

「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

H28 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

(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
 - ・ 言明・知識間の論理関係の設計と評価用コーパスの作成
 - ・ 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(黒橋グループと共同)

§ 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。プロジェクトの第四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「cgg2lambda」を両輪として、自然言語処理、理論言語学、言語哲学のいずれの分野においても最先端であるような、理論および処理系の開発を進めている。DTS の特徴は「証明論的意味論」である点にあり、現代的な論理学の成果を用いて推論の効率的な計算を可能とするが、近年の機械学習の技法に対しても有効な「深い意味」の情報を提供する。本年度の主な成果としては、「cgg2lambda の日本語への適用」[1]が挙げられる。

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。テキストから因果関係等の個別の知識を正確に抽出するには省略解析の高精度化が必須である。大規模コーパスから学習した選択選好等の様々な手がかりをニューラルネットワークによって柔軟に融合する手法を考案し、日本語同一文内の省略解析精度を従来手法の F 値 42.1 から F 値 53.4 に大幅に向上させた[2]。さらに、「洗濯機を回す」と「洗濯機を使う」のような項を共有する同義表現、「X が破綻する→Y が X を買収する」のような基本事態対の知識獲得についても研究を進めた。また、あるトピックに関連する出来事を時間軸にそって整理するタイムライン構築について、テキスト中の出来事表現に時間タグを付与したコーパスを構築するとともに、二段階の処理によって大域的情報を活用するタイムライン処理手法を考案し英語の標準的テストセットにおいて世界最高精度を達成した。

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。本グループでは、大規模言語データから語の意味の分散表現(ベクトル表現)を教師なしで学習する機構の研究を展開し、ベクトルの加法による意味合成の一般理論を完成させるとともに、これを論理的意味操作が扱える表現に拡張した[3]。また、Wikipedia と一般テキストを知識の対象物(エンティティ)に基づいて統合的に構造化する知識インフラの構築に着手し、日本語では初めてのエンティティ・リンキング・コーパスを構築するとともに、解析モデルの学習が可能なることを示した。さらに、テキストから多様なエンティティの特性や因果関係等の知識を抽出することを目的として、shift-reduce 型構文解析法と単語分散表現を融

合させたセマンティック・パーザの研究を進め、特定の領域での効果を実験的に示した。

[1] Koji Mineshima, Ribeka Tanaka, Pascual Martinez-Gomez, Yusuke Miyao, and Daisuke Bekki. Building compositional semantics and higher-order inference system for a wide-coverage Japanese CCG parser. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP2016), pp. 2236-2242, 2016.

[2] Tomohide Shibata, Daisuke Kawahara and Sadao Kurohashi. Neural Network-Based Model for Japanese Predicate Argument Structure Analysis. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016), pp.1235-1244, Berlin, Germany, 2016.

[3] Ran Tian, Naoaki Okazaki, Kentaro Inui. The Mechanism of Additive Composition. Machine Learning, pp.1-48, 2017.