「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

H27 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

- Wikipedia 文脈注釈付与コーパス作成
- クラウドソーシングによる同義表現・基本事態対の作成
- 知識に基づく省略・談話構造解析モデルの構築

(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・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。プロジェクトの第三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、他グループに対して知識情報処理の基盤技術を提供する。大きな成果として形式意味論に基づく高階論理表示の意味合成と推論を行うシステムを構築した。この本研究では、自然言語の複数の文に対して、頑健な CCG パーザの出力から形式意味論に基づいて高階論理表示の意味合成を行い、定理証明器 Coq による自動推論を行う推論システムを構築した。検証は FraCaS テストセットによって行い、一階述語論理に基づく最先端の推論システム(Nutcracker)の正解率 50%に対して、69%の正解率を達成した。また、これまで高階論理による推論は効率的ではないと考えられてきたが、本システムでは Natcracker の速度 11.23 秒/問に対して速度 3.72 秒/問を達成した。本成果は言語処理分野のトップカンファレンスである EMNLP2015 に論文採択された[1]。

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する。文章解析の基盤となる単語を正確に認識するため、大規模語彙情報と意味的汎化言語モデルに基づく新たな形態素解析手法を考案した。高次の文脈解析を行うためには、従来の形態素解析の98%前後の単語・品詞認識は十分な精度とはいえない。本研究では、Wikipedia、Wiktionary 等から大規模な語彙獲得を行うことで辞書を拡充し、さらに、意味の汎化に基づく言語モデル、RNNLM (Recurrent Neural Network Language Model) を導入することにより、形態素解析の大幅な精度向上を達成し、構文解析・文脈解析に悪影響を与える誤りを 70%削減することに成功した。本成果についても EMNLP2015 で論文発表を行った[2]。

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。言明間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。本研究では、大規模言語データから教師なし学習で得られる語の意味の分散表現(ベクトル表現)をベースとして、そこから句の意味の分散表現を構成的に計算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。また、cause(X, Y)のような関係知識の獲得に適した意味表現の学習にも取り組んだ。句の意味的類似性を推定するタスクや関係抽出タスクなど、複数のベンチマークデータで既存の手法を越える最高性能を実現している。成果の一部は、国際ジャーナル[3]に採択された他、ACL2016 でも発表する予定である。

[1] K. Mineshima, P. Martinez-Gomez, Y. Miyao and D. Bekki, Higher-order Logical Inference with Compositional Semantics, Proceedings of EMNLP2015: Conference on Empirical Methods in Natural Language Processing, Lisboa, Portugal, pp.2055-2061, 2015.
[2] H. Morita, D. Kawahara and S. Kurohashi. Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, Proceedings of EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Lisboa, Portugal, pp.2292-2297, 2015.
[3] Sho Takase, Naoaki Okazaki, Kentaro Inui. Modeling Semantic Compositionality of Relational Patterns. Engineering Applications of Artificial Intelligence, vol. 50, pp.256-264, 2016.