科学研究費補助金（特別推進研究）公表用資料
〔研究進捗評価用〕

平成 18 年度採択分
平成 21 年 3 月 31 日現在

研究課題名（和文）
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
研究課題名（英文）
Research on Advanced Natural Language Processing and Text Mining
研究代表者
氏名　辻井　潤一（Junichi Tsujii）
所属研究機関・部局・職　東京大学大学院情報理工学系研究科・教授

研究の概要：本研究は、過去 10 年間、文解析研究で成功してきた手法、すなわち、巨大な文書集合からの機械学習技術と記号処理アルゴリズムとを融合する手法を、意味・文脈・知識処理に適用することで、言語処理技術にブレークスルーをもたらすことを目標とする。このために、テキストへの意味アノテーション付与、分野オントロジーの自動構築、意味・知識に基づく文解析手法、資源共有型の分散計算機環境の構築、の研究を行う。

研　究　分　野：総合領域
科研費の分科・細目：情報学・知能情報学
キ　ー　ワ　ー　ド：言語理解、意味処理、テキストマイニング、文脈処理、知的検索

１．研究開始当初の背景

ウェッブ中のテキスト量の急激な増大、電子出版の一般化など、膨大なテキスト集合を高効率、高精度で処理する言語処理技術への期待が高まると同時に、膨大なテキストの存在は、機械学習・確率モデルに基づく言語処理技術に急速な進展をもたらしていた。しかし、膨大なテキストの存在のみに依拠し、言語構造に関する理論、あるいは意味知識処理に必要な大規模なリソース（意味コーパス、オントロジー）の構築を無視した技術開発の限界も次第に明らかになってきていた。また、意味・知識処理には、膨大なテキスト中の個々の文が持つ微細な構造まで処理するための強力な計算環境が不可欠との認識も共通のものとなっていた。しかし、強力な計算能力を提供しえる PC クラスタなどは存在するものの、それらを現実の言語処理に使うためのシステム・ソフトウェアは存在しなかった。

２．研究の目的

上のような背景から、本研究では、本格的な意味知識処理を含む高度言語処理にとって必要な３つの基盤、（１）構造に関する理論と確率・機械学習の理論を有機的に統合した理論、（２）大規模な意味・知識リソース、（３）大規模データを処理する計算環境を確立した上で、（４）意味・知識処理技術の研究を系統的に行うことを目的とした。また、意味や知識という抽象度が高い対象の研究には、成果の有効性を実証できる応用システムの存在が不可欠と考え、そのための実証システムとして、（５）生命科学分野のテキストマイニングと高品質機械翻訳のシステムの開発をプロジェクト開始当初から同時進行的に行うこととした。

３．研究の方法

（１）理論：言語の構造と意味の関係を系統的に取り扱うために、理論言語学からの文法を計算言語学の「深い文解析」に適用する研究を行う。文法のための確率モデル、浅い文解析と深い文解析の融合手法などについて研究し、理論言語学の文法を深い文解析に適用する基盤技術を確立する。また、深い文解析・意味処理のための機械学習の研究を使った高効率な文解析、意味処理の基盤技術を確立する。

（２）リソース構築：テキスト情報と分野知識との関係をデータ中心に研究するために、テキスト中の表現を分野知識（オントロジー）に結びつける意味コーパス（１００万語規模）を構築する。具体的には、固有名、生命事象などを付与した生命科学分野の意味コーパス（GENIA）を構築する。また、言語処理ソフトウェア共有枠組みを構築する。

（３）計算環境：複雑な意味知識処理を大規模に実行するために、多様な処理モジュールが処理結果を交換しながら分散的に仕事を進めるデータ中心の大規模処理モデルのための計算環境を構築する。

（４）意味・知識処理：（１）～（３）の成果を使い生命科学分野の文献の意味を分野知識で解釈する処理技術（固有名認識，事象認識，プロセス認識）を開発し、これを生命科学のための知的な知識管理システムに統合する。

４．これまでの成果

１．深い文解析と意味知識処理：深い文解析を本格的な情報抽出（タンパク質相互作用の抽出）に適用し、従来のシステムの精度を格段に向上させた。深い解析が情報抽出に有効との結果を世界で最初に実証した。

２．系列 tagging 学習器：隠れ変数を使った機械学習を言語処理へ適用し、深い文解析の速度を 20 倍向上させるとともに、固有名認識などの意味処理タスクでも、世界最高水準のパフォーマンスを達成した。

３．GENIA コーパス：構築した GENIA コーパスは、これを使った国際コンペティションに２４チームが参加するなど、生命科学分野でのデ・ファクトの国際標準となっている。

４．U-Compare：言語処理ソフトウェア共有枠組み（U-Compare）は、世界で最大(組み込みツール４０超)の共有枠組みとなっている。この研究は、UIMA Innovation Award を IBM Watson 研究所より受賞(2009 年）。

５．計算環境：並列処理の記述を殆どしなくてよい汎用的ワークフロー処理系、任意の計算資源の上に分散ファイルシステムを構築するシステムという、汎用性の高いデータ処理の枠組みを確立した。これは、我々の研究に日常的に使われているだけでなく、今後のクラウド環境など大規模な計算資源を柔軟に使う基礎技術となっている。

５．今後の計画

（１）理論：深い文解析のための新たな枠組み（統合的パイプライン・モデル）で処理効率、精度とも世界最高の深い文解析器を完成させる。深い文解析、意味処理の基礎技術としての隠れ変数をもつ系列 Tagging の機械学習モデルの高効率化を達成する。

（２）リソース構築：全体部分関係など静的関係に基づく間喩表現を対象にした静的関係のアノテーション、および、フル・ペーパーの意味アノテーションを完成し、公開する。

プロジェクトで開発されたソフトウェアを我々が主導するツール共有の国際コンソーシアム U-Compare から公開する。

（３）計算環境：　多様な言語処理モジュールが中間ファイルを介して通信し、分散並列処理を実行する新しい方式の大規模テキスト処理の形態を実現する。このシステム構築を通して、スーパーコンピュータ、クラスタ、クラウドなどの資源を自在に組み合わせて用いることが可能な大規模テキスト処理基盤を確立する。

（４）意味知識処理：現在世界最高水準の生命科学分野での NER/ER をもとに関係認識の一般的手法とする。また、複数事象をまとめて構造化知識に写像する文脈処理技術を生命科学の Pathway 表現を対象にして実現する。

（５）応用システム：生命科学の知的検索システム MEDIE を高機能化した MEDIE2.0，および，Pathway 表現とテキストを結ぶ PathText システムを一般ユーザに公開する。また、研究成果を日中械翻訳システムに適用し有効性を確認する。

６．これまでの発表論文等

[受賞]

(1)IBM Innovation Award (IBM Watson Center), 2009
(2)Research Award (Microsoft research Asia), 2009
(3)Best Paper Award (ICGL 08, Hong Kong), 2008

[発表論文]

(1) Sun, X, **J. Tsujii**, et.al: Latent Variable Perceptron Algorithm for Structured Classification, IJCAI. 2009.

(2) Kano, Y., **J.Tsujii**, et.al: U-Compare: share and compare text mining tools with UIMA. Bioinformatics, Oxford University Press, 2009.

(3) Miyao, Y, **J.Tsujii**, et.al: Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction. Bioinformatics. OUP, 2009.

(4) Okazaki, N., **J.Tsujii**, et.al: A Discriminative Candidate Generator for String Transformations. EMNLP 2008.

(5) Miyao, Y. and **J.Tsujii**. Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics. 34(1). MIT Press, 2008.

(6) Matsuzaki, T., et.al: Comparative Parser Performance Analysis across Grammar Frameworks through Automatic Tree Conversion using Synchronous Grammars. COLING, 2008

(7) Hironaka, K., K.Taura. et.al: gluepy: A Simple Distributed Python Framework for Complex Grid Environments. LCPC 2008

(8) Sagae, K., Y.Miyao, et.al: Challenges in Mapping of Syntactic Representations for Framework Independent Parser Evaluation. ICGL, 2008.

(9) Kim, JD., **J.Tsujii**, et.al: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 9(1). 2008. ISSN 1471-2105.

(10) Oda, K., **J.Tsujii**. et.al: New challenges for text mining: Mapping between text and manually curated pathways. BMC Bioinformatics. 2008.

(11) Matsuzaki, T., et.al: Efficient HPSG Parsing with Supertagging and CFG-filtering. IJCAI 2007.

ホームページ等

http://www-tsujii.is.s.u-tokyo.ac.jp/index-j.html