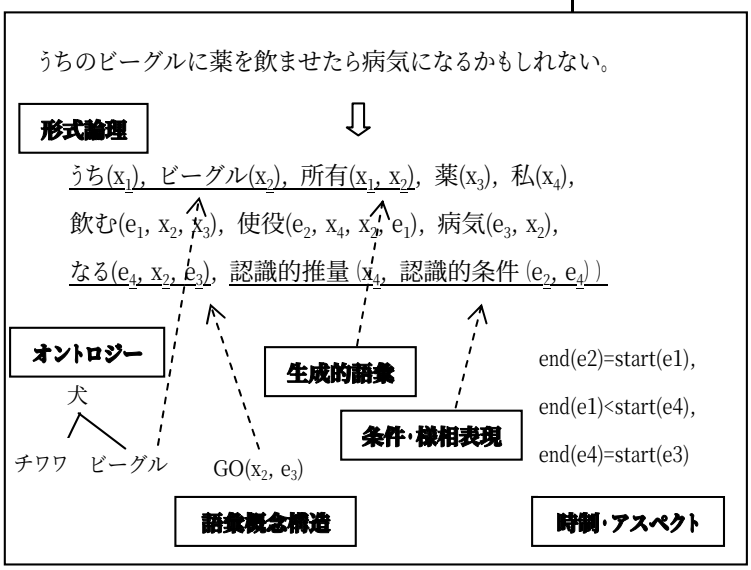
できる。例えば、この図においては、同義関係（絶頂期⇔最盛期）、含意関係（～率いる→～の時代、迎える→ある）、時制・アスペクト（迎えていた→あった）、格関係（～率いる→～が率いる、率いる～→～を率いる）、といった意味的現象が現れている。研究項目(3)、(4)は対象とする意味的現象が多岐に渡り膨大であるため、含意関係コーパスを分析し、そこに頻出する現象から優先度を付けて研究を行う。これにより、全ての意味的現象を説明するわけではないが、近似的に実世界テキストを広くカバーする意味表現を効率的に構築することができる。

(3) 意味理論の形式的な記述

意味に関する言語理論の大きな欠点として、形式化が不十分であることが挙げられる。統語論においては形式文法や型理論に基づく形式化がある程度進んでおり、形式的記述方法が確立されている。一方、意味論についてはそのような形式化があまりされていないのが現状である。例えば、様相表現に関する研究では、形式化された理論は今のところ様相論理を拡張した程度であり、論理的・数学的性質の研究は進んでいるが、実際の自然言語に現れる様々な様相（真偽判断、価値判断、発話などのモダリティ）についての形式的理論は存在しない。しかし、これらについては言語学において詳細な分析が行われており、言語学的知見は蓄積されているため、これらの言語理論を形式的枠組みで再記述することを目指す。

(4) 意味の統一的記述枠組みの研究

研究項目(1)、(3)に基づき、現在は独立に研究されている意味論を統一的に記述する形式表現について研究を行う。基本的には、形式論理の成果（動的論理など）をベースとして、オントロジー、語彙概念構造、生成的語彙、様相表現、時制・アスペクトなどの記述を追加して拡張する方針を想定する（下図参照）。



しかし、このような拡張を行うと元々の形式論理表現における推論規則は当然成り立たなくなり、推論規則についても拡張が必要と思われる。しかし、本研究では推論規則の厳密な定義や意味論の構築はあきらめ、研究項目(2)の分析に基づき実例データを広くカバーする現象について意味の同値性・差異を計算できるような推論規則を定めるにとどめる。従来の論理学・言語学における研究では、理論の厳密性を指向していたために幅広い現象を説明する理論の構築がほぼ不可能であったと考えられる。そこで、本研究では厳密性を犠牲にする代わりに幅広い意味的現象を説明する統一的枠組みを構築するというアプローチを採る。本研究項目で構築した意味表現は、含意関係コーパスの実際のテキストに適用し、含意関係が正しく予測できるかどうか評価を行う。

４．研究成果

研究項目(1)については、述語論理や動的論理などの形式意味論、オントロジー工学を中心とした語彙的意味論、モダリティに関する言語理論、時間・アスペクトに関する言語理論や言語処理研究、談話関係に関する理論・言語処理研究について調査を行った。形式意味論と語彙的意味論については形式モデルが提案されており、ほぼそのまま利用することができる。特に、述語論理、型論理、後述する集合間関係に基づく構成的意味論などの形式意味論の記述枠組みは、語彙的意味論を実装するのに十分な記述力があるため、これらは統合的に実装することが可能である。ただし、述語論理などにおける推論手法は計算量の問題が大きいため、計算量を抑えるための形式化や推論手法が必要である。一方、モダリティや時間・アスペクトに関する研究では、非形式的な言語学研究やリソース構築の研究が主である。目的、理由、手段、原因、結果などの談話関係については、形式理論は存在しないが、いくつかの関係については談話関係として研究が行われており、RST Treebank や Penn Discourse Treebank といったコーパスが開発されている。ただし、これらの研究では関係ラベルの種類を天下り的に決めており、それを支持する基準や相互作用については研究がされていない。例えば、理由と原因は似た性質を持っているため、単に関係ラベルを列挙するのではなく、これらの相互作用を説明できる枠組みが必要である。

研究項目(2)については、評価型ワークショップ NTCIR RITE タスクで提供されている含意関係コーパスを用いて、実テキストの含意

関係において現れる意味的関係の分析を行った。英語における既存研究を参考に、含意関係を基本文関係にブレイクダウンする手法を開発し、実際にアノテーション作業を行った。その結果、このデータにおいては上位・下位、同義、含意などのオントロジー的関係、特に単語間ではなくフレーズ間の関係が高頻度で現れることが明らかとなった。一方、モダリティや時間・アスペクトに関する意味的関係、および目的・理由などの談話関係は、出現頻度が低いことが明らかとなった。したがって、これらについては、形式的表現方法についての調査・検討（研究項目(3)）は続けるものの、含意関係認識システムへの実装（研究項目(4)）の優先順位を下げることとした。

研究項目(3)については、集合間関係に基づく構成的形式意味論に基づき、上述した各種意味論を形式的に記述する方法について調査・検討を行った。語彙的意味・オントロジー的関係については、前述のとおりほぼ自明である。一方、時間、アスペクト、モダリティの実装については、それぞれ集合間関係として記述することができるため、同じ枠組みで記述することが可能である。ただし、これらの間の相互作用については理論的に不明な点が多く、今後引き続き検討が必要である。談話関係については、言語学・自然言語処理において形式的記述に関する研究がほとんどなく、改めて理論を構築する必要がある。実テキストの分析に基づき談話関係を形式的に分類・記述する理論について検討を始めたが、形式意味論の記述枠組みに統合するには至っていない。これについては今後の課題とする。

研究項目(4)については、含意関係認識のための意味記述および推論手法を提案し、含意関係認識評価データにおいてその有効性を示した。本システムでは、形式論理に基づく一般的な推論（三段論法など）に加え、時間や一般化量化子に関する推論、さらに WordNet などの語彙的・オントロジー的知識や単語の分散意味表現に基づくパラフレーズ認識を統合した推論を実装している。モダリティやアスペクトに関する推論は未実装であるが、上記のとおり理論的には記述可能である。実験では、これまでに提案された表層的類似度に基づく手法や一階述語論理に基づく推論システムよりも高速かつ高精度を達成することが示された。一方、含意関係認識の精度は実用レベルからはいまだ遠いのが現状である。データやシステム出力の分析によると、含意関係認識が難しいケースの多くは、フレーズ間の同値関係・排他関係の認識が不完全であることによる。逆に言うと、論理推論や時間関係の推論が原因であることは少ない。したがって、より高精度な含意関係認識システムを実現するためには、パラフレーズ認識の高精度化が必須であるとの結論に至った。

５．主な発表論文等

〔雑誌論文〕（計　3 件）

1. Encoding Generalized Quantifiers in Dependency-based Compositional Semantics.
Yubing Dong, Ran Tian, Yusuke Miyao.
Proceedings of PACLIC 28. 2014 年. pp. 585-594.
査読有り

2. Logical Inference on Dependency-based Compositional Semantics.
Ran Tian, Takuya Matsuzaki, Yusuke Miyao.
Proceedings of ACL 2014. 2014 年. pp. 79-89.
査読有り

3. Building Japanese Textual Entailment Specialized Data Sets for Inference of Basic Sentence Relations.
Kimi Kaneko, Daisuke Bekki, and Yusuke Miyao.
Proceedings of ACL 2013. 2013 年. pp. 273-277.
査読有り

〔学会発表〕（計　2 件）

1. 基本文関係に分解した日本語含意関係認識アノテーション.
金子 貴美, 戸次 大介, 宮尾 祐介.
人工知能学会全国大会（第 27 回）. 2013 年 6 月 7 日. 富山国際会議場.

2. 基本文関係に分解した含意関係認識日本語評価データの構築.
金子貴美, 宮尾祐介, 戸次大介.
言語処理学会第 19 回年次大会. 2013 年 3 月 15 日. 名古屋大学.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計　1 件）
名称：自然言語推論システム、自然言語推論方法及びプログラム
発明者：宮尾祐介、田然
権利者：情報・システム研究機構
種類：特許
出願番号：特願 2013-108335
出願日：平成 25 年 5 月 22 日
国内外の別：国内

○取得状況（計　0 件）

６．研究組織

(1)研究代表者
宮尾　祐介　（MIYAO, Yusuke）
国立情報学研究所・コンテンツ科学研究系・准教授
研究者番号：00343096

(2)連携研究者
該当無し