審 査 の 結 果 の 要 旨

氏名　薬師寺　あかね

本論文は、生物医学テキストベース(Medline)から蛋白質間の相互作用に関する方法を自動抽出する研究についてまとめたものである。特に、浅い、部分的な統語解析のみを行う従来の手法と異なり、深く、かつ、完全な統語解析を行った後に情報抽出を行う、これまでにない野心的な枠組みを提案し、700 万抄録（7 千万文）を超える巨大な Medline テキストベースを使って提案枠組みの実用性を実証している。このように、本論文は、情報抽出の新しい枠組みの提案、および、その有効性の大規模テキストベースによる実証に、主たる貢献がある。以下に、各章について説明する。

第１章では、生命科学分野、特に、蛋白質間相互作用に関する情報抽出の必要性をのべ、従来手法である機械学習と規則主導の２つの手法を紹介している。第２章では、１章での２つの手法が、ともに、分野とタスクに大きく依存し、研究成果の汎用性に問題があることを指摘し、深い統語解析を前段階で実行する本研究のシステム構成が、従来手法の欠陥を補うものであることを指摘している。第３章では、本研究が比較の対象とするベースライン・システムとして、テキサス大学と日立製作所の２つのシステムを中心に関連研究を簡潔にまとめている。

第４章、第５章では、深い統語解析が情報抽出という実用システムに有効な技術であることを、ペンシルベニア大学で開発された XTAG 解析プログラムを使った抽出実験、および、動詞を中心とした規則自動抽出の実験を行うことで、実証している。４章の実験は、英語解析文法の被覆率が本研究の枠組みに必須の条件であること、また、５章の実験では、深い統語解析が動詞中心のパターン規則の学習を少量のテキストで可能とすることを示しており、この２つの章は、６章の本格的なシステム構築を行う際の指針を与えている。

第６章では、（１）パターン規則構築を動詞中心の事象関係のパターン、名詞句中心の「もの」に関するパターンに分離して学習すること、（２）パターンの信頼度を算出することで、要求される Precision と Recall 率にあわせた規則集合が構築できること、また、（３）これらの部分パターンと規則中の単語を素性とした Classifier（SVM）を使用することで、規則集合だけの場合よりもはるかに優れた結果が得られること、を示している。また、人手によって構築された Reactome のデータベースと比較することで、構築したシステムの性能を評価している。特に、システムが発見した蛋白質対で Reactome に未登録の対が 7620 対存在し、そのうち、60%が Reactome に登録されるべきものであることが確認できるなど、システムのパフォーマンスが非常に優れたものであったことを報告している。結果が膨大なため、正確なパフォーマンス測定はないが、サンプル調査の結果としての数値は、Precison64.4%、Recall85.3%と非常に良好なものであった。

第７章では、提案手法の特徴である分野非依存性を実現するために、深い文解析システムがもつべき特徴に関して整理している。言語自体の分野依存性は、タスク依存性よりもさらに困難な課題であり、ここでの議論は、第９章での今後の課題に引き継がれている。

以上のように、本研究は、深い統語解析の結果を情報抽出という現実的な課題に適用し、しかも、７千万文という膨大な文集合を処理することで、その有効を確認した世界で最初の研究となっている。システムの性能も、世界水準に達するものであり、今後のこの分野での新たな研究方向を示唆する貴重なものとなっている。

以上より、本論文は、博士（情報理工学）の学位請求論文として合格と認められる。