審 査 の 結 果 の 要 旨

氏　　名　　増田　勝也

近年、品詞タガーや構文解析器、固有表現認識器などの基本的な自然言語処理システムの開発整備が進み、十分な精度のより高度な自然言語処理アプリケーションの構築に利用されている。しかしながらそれらの基本的な自然言語処理システムの結果を統合的に利用・管理する効果的な枠組みが存在しないため、簡単な絞り込みによる検索の後に自然言語処理を行うという手法が主流である。一方、情報の利用・共有のために、テキストに対して多種多様な情報をアノテーションとして付与する試みが、自然言語処理分野のみならず様々な分野において行われている。また情報検索の分野では、文字列情報のみを扱う従来のキーワード検索ではなく、アノテーションなどの高度な情報を活用して、利用者の検索意図に適合する結果を提示する意味検索と呼ばれる手法が研究されている。

本論文は、このような意味検索の一つとして種々の自然言語処理を統合的に利用する高度な意味検索システムを実現し、情報検索に対する自然言語処理技術の有用性を示している。本論文は「SEMANTIC SEARCH USING ANNOTATIONS BY NATURAL LANGUAGE PROCESSING: PAPER SEARCH BASED ON EVENTS IN BIOMEDICAL SCIENCE（自然言語処理アノテーションを利用した意味検索: 生命医学系論文に対する事象に基づく検索）」と題され、6章からなり英文で書かれている。

上記の意味検索システムを実現するための手法として、自然言語処理モジュールを利用して言語的情報が付与されたテキストに対し、その情報を利用して従来のキーワードベースの検索に比べより高度な検索を実現する枠組みを提案している。アノテーションにより構造化されたテキストに対する検索枠組である領域代数を拡張し、自然言語処理アノテーションの特徴である入れ子構造に対応した検索アルゴリズム、および変数による参照を利用可能な検索枠組を提案している。また、従来のキーワードベースの検索で使用される確率的言語モデルを拡張し、依存関係が存在する、構造を持つクエリ集合を利用した検索に対するランキング検索手法を提案している。

また、提案システムの実世界への適用例として、生医学論文の要旨データベースであるMEDLINEに対する論文検索システムの構築を行っている。自然言語処理の基本的なモジュールとして、深い構文解析器、固有表現認識器を利用、さらにはより高度なモジュ

ールとしてタンパク質間相互作用等の生医学研究におけるeventの認識器を利用し、それらの処理結果が付与されたテキストを対象データとした意味検索システムとなっている。付与されたアノテーション情報を利用した検索を可能とすることで、生医学研究において重要とされる物質間の相互作用等の検索を可能としている。

実験では、情報検索評価のテストコレクションであるTRECのデータと独自に作成したテストデータを用いて、提案した意味検索システムの検索精度を評価している。従来のキーワード検索に比べ、高精度な検索が可能であることが確認されている。また提案した検索枠組およびアルゴリズムの評価として、既存のXMLデータベースと速度面での比較を行い、提案手法が複雑な検索を高速に実行できることを示している。

以上のように、本研究は、自然言語処理を統合的に利用した情報検索の手法を提案し、その有用性を実用的な生医学分野の論文に対する検索システムの実現、および精度面・速度面の実験を通して示しており、コンピュータ科学の分野、特に自然言語処理及び情報検索の分野において貢献するところが極めて大きい。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。