「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
平成 25 年度採択研究代表者

H29 年度
実績報告書

黒橋　禎夫

京都大学大学院情報学研究科
教授

# 知識に基づく構造的言語処理の確立と知識インフラの構築

## §1. 研究実施体制

(1)黒橋グループ
① 研究代表者:黒橋　禎夫 （京都大学大学院情報学研究科　教授）
② 研究項目
・クラウドソーシングによる同義表現・基本事態対の作成
・知識に基づく統語的文章解析モデルの構築と高度化
・行政サービス対話ボットのための対話フローチャートの自動生成(乾ｸﾞﾙｰﾌﾟと共同)

(2)戸次グループ
① 主たる共同研究者:戸次　大介 （お茶の水女子大学大学院人間文化創成科学研究科　准教授）
② 研究項目
・日本語 CCG パーザへの依存型意味論の実装
・依存型意味論に基づく注釈付与コーパスの作成
・依存型意味論に基づく意味計算システムの構築

(3)乾グループ
① 主たる共同研究者:乾　健太郎 （東北大学大学院情報科学研究科　教授）
② 研究項目
・言明間関係解析のための意味表現の検討
・知識の関係付けを実現する知識推論機構の構築
・企業コンタクトセンター等への適用(黒橋ｸﾞﾙｰﾌﾟと共同)

## §2. 研究実施の概要

本研究の目標は、人間の知識表現の根幹である言語の計算機処理を進化させ、知識に基づく頑健で高精度な構造的言語処理を実現し、これによって様々なテキストの横断的な関連付け、検索、比較を可能とする知識インフラを構築することである。プロジェクトの第五年度となる今年度は各研究項目について以下の成果を得た。

**文の意味の表現・計算モデルの構築: 戸次グループ(お茶の水女子大学・NII)**

戸次グループは文の意味の表現・計算モデルの構築を担当している。平成 26 年度に提唱した自然言語の意味の理論「依存型意味論(DTS)」と、平成 27 年度に完成させた自然言語推論システム「ccg2lambda」を両輪として、自然言語処理、理論言語学の最先端研究を目指し、理論および処理系の開発を進めている。平成 29 年度は、まず CCG 木のための新たな確率モデルと深層学習を組み合わせた CCG 構文解析器 depccg を開発し、英語・日本語において最高精度を達成した。また、日本語含意関係認識のためのテストセット JSeM に、より多様な機能語の意味把握をテストしうるデータを追加した。さらに、ccg2lambda が対象とするタスクを、これまで含意の有無という離散的な分類問題に限られていたところを、文類似度のような連続的回帰問題に拡張する研究を行った[1]。総じて、証明論的意味論の枠組みに基づく意味処理が、情報量の豊かさと柔軟性において有望であるという知見が得られた。

**知識に基づく文脈解析の実現と因果関係知識の抽出: 黒橋グループ(京都大学)**

黒橋グループは本研究の中心的課題として、知識に基づく文章解析モデルの構築を担当している。テキスト解析の基本となる構文解析(中国語)において、単語分割・品詞付与と依存構造解析を遷移型統語解析において同時に行うモデルを考案し、世界最高精度を達成した[2](ACL2017 Outstanding Paper Award 受賞)。さらに、テキストから因果関係等の個別の知識を正確に抽出するための必須要素である省略解析において、省略解析と共参照解析を統合し、文章中のエンティティの意味表現を両解析の進行とともに動的に更新することにより、従来極めて困難であった文をまたぐ省略解析の精度を飛躍的に向上させた。また、敵対的学習の考え方により省略解析においてタグ等の付与されていない生コーパスを用いて解析精度が改善できることを示した(いずれも ACL2018 で発表予定)。さらに、株式会社 Insight Tech、日立アプライアンス株式会社、LINE 株式会社、兵庫県等との共同研究を推進した。

**テキスト横断的な知識の関係付けによる知識インフラの構築: 乾グループ(東北大学)**

乾グループは最上位のレイヤーで、テキスト横断的な知識の関係付けによる知識インフラの構築を担当する。知識を記述した言明どうしの論理関係を計算するためには、言語表現間の意味の類似性を柔軟に計算する仕組みが必要である。本年度は、語の意味の分散表現学習に関する前年度までの成果をさらに発展させ、種々のメタ情報を用いて意味の異なる側面を捉える分散表現や、感情極性や対義性を取り入れた分散表現など、単語の意味をより精緻に捉える分散表現の学習方法の研究を進めた。また、知識どうしを繋げる推論機構の研究の例題として知識ベース補完タ

スクを、世界知識をテキストの解析に活用する仕組みの研究の例題として意見分析タスクを取り上げ、知識獲得から知識適用までのプロセスを統合する研究に取り組んだ。知識ベース補完では標準的なベンチマークデータにおいて世界最高性能を達成し（国際会議ACL2018で発表予定）、意見分析タスクでは行列因子分解やFactorization Machinesによって同領域における世界知識の獲得とテキストの意見解析を統合する新しい枠組みを構築した（国際会議 ACL2017 で発表、COLING2018で発表予定）。

代表的原著論文

[1] Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, Daisuke Bekki. Determining Semantic Textual Similarity using Natural Deduction Proofs, In Proceedings of Empirical Methods in Natural Language Processing (EMNLP2017), pp.681-691, September 7-11, Copenhagen, Denmark, 2017.

[2] Shuhei Kurita, Daisuke Kawahara and Sadao Kurohashi. Neural Joint Model for Transition-based Chinese Syntactic Analysis. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017), pp.1204-1214, Vancouver, Canada, 2017.

[3] Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki and Kentaro Inui. Other Topics You May Also Agree or Disagree: Modeling Inter-Topic Preferences using Tweets and Matrix Factorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017), pp.398-408, 2017.