様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月　9 日現在

研究種目：若手研究(B)
研究期間：2007 年度～2009 年度
課題番号：19700130
研究課題名（和文）
大規模な語彙意味構造辞書をコーパス主導で開発する研究
研究課題名（英文）
Corpus-driven development of a large-scale lexicon of lexical semantics structures
研究代表者
宮尾　祐介　　（MIYAO YUSUKE）
東京大学・大学院情報理工学系研究科・助教
研究者番号：00343096

研究成果の概要（和文）：
自然言語テキストの意味の構造を自動解析するために，単語が本来持っている意味（語彙的意味）の大規模辞書を開発した．このとき，一般的な辞書では，辞書に登録されていない単語（未知語）に対しては意味が与えられないという問題がある．そこで，本研究では単語と意味との対応関係を確率モデルで表現する手法を提案した．既存の辞書を学習データとし，大量のテキストから抽出した統計情報を特徴量とすることで，大規模辞書の確率モデルが学習できることを実証した．

研究成果の概要（英文）：
We aim at the automatic analysis of semantic structures of natural language texts, and developed a large-scale lexicon of lexical semantics, which represents meanings a word inherently owns. Conventional lexicons cannot assign semantic structures to unknown words (words that do not exist in a lexicon). Therefore, this research proposed a method for representing word-to-semantics mappings as a probabilistic model. We empirically demonstrated that we can obtain a probabilistic model of a large-scale lexicon by using an existing lexicon as training data and statistics extracted from large texts as features.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,100,000	0	1,100,000
2008 年度	1,200,000	360,000	1,560,000
2009 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	2,390,000	630,000	3,830,000

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　意味構造

１．研究開始当初の背景

構文・意味解析は，自然言語処理において古くから最重要課題であり，さかんに研究が行われている．最終的な目標は，実世界のテキストに対し，その「意味構造」を計算する

ことである．しかし，自然言語の複雑さ，さらに膨大な知識を記述する必要があることから，実世界テキストの意味構造を高被覆かつ高精度に計算できるシステムは未だに実現されていない．

現在までの意味解析の研究は，理論的研究にとどまるか，もしくは極めて単純化された「意味」を扱うものであった．前者の研究としては，形式意味論，談話理論，語彙意味論などが提唱され，一定の理論的成果を収めている．しかし，実世界のテキストに対してこれらの理論に基づく意味構造を計算するシステムは現在までに実現されておらず，既存のアプローチでの実装は事実上不可能と考えられる．一方，後者の研究としては，古くは WordNet があり，最近では意味的関係（述語項関係，照応関係など）をコーパスに注釈付け（アノテーション）する研究がさかんに行われ，様々なリソースが利用可能となっている．しかしながら，これらはあくまでも構造化されていない意味情報を単純にデータベース化したものであるため，そのままでは上記の理論に基づく意味構造を計算することはできない．また，これらのリソースを利用して格フレームの自動獲得，意味の曖昧性解消タスクなどがさかんに研究されてきた．しかし，これらは上述のリソースをそのまま統計的機械学習の学習・評価データとして利用するものがほとんどである．統計的機械学習に頼ったこれらの手法で扱う「意味」は，言語理論や複雑な構造を仮定せず，単語の共起の選好性や，予め与えられた意味ラベルの付与といった，単純化された意味情報を扱うものに限られている．

一方，研究代表者は，語彙化文法理論の一つである Head-Driven Phrase Structure Grammar (HPSG) に基づく構文解析器の研究に従事してきており，現在までに高被覆かつ高精度な英語構文解析器の開発に成功している．HPSG 構文解析器の開発においても，大規模な辞書の構築がボトルネックとなっていた．この問題に対して研究代表者は，コーパス指向の文法開発手法を提案した．この文法開発手法では，既存のコーパス(Penn Treebank)に HPSG 理論に基づく構文情報を注釈付けし，それに HPSG の文法規則を逆適用することで，大量の HPSG 語彙構造を獲得することに成功した．

意味構造辞書の構築では，HPSG 文法の開発と同様の困難さがある．したがって，構文解析器の開発で成功した方法論を発展させ，上に挙げた既存の意味情報リソースを利用することで，語彙意味論に基づく意味構造の大規模辞書を獲得することができると考えられる．

２．研究の目的

本研究課題は，語彙意味論に基づく意味構造の大規模辞書を構築し，実世界テキストの意味構造を高被覆かつ高精度で計算することを目的とする．語彙意味論は，単語の意味を語彙意味構造に構造化し，それらの相互作用を意味構造合成規則で記述することで，文の意味構造を計算する理論である．言語の多様性を語彙意味構造の多様性で表現することで，一般化された規則性を少数の合成規則で理論化する枠組みである．したがって，実世界テキストに対して意味構造を計算するためには，全ての単語に対してその語彙意味構造を記述しなければならない．しかしながら，このような大規模な意味構造辞書の構築は，人手による開発は事実上不可能であり，また単純な統計的手法で自動学習することも困難である．

そこで，コーパス（実世界のテキストに言語学的情報を注釈付けしたもの）に意味情報を注釈付けし，注釈付きコーパスから語彙意味構造辞書を獲得する研究を行う．即ち，意味構造辞書を直接人手で作成するのではなく，まず実際の文がどう解析されるべきかを注釈付けすることで，実際の文を理論に基づき説明するような語彙意味構造を獲得する．

本課題は，語彙意味構造辞書の開発，およびそれを用いた高被覆かつ高精度な意味解析に焦点を当てる．HPSG 構文解析器により構文的変形（受身や長距離依存関係など）によらない構造（述語項構造）を計算することは可能となったが，文の構造とは独立な単語間の意味的関係は計算することができない．語彙意味論は，単語間の意味的関係およびその文中における意味の合成を説明する枠組みであり，同じ意味を様々な語句の組み合わせで表現することができる自然言語テキストを自動解析するためには必要不可欠な研究対象である．本研究では，実世界テキストを解析できる意味構造辞書を構築することで語彙意味論の妥当性を明らかにし，また実世界テキストに対して実際に意味解析を行うことにより，このようなアプローチによる意味解析の有効性を明らかにする．また，研究過程では，コーパス主導のリソース開発をサポートするツールの開発を並行して行う．

３．研究の方法

(1)　意味情報リソース，関連研究の調査

意味情報が付与されたコーパス(PropBank, NomBank, FrameNet など)，データベース(WordNet, VerbNet, OpenCyc など)，その他のリソースについて調査を行う．基本的には，PropBank, NomBank, FrameNet は，構文構造

と述語項構造との関係を記述したリソースであり，一方，WordNet，VerbNet や OpenCyc は，文から切り離された単語の意味的関係のデータベースである．特に後者は，現在の HPSG 構文解析器が持たない語彙的意味情報を持つので，本研究で目指す意味構造辞書の構築に利用すべきリソースであると考えられる．さらに，近年意味情報リソースの開発がさかんに行われているため，最新のリソースの開発状況について調査を行う．また，意味表現に関する言語理論(フレーム理論，生成語彙論，語彙概念構造理論など)について最新の研究を調査し，既存のリソースや構文解析器がどのように利用可能か検討する．

語彙意味論に基づく意味構造辞書の構築のためには，既存の意味情報リソースでは不十分である可能性がある．そのようなリソースについては，本研究において構築を行うか，もしくは十分な規模のデータを作成するのが困難な場合は，機械学習等による自動獲得のための学習データを構築することを検討する．

(2)　語彙意味構造辞書の自動獲得の研究

上述のリソースを用いて意味情報をコーパス(Penn Treebank)に注釈付けし，その注釈情報から語彙意味論に基づく意味構造辞書を獲得するアルゴリズムを開発，実装する．どのような情報をコーパスに付与すべきかは語彙意味論からの要請である程度決定される．しかしながら，HPSG 文法開発における経験から，実際には理論が想定していなかった現象が発見されると考えられる．コーパス主導のリソース開発においてはそのような現象が実例ベースで検出できるので，実験結果の分析を通して注釈付けの追加・改良をしていく方針を採る．

(3)　コーパス主導のリソース開発を支援するツールの開発

HPSG 構文解析器を開発した際，コーパス指向文法開発を支援するツールキットも同時に開発を行った．しかし，それらは機能・速度面で改良の余地がある．また，高被覆な意味解析のためには様々な分野のテキストを解析できる構文解析器が必要である．そこで，文法開発ツールキットを拡張し，注釈付きコーパスを効率的に構築するためのツールを開発する．HPSG 構文解析器を開発した際のノウハウにより，ツールキットに要求される機能や基本アイディアは既にある程度検討済みである．したがって，現在のツールキットを発展させることで，より一般的な言語リソース開発ツールが開発できると期待される．

４．研究成果

(1) 意味情報が付与されたコーパスやシソーラス，およびこれらを利用した最近の研究について調査を行った．PropBank，NomBank，FrameNet などは意味役割認識の研究が広く行われており，2007 年の CoNLL shared task でも採用されている．一方，WordNet などのシソーラスを用いた研究として，シソーラスの自動構築や語義曖昧性解消が行われている．これらの研究は，これらのリソースを機械学習の学習データとして直接的に利用する研究である．

英語においては，動詞の語彙意味構造クラスの辞書として VerbNet が開発されており，それと FrameNet，PropBank，WordNet の相互関係を付与した SemLink の開発が行われている．両リソースとも対象のテキスト（主に新聞記事）に対しては被覆率が高く大規模なリソースである．VerbNet は意味クラスごとに語彙的意味テンプレートを記述しており，さらに，それを利用して語彙概念構造辞書を半自動構築する研究が存在する．したがって，本研究では VerbNet および SemLink を用いて意味構造辞書獲得の実験を行う方針とした．

(2) 従来研究より意味構造に近いレベルでの構文解析器の評価を行い，HPSG に基づく構文解析器が高精度を達成することを示した．これは，HPSG に基づく意味構造に対して語彙意味構造を統合することの有効性を示していると考えられる．また，WordNet の意味クラスを HPSG 構文解析器に統合する実験を行った．意味クラスを曖昧性解消の特徴量として利用する実験を行ったところ，構文解析の精度が有意に向上することを確認した．これは，構文解析と意味解析を統合して行うことの優位性を示している．

また，異なるフレームワークに基づく構文解析器を横断的に比較するために，タスクに基づく構文解析比較手法を提案した．具体的には，構文解析の出力を様々なフォーマットに変換し，PPI 抽出器の入力として与え，PPI 抽出の精度を比較する．これにより，各構文解析器が後段の自然言語処理タスクに寄与する度合いを比較することができる．実験により，述語項構造やそれを近似する依存構造を出力する構文解析器が，PPI 抽出への寄与度が高いことが示された．すなわち，意味解析がこれらの実用タスクに対して有効に働くことを強く示唆している．

(3) 成果(1)に基づき，実テキストに対して動詞の語彙意味クラスを自動認識する技術について調査・研究を行った．SemLink を利用して Penn Treebank に語彙意味クラスを注

釈付けし，このコーパスを学習データとして VerbNet 意味クラスを自動認識する実験が既に行われており，高い精度（90%以上）を達成することが報告されている．この研究の追試を行ったところ，学習コーパスに存在する語（既知語）に対しては高精度を達成するが，学習コーパスに存在しない語（未知語）や，学習コーパスと異なる分野のテキストに対しては著しく精度が低下することが観察された．この結果については，生命科学テキストに対して同様の実験を行った関連研究においても同様の観察がされている．したがって，一つの分野のテキストにおいては語とその意味クラスとの対応関係が固定的であり，意味クラスの高精度な自動認識のためには高被覆な辞書を構築することが本質的であると考えられる．これらの調査結果から，未知語や異なる分野のテキストに対して頑健な意味構造辞書を獲得することが本質的な問題であることが分かった．

(4) VerbNet 動詞意味クラスの確率的辞書を獲得する手法を提案した．確率的辞書とは，未知語も含めて，単語とそれが取りうる意味クラスとの対応関係を確率分布で表現したものである．提案手法では，VerbNet を学習データとし，生コーパスから抽出した統計量（下位範疇化フレーム，項となる語，など）を特徴量として用いる．これにより，未知語（VerbNet には存在しないが生コーパスには存在する語）についての確率が推定されるため，生コーパスに存在する任意の動詞について意味クラスの高精度な認識が可能となる．VerbNet を用いた実験において，本手法により大規模な意味クラス辞書の確率分布を学習することができることを示した．

(5) コーパス主導のリソース開発のためのツールの開発

HPSG 構文解析器を用いて，新たなテキストに対して効率的に注釈付きコーパスを開発するツールを開発した．構文解析結果をグラフィカルに表示し，それをインタラクティブに修正することで，効率的な注釈付け作業を可能にする．これにより，様々な分野のテキストに対して頑健な構文解析器が効率的に開発できると期待される．

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）

〔雑誌論文〕（計６件）

① Yusuke Miyao and Jun'ichi Tsujii. Supervised learning of a probabilistic lexicon of verb semantic classes. Proceedings of EMNLP 2009. 査読有. 2009. 1328-1337.

② Kenji Sagae, Yusuke Miyao, Takuya Matsuzaki, Jun'ichi Tsujii. Challenges in Mapping of Syntactic Representations for Framework-Independent Parser Evaluation. Proceedings of Workshop on Syntactic Annotations for Interoperable Language Resources. 査読有. 2008. 63-66.

③ Jun Hatori, Yusuke Miyao, and Jun'ichi Tsujii. Word Sense Disambiguation for All Words using Tree-Structured Conditional Random Fields. Proceedings of COLING 2008. 査読有. 2008. 43-46.

④ Yusuke Miyao, Kenji Sagae and Jun'ichi Tsujii. Towards Framework-Independent Evaluation of Deep Linguistic Parsers. Proceedings of GEAF 2007. 査読有. 2007. 238-258.

⑤ Kenji Sagae, Yusuke Miyao and Jun'ichi Tsujii. HPSG parsing with shallow dependency constraints. Proceedings of ACL 2007. 査読有. 2007. 624-631.

⑥ Yusuke Miyao, Rune Sætre, Kenji Sagae, Takuya Matsuzaki and Jun'ichi Tsujii. Task-Oriented Evaluation of Syntactic Parsers and Their Representations. Proceedings of ACL-08:HLT. 査読有. 2008. 46-54.

〔学会発表〕（計１件）

① Alastair Butler, Yusuke Miyao, Kei Yoshimoto, Jun'ichi Tsujii. A Constrained Semantics for Parsed English Sentences. 言語処理学会第 16 回年次大会発表論文集. 査読無.

６．研究組織

(1)研究代表者

宮尾　祐介　　(MIYAO YUSUKE)

東京大学・大学院情報理工学系研究科・助教

研究者番号：00343096

(2)研究分担者

(3)連携研究者