１．研究課題名：高度言語理解のための意味・知識処理の基盤技術に関する研究
２．研 究 期 間：平成 18 年度～平成 22 年度
３．研究代表者：辻井　潤一（東京大学・大学院情報学環・教授）

４．研究代表者からの報告

（１）研究課題の目的及び意義

巨大な文書集合の利用により、過去１０年の間に、言語の計算理論と処理技術に大きな進展があった。とくに、文解析研究では、記号処理と機械学習・確率モデルの融合により、高耐性・高効率な処理系の開発が可能となった。しかし、その文解析においても、意味・知識から切り離された現在の手法は、係り受け関係の精度で 90%を越えられず、多くの応用にとって不十分なものとなっている。本研究は、文解析で成功した記号処理と確率モデルの融合枠組みを意味・知識処理へと拡張し、その技術を確立することを目的とする。その成果は、テキストからの情報抽出、高品質機械翻訳、知的検索、高機能な知識管理など、構造を超えた処理が必要な言語処理応用システムの性能を飛躍的に向上させる。

意味・知識処理に基づく言語処理のブレークスルーは、従来研究の単純な延長では達成できない。本研究では、次の３つの課題設定により、この難問に系統的にアプローチする。 まず、第一に、言語表現との対応が自明でない意味・知識に関する研究を系統的取り組むために、知識リソースとそれに基づく意味アノテーションつきコーパスを構築する。第二に、文単位の局所処理である文解析に対して、意味・知識処理には文脈という大域構造が関与する。このための言語モデル、局所モデルと大域モデルの融合したモデルと処理の理論を構築する。第３に、膨大なテキスト集合に対する大域的な確率モデルを構築するための計算機環境、大規模テキスト処理のための GRID 環境を構築する。最後に、意味・知識の処理に関する研究を架空の、抽象的な研究に終わらせないために、研究成果を反映した、実ユーザ（生命科学者）が実際に使うシステムを構築する。

（２）研究の進展状況及び成果の概要

それぞれの項目の研究は順調に進行し、研究初年度、次のような成果を上げた。
①言語・知識資源の構築：１０００件の論文抄録（約 8000 文）に対する意味事象と文脈参照表現のアノテーションを完了し、生命科学分野の代表的オントロジーである GO(Gene Ontology)に基づく知識資源の確立を行った。また、テキスト・アノテーション用のツール(XComc)を開発した。
②意味・知識処理の言語処理も出る：Reference Distribution の活用による、分野適応の確率モデルの枠組みを設定、また、文解析過程への依存構造に基づく確率モデルの導入枠組みを設定し、次年度以降の研究枠組みを規定した。
③計算環境とソフトウェア・ツールの開発：NLP に典型的な複雑な workflow 全体から，web を経由した検索処理までをひとつの枠組みで統合し，分散環境で並列実行できるスクリプト言語 DDS を設計・実装し、その有効性を 1000 台のＰＣから構成されるクラスタ￥に適用し、有効性を確認した。
④統合システムの構築：先行プロジェクトで開発したシステム(MEDIE、Info-Pubmed)をモジュラリティの高い構成に改編し、今後の研究プラットフォームとして整備した。また、病疾患-遺伝子関係認識、蛋白質相互作用認識システムなど、実ユーザからの要求が明確な分野でのシステム構築を行った。

５．審査部会における所見

Ａ（現行のまま推進すればよい）

意味・知識・文脈を扱うには、データ主導と文法理論主導を組み合わせた理論的枠組みの研究とともに、コーパスに意味情報を付与した大規模言語資源の開発が必要である。さらに、大規模分散並列計算環境の活用は必須であり、このための研究を進める必要もある。これらの研究を並行して進めているところに、この組織の実力が伺える。また、利用者が明確な生命科学を応用領域として選び、それに焦点を当てることで着実に研究を進めている。生命科学の文献を対象に、共参照やイベントに対する意味的アノテーションの付与や GENIA オントロジーの構築を完了している点､確率モデルの導入により解析精度の向上が得られた点より、研究は着実に進展していると考えられ、それぞれの成果を統制することによって、さらなる加速が期待できる。以上より、現行のまま推進すればよいと判断した。