科学研究費助成事業（特別推進研究）公表用資料〔追跡評価用〕

「意味処理技術の深いマイニング技術への適用」

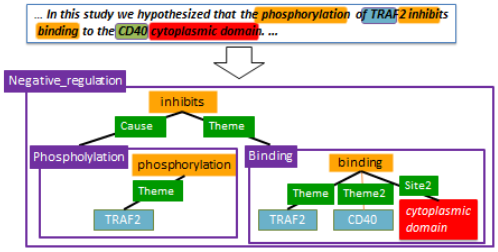
（平成 18～22 年度 特別推進研究（課題番号：18002007）
「高度言語理解のための意味・知識処理の基盤技術に関する研究」）

所属（当時）・氏名：東京大学・情報理工学系研究科・
教授・辻井　潤一
（現職：国立研究開発法人産業技術総合研究所・人工知能研究センター・センター長）

１．研究期間中の研究成果

・背景　文の意味は、単語の集合だけで決まるわけではなく、その組み合わせ方（構文構造）で決まる。従来のテキストマイニング技術は、このような文の構造を考慮せず、テキストを単語集合として取り扱ってきた。これは文の構文解析の精度が低く、また、そのコストが高いことが主たる理由であった。本研究では、単語の列である文が持つ**潜在的な構文構造を認識し、それを意味や知識へと結び付ける基盤技術**を開発することであった。

図１　文からの意味の抽出（事象認識）



・研究内容及び成果の概要　言語学からの文法(HPSG)をテキストから自動学習させ、かつ、本研究で研究した Supertagging、CFG 近似、素性構造の確率モデルの理論などを使うことで、処理速度、精度の点で実用的な構文解析プログラムを構築した。また、構文構造を固有名、事象認識(図 2)と結び付ける意味・知識処理の基盤技術を開発した。これらの技術の有効性を生命分野のテキストマイニングに適用し、本研究の成果が実世界の応用にも重要であることを実証した。

２．研究期間終了後の効果・効用

本研究の成果をプログラムの形で公開した。これらの公開プログラムは、**国内・国外の数多くのグループ**に使われている。また、意味に基づく深いマイニング技術とその生命科学分野への応用は、**言語処理の一大応用分野となり、ACL（計算言語学会）のもとに SIG－BioMedNLP を組織する**ことになった。この SIG は、現在、ACL の大きな SIG の一つとして、活発に活動している。また、本研究で開発したマイニングのシステムの多くは、**英国の NaCTeM(National Centre for Text Mining)から実ユーザに提供されている。**とくに、そのシステムの一つである PathText の問題意識は、現在進行中の米国 DARPA プロジェクト（Big Mechanism）に引き継がれている（図 2)。

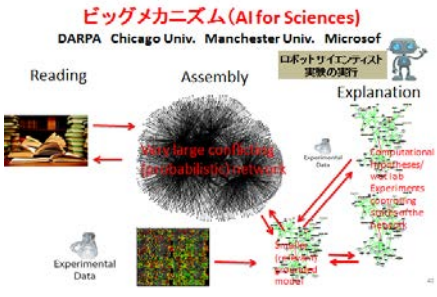


図２　米国 DARPA　ビッグメカニズム

・波及効果

素性構造文法による複雑な記号処理と機械学習手法とを結びつける最初の試みとなった。この研究は、その後、局所モデルと大域的な確率モデルとの **Joint モデルの研究**など、豊かな研究分野となった。また、**意味に基づく深いマイニングの研究**は、生命科学から医療、物性科学でのマイニングへとその応用範囲を広げており、**言語処理の大きな応用分野に育っている。**