研究課題別　事後評価結果

１．研究課題名： 次世代テキストマイニングの技術基盤に関する研究

２．研究代表者： 辻井　潤一　（東京大学　大学院情報理工学系研究科　教授）

３．研究内容及び成果

これまでのテキストマイニング（TM）の技術は、テキストを単なる単語集合(Bag of Words; BOW)とみなして、これに確率・統計モデルに基づくマイニングを適用する「量の技術」であった。これは、大量の言語データを処理するのに十分な処理効率と耐性とを備えた言語処理の技術、意味処理・知識処理に必要となる大量のリソース（意味辞書、知識ベース）の構築、および質・量ともにスケール・アップした処理を支える計算機インフラの開発、の3つの技術分野が未成熟であったためである。しかしながら、構造・意味を取り扱う処理手法と量的側面を取り扱う機械学習からの処理手法との融合により、大量・広範なテキストを処理するのに必要な高耐性・高効率な言語処理技術が現時点で十分達成可能なものとなりつつある。また、大量テキストからの意味辞書・知識ベースを構築する半自動的な手法の発展により、生命科学をはじめとするいくつかの専門分野では、現実に膨大な言語・知識リソースの構築が開始されている。さらに、処理と記憶の能力を格段に向上させるGRID技術（並列処理技術）の発展も著しい。

本研究では、文の統語・意味構造、テキストの文脈構造、および、明確には表現されない背景知識を取り扱う質の技術に焦点をあて、これを量に基づく技術に統合することで、従来の技術をはるかに凌駕するTMの技術基盤を確立することを目的とした。また、膨大なテキストの集積と複数分野での知識統合が進展し、TM技術の需要が顕在化している生命科学で開発したTM技術の有効性を実証した。

以下、研究成果の概要をまとめる。

1）生命科学の課題解決型システムの開発

CREST研究で開発した2つの統合システム（Info-Pubmed, MEDIE）について、実ユーザ（国立遺伝学研究所、理化学研究所、マンチェスター大学バイオ研究センター等の生命科学者）からのフィードバックに基づき、個別課題のための機能を拡充した。特に、疾患と遺伝子との関係をテキストからマイニングするシステム、および1500万件のMedlineテキストベースからタンパク質相互作用に関する情報抽出を行うシステムを構築した。

2）高速・高精度な分野適応型言語処理技術の開発

本研究の目標は、従来のBOWモデルの限界を突破する言語処理に基づいたTM技術

の開発である。このため、本格的な言語理論(HPSG; Head-Driven Phrase Structure Grammar)に基づいて1500万件の論文抄録（7000万文、14億語）という巨大なテキスト集合を処理し、かつ、1文あたりの処理時間が15msecという高速な英語解析システム(Enju)を開発した。また、Medlineに特化した統計モデルを学習する領域適応の技術を利用することで、F-値が９０%を超える精度を達成出来ることを実証した。1)の課題解決型システムは、全てこのEnjuを使った成果である。

3) 知識・言語リソースの構築

CREST研究で開発したアノテーション・コーパス（GENIA）は、生命科学におけるTM技術を開発するための言語資源として世界の研究者に利用されている。本研究では、このGENIAコーパスを知識・意味処理のための基礎コーパスとして整備するために、これまでの名詞的概念のアノテーションを動詞的概念（事象）に拡張し、1000論文抄録に対する生命事象アノテーションを完成した。また、GENIAオントロジを生命科学の主要なオントロジ（GO; Gene Ontology、Mesh; Medical Subject Headings、BioPax; A Standard Data Format for Pathway Data Exchange）とリンクし、リソース間の相互利用を可能とした。

４．事後評価結果

４－１．外部発表(論文、口頭発表等)、特許、研究を通じての新たな知見の取得等の研究成果の状況

期間中の外部発表、特許等の実績を示す。

発表論文： 6 件

口頭発表： 23 件

本プロジェクトは、研究代表者が別の研究プロジェクトを獲得したことで、研究期間が実質的に1年5ヶ月に短縮された。しかし、短縮された研究期間のなかで、先行するCREST研究の活力を維持しつつ、構文・意味処理に基づく本格的な自然言語処理技術によるTM技術を研究開発した。これは、テキストを単語集合として近似する従来のTM技術の限界を超える次世代技術として位置づけられるものであり、その性能を実用的なレベルまで高めたことの意義は大きい。また、他の分野への貢献も大いに期待出来ることから、高く評価される。外部発表は、第一級レベルの国際会議に着実に採択されており、適切な発表状況であるといえる。特許は出願されていないが、本研究で開発されたソフトウェアはオープンに公開され、またアノテーション付コーパスも他研究グループで利用される等、人類の共有財産として適切に活用されていると推察される。

４－２. 成果の戦略目標・科学技術への貢献

情報メディアの中で最も重要であると考えられる言語情報に関して、テキスト処理による知識発見と管理、活用へと発展させる、新たな方向性を打ち出した研究として極めて高く評価出来る。TM統合アプリケーションに関しては、生命科学分野で実用可能な性能を持つことを示した。HPSG構文解析に関しては，従来手法のほぼ10倍の高性能化、構文解析精度の向上、分野適応のための学習性能の向上等の成果を上げた。GENIAコーパスの構築に関しては、世界で初めて事象アノテーション付コーパスを構築する等、世界をリードしている。本研究は、元々生命科学での汎用TM技術の確立を目的としていた。研究期間短縮と組織変更により生命科学者に提供する統合システムの完成に焦点を絞ったものになったが、生命科学分野で実用に供する性能を持つ次世代TM技術を確立したことの意義は極めて大きい。構築された技術の中には、MEDIEのように他分野に適用可能なものもあり、汎用TM技術の展開が期待される。しかし現時点では、ある程度限定された統語・意味パターンを対象とする検索に止まっており、利用範囲も限定的である。文脈を考慮した検索や同義異表現の問題等、本研究では対象にしていない困難な問題への対処を期待する。

４－３. その他の特記事項（受賞歴など）

英国が設立した国立 TM センターの研究所長に研究代表者が任命されたことは、本研究グループが言語処理技術を使った生命科学分野の TM において世界的に注目を浴び、リーダーとして認められたことを示唆するものであると言える。