戦略的創造研究推進事業　CREST
研究領域「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」
研究課題「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」

研究終了報告書

研究期間　2013年10月～2020年3月

研究代表者：黒橋　禎夫
（京都大学大学院情報学研究科、
教授）

§1 研究実施の概要

(1)実施概要

言語、テキストは情報・知識の表現と利活用のための根幹的メディアである。その高度な計算機処理が実現できれば、人間の知的活動を強化し、社会におけるコミュニケーションと知識循環を円滑化し、ひいては異なる分野間での知識の相互関連性の発見や新しい知識・法則の発見を支援することが可能となる。しかし、言語は社会、世界の複雑さを十分に表現しうることの裏返しとして、曖昧性を持ち、例外に満ち、複雑多様である。そのため、従来の IT システムでは、言語で表現される意味内容を Bag of words を始めとしてごくごく近似的に扱うに留まっていた。

本プロジェクトは、言語表現を形式意味論と論理の視点から捉える研究を進めてきた戸次グループ(お茶の水女子大学)、実テキストの頑健な構造解析を進めてきた黒橋グループ(京都大学)、言語処理に立脚して知識と推論の研究を進めてきた乾グループ(東北大学)が協力することにより、言語表現から知識、基礎理論から実システムまでを包括的に捉え、言語の計算機処理を大幅に進化させ、計算機がテキスト情報を知識として活用できる基盤(知識インフラ)の構築を目指して研究を推進した。また、本プロジェクト構想時には想定できなかったスピードと精度で進展した言語処理分野におけるニューラルネットワークと分散意味表現の利用についても、当初計画を適宜修正して積極的に研究項目として取り入れた。

その結果、戸次グループでは、組合せ範疇文法に基づく統語解析、形式意味論に基づく意味解析、高階論理に基づく自動証明を行う統合的システムを構築し、高階論理によって複雑な言語現象が関与する推論を効率的に行うことが可能であることを示し、論理と形式意味論に基づく深い言語解析に新たな道を切り開いた。黒橋グループでは、ニューラルネットワークに基づく形態素・構文・格・省略・共参照・談話解析の大幅な精度向上を実現し、述語項構造をベースとしてテキスト中で表現される事象や行為とその因果関係等をグラフ表現に変換する言語解析リソース群を整備することにより、言語処理アプリケーションにおいてテキストの深い意味情報を活用することを現実化した。乾グループでは、単語ベクトルから句の分散表現を構成する加法構成の一般理論を構築するとともに、構成的分散表現空間の学習に基づき、ユーザの質問に対して知識ベース上の知識と大規模テキスト集合に散在する記述をシームレスに組合せながら総合的に答えを推論する大規模知識インフラの構築を実現した。

さらに、これらの基盤技術を社会の実問題に適用し、生活者の意見集約サービスや、行政への問合せに対する対話応答システム等においてその有効性を示した。また、研究コミュニティによる一層の研究の加速を実現するために、本プロジェクトで開発した注釈付与コーパス、知識ベース、解析システム等の言語リソースを公開した。

一連の成果は、ACL、NAACL、EMNLP 等の当該分野のトップ国際会議において多数の発表を行い、国際会議での Outstanding Paper Award、科学技術分野の文部科学大臣表彰、言語処理学会論文賞等の多数の受賞を得た。

さらに、高度な意味処理の利用を普及させ、イノベーション創出に寄与することを目的として研究期間を 1 年間延長し、含意関係認識統合システム ccg2lambda のツールキット化と公開、因果関係グラフツールの整備、DCS-Vec に基づく知識統合環境の日本語化とソフトウェアパッケージの整備・公開を行った。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. 範疇文法と高階論理に基づく意味処理の理論と実装

概要:自然言語文に対して、組合せ範疇文法(CCG)に基づく統語解析、形式意味論に基づく意味解析、高階論理に基づく自動証明を行う統合的システムを構築し、含意関係認

識タスク・文類似度判定タスクにおいて世界最高水準の精度を達成した。一連の研究成果は EMNLP2015、ACL2016、EMNLP2016、EACL2017、ACL2017、EMNLP2017、NAACL-HLT2018、AAAI2019、ACL2019 と連続して当該分野のトップ会議に採択された。この研究は高階論理によって複雑な言語現象が関与する推論を効率的に行うことが可能であることを示すものであり、論理と形式意味論に基づく深い言語解析に新たな道を切り開いた。

2. 言語の構成的意味分散表現の学習に関する理論的・実証的研究
概要:語句の意味の類似性を計算する仕組みとして、単語ベクトルから句の分散表現を構成する加法構成の一般理論を構築し、加法構成の成立条件と予測誤差を世界で初めて数理的・実験的に解明した。また、この成果をもとに依存構成意味論 DCS の分散表現モデルを設計し、3種類の意味計算課題(句の類似度推定、文からの関係抽出、文の穴埋め)において世界最高精度を達成した。一連の成果は、機械学習のトップ国際雑誌 Machine Learning、NLP トップ国際会議 ACL2016(2 件)、NAACL2016、ACL2018 等に採択された。

3. ニューラルネットワークを利用した言語処理基盤の構築
概要:RNN 言語モデルを利用した日本語形態素解析、エンティティ分散表現の動的更新に基づく日本語省略・共参照統合解析、中国語単語区切り・品詞同定・依存構造統合解析等のニューラルネットワークを利用した言語処理基盤構築において先駆的な研究を行い、いずれも世界最高性能の解析精度を実現し、その成果を EMNLP2015、ACL2016、ACL2017、ACL2018 等のトップ国際会議で発表した(ACL2017 の論文は Outstanding Paper Award 受賞)。

＜科学技術イノベーションに大きく寄与する成果＞

1. 日本語言語処理リソースの整備・公開
概要: ウェブ上の多様な文書、15,000 文に対して述語項構造から談話関係までを付与した注釈コーパスの構築・公開、ニューラルネットワークに基づく形態素・構文・格・省略・共参照・談話解析システムの構築・公開を行った。さらに、テキストの意味を扱う標準単位として述語項構造を基本とする「イベント」を定義し、前述の言語解析結果をイベントを単位とするグラフ構造に変換するツールを構築・公開した。これらの言語解析リソース群によって、言語処理アプリケーションにおける意味情報の利用を大幅に簡単化した。

2. 知識獲得と知識推論を融合した知識インフラの実現
概要:エンティティ間の関係の集合からなる知識ベース全体を構成的意味分散表現空間に埋め込むことによって、知識補完と呼ばれる推論を世界最高精度で実現した(ACL2018 で発表)。また、この知識ベース埋め込みとテキストからの知識獲得を同じ構成的分散表現空間の学習として統合することにより、ユーザからの質問に対して、知識ベース上の知識だけでなく、大規模テキスト集合に散在する記述も手がかりとしてシームレスに組合せながら総合的に答えを推論できる大規模知識インフラの構築が初めて可能になった。

3. 構造的言語処理の実社会課題への適用
概要:本プロジェクトで開発した種々の技術を社会の実問題に適用した。Insight Tech 社の生活者の意見集約サービス「不満買取センター」においては、投稿された不満意見の集約に本言語解析基盤が活用されている。また、兵庫県丹波市・尼崎市および NII、LINE との協力により、行政ホームページ上の FAQ を基本知識源として行政サービスの問合せに回答する対話ボットを構築し、2018 年 6 月からサービスを開始した (2019 年 5 月現在、丹波市、尼崎市それぞれの利用ユーザ数 450 名、1200 名)。

＜代表的な論文＞

[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] Ran Tian, Naoaki Okazaki, Kentaro Inui. The Mechanism of Additive Composition. Machine Learning, Volume 106, Issue 7, pp. 1083-1130, 2017.7.

[3] Tomohide Shibata and Sadao Kurohashi. Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL2018), pp. 579-589, Melbourne, Australia, 2018.7.

§2 研究実施体制

(1) 研究チームの体制について

① 黒橋グループ

・研究代表者 黒橋禎夫(京都大学大学院情報学研究科 教授)
・研究項目 知識に基づく文脈解析の実現と因果関係知識の抽出
- 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
- 知識に基づく統合的文章解析モデルの構築と高度化

実社会課題への適用(乾・戸次ｸﾞﾙｰﾌﾟと共同)
- 企業コンタクトセンター等への適用
- 行政サービス対話ボットの構築

② 戸次グループ

・主たる共同研究者:戸次大介(お茶の水女子大学基幹研究院自然科学系 准教授)
・研究項目 文の意味の表現・計算モデルの構築
- 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

・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)
・研究項目 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
- 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

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

黒橋グループでは、柴田講師を半年間 CMU に派遣し、Event 構造についてプロジェクトを推進している Teruko Mitamura 教授、Eduard Hovy 教授との研究交流を行った。さらに、本プロジェクトと近い問題意識を持つ EU プロジェクト、EXCITEMENT を推進している FBK の HLT 研究ユニット・リーダである Bernardo Magnini 博士とも協力関係を築いた。また、国際強化支援を受け、ゼロ照応解析、事態間知識抽出、推論等の研究を行っている University of Texas at Dallas の Vincent Ng 教授のグループとの協力関係を築いた。2017 年度には Ng 教授が 1 週間、Ng 教授の研究室の博士課程学生 Jing Lu 氏が 1ヶ月間、京都大学に滞在し、深層学習によるイベント解析の共同研究を行った。また、2018 年度には黒橋研博士課程学生 Yin-Jou Huang 氏が 2 ヶ月間 Ng 教授のもとに滞在し、共同研究を継続している。

乾グループでは、仮説推論について Jerry Hobbs 教授(University of Southern California, ISI) と共同研究を行っており、本プロジェクトとしても国際ワークショップに招聘するなど、連携を進めた。また、University College London の Pontus Stenetorp 博士と密な連携関係を築き、エンティティに基づくテキスト解析に関する研究を共同で行っている。その成果の一つはトップ国際会議 EACL2017 に採択され、Outstanding Paper Award を受賞した。

戸次グループでは、数理言語学の国際ワークショップ LENLS を 7 度(2013 年～2019 年)主催し(人工知能学会との共催)、自然言語処理・理論言語学・数理論理学・言語哲学の学際領域にお

ける本CRESTプロジェクトのプレゼンスを向上させるべく努めた。また、2014年9月には田中リベカ（当時戸次研M2）を CNRS, Institut de Recherche en Informatique de Toulouse / Université Paul Sabatier に、2015年3月−6月には叢悠悠（当時戸次研M2）をNew York Universityに、2015年6月−9月には金子貴美（当時戸次研D2）をRakuten Institute of Technology（RIT）にそれぞれ派遣し、研究交流を行っている。2014 年 11 月には CREST International Workshop on Formal and Computational Semanticsを開催し、Chris Barker 教授(New York University)、 Matthew Stone 教授 (Rutgers University)を招聘して交流を深めた。また、2015 年 4 月には傍士元教授（University of Southern California）を招聘し、合宿形式で研究技術の吸収に努めた。

産業界との連携については、本研究プロジェクトの成果展開に関して、不満買取センターの意見表現分析について株式会社 Insight Tech、日立お客様相談センターへの問い合わせ分析について日立アプライアンス株式会社、金融ドキュメントのチェック処理について野村総研、行政サービスの対話ボット構築についてLINE株式会社とそれぞれ共同研究を行った。