論文の内容の要旨

論文題目

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）

氏　　名　　増　田　勝　也

本論文では自然言語処理技術を用いた高精度な意味検索システムの実現を目的として、自然言語処理技術を用いてアノテーション情報が付与されたテキストに対し、そのアノテーション情報を検索時に統合的に利用するための意味検索枠組みを提案する。近年、品詞タガーや構文解析器、固有表現認識器等の自然言語処理分野における基本的なシステムが開発され、十分な精度でより高度な自然言語処理アプリケーションにおいて利用されている。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。また一方でテキストに対し多種多様な情報をアノテーションとして付与し、利用・共有を行うという動向が自然言語処理分野のみならず様々な分野において存在する。

そこで本論文では、これらの種々の自然言語処理技術を統合的に利用した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すことを目的とする。また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テキストに対して言語的情報を付与し、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する。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構築する。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する。

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 に対し検索システムを実装する。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し、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、それらの処理結果をテキストにアノテーションとして付与する。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するシステムを構築し、本論文において提案する意味検索システムの有用性を示す。

実験において、提案する意味検索システムの検索精度の評価を独自に作成したテストデータ、および情報検索評価用のテストコレクションを用いて行い自然言語処理を利用した意味検索システムの有効性を示す。また、既存の XML データベースとの比較や種々のクエリによる実験を行い、アノテーションが付与されたテキストを対象とした、検索枠組・アルゴリズム自体の評価を行い、提案枠組の有用性を示す。