「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

H26 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の作成
- 言語解析のための基盤知識の自動獲得と知識に基づく統合的文章解析モデルの構築
- 事態間知識の構築

(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
- 言明間関係解析のための意味表現の検討
- 言明間の論理的関係を解析する計算モデルの構築

§2. 研究実施の概要

テキストは、専門家によるデータの分析結果や解釈、ステークホルダーの批判・意見、種々の手続きやノウハウなどが表出されたものであり、人間の知識表現の根幹をなすものである。言語の計算機処理はウェブをはじめとする大規模テキストの活用によって長足の進歩を遂げつつあるが、本研究ではこれをさらに発展させ、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。また、構築した注釈付与コーパス、辞書、言語解析システムの公開によって研究コミュニティによる一層の研究の加速を実現するとともに、これらの研究成果を企業のカスタマセンター業務等の社会の実問題に適用し、その有用性を評価する。今年度は本研究プロジェクトの第二年度として、各研究項目について以下の研究を実施した。

文の意味の表現・計算モデルの構築

新たな意味の表現・計算モデルの構築を目指して、「組合せ範疇文法(CCG)に基づく日本語パーザ＋依存型意味論(DTS)の証明システムによる計算」という枠組みで研究を推進した。日本語 CCG パーザのためのリソース（主にツリーバンク）の構築・改良、構文解析結果のエラー分析に基づく曖昧性解消モデルの改良により、前年度は 80%程度だった係り受け解析精度を 89%程度まで向上させた。また、DTS に基づく注釈付与コーパスの作成、DTS の理論的・計算論的基盤の確立、さらに、より広範囲の意味・推論現象を扱うことを目指して日本語の敬語表現、モダリティを伴う照応・前提への DTS の拡張を行った。

知識に基づく文脈解析の実現と因果関係知識の抽出

省略解析、談話解析などの文脈解析を高精度化し、これによってテキストの各論述から直接的に言明とその間の因果関係等の知識を抽出することを目標として研究を進めた. 現状の文脈解析の困難さの要因が、より低次の形態素解析、構文・格解析の誤りの蓄積に起因するという分析に基づき、Wikipedia 等のウェブ上のリソースの活用により語彙データベースを 232 万語に拡張し、未知語による形態素解析誤りを大幅に削減した. また、構文・格解析での重要な知識源である格フレームについて、動詞の用法のクラスタリングを CRP (Chinese Restaurant Process)に基づく方法に改良し(B-11)、加えて、同様のプロセスで動詞意味クラスを学習することに成功した(B-7). さらに、格フレーム学習においてより広い文脈を見る方法や、分散意味表現に基づく方法の検討をすすめ、さらなる改良の手がかりを得た。

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

テキストの各論述から抽出される知識は、様々な抽象度・粒度の知識がばらばらに混在したものであるため、これらの知識を相互に関連づける推論機構が必要となる。この研究を進めるための基礎データとして、医療/健康ドメインを中心とした約 7000 文対に 6 種類の論理関係ラベル、19 種類の理由ラベルを付与した大規模注釈付与コーパスを作成した。さらに、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ことから、語の意味の分散表現（ベクトル表現）の構成性の研究に着手し、加法構成性の原理に理論的枠組みを与え、短いフレーズの意味が構成できる条件を数理的・実験的に解明した。