様式 C-7-2

自己評価報告書

平成２１年５月７日現在

研究種目：　特別推進研究
研究期間：　2006　～　2010
課題番号：　18002007
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　辻井潤一
東京大学・大学院情報理工学系研究科・教授
研究者番号：20026313

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．研究計画の概要

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。

２．研究の進捗状況

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。

３．現在までの達成度

（１）理論面では、系列 tagging の機械学習とパイプライン処理方式とを組み合わせることで、20 倍という当初計画を超える飛躍的な効率向上を得た。この方式は、曖昧さ保持のデータ構造が不要な単純な枠組みで意味・統語の融合処理が実行できる優れた特徴を持つ。今後、系列 Tagging での新たな成果を取り込むことで、精度も飛躍的に向上することが期待できる。この方式は、単語分割処理の困難さのために高精度解析ができなかった中国語、アジア諸語の処理にも大きな効果をもつ。このための実証実験を計画している。これは、当初計画にはない新たな成果となろう。

（２）GENIA コーパスが生命科学分野でのデ・ファクト世界標準となったことで、このコーパスへの意味アノテーションが他のグループでも行われるようになり、GENIA 意味リソースの構築が加速度的に進展している。NUS(シンガポール)の共参照関係、BioScope(University of Szeged、ハンガリー)によるモーダル情報、Linköpings 大学(スウェーデン)による GENIA オントロジーの写像、UCL(英国)による依存構造付与、ウィスコンシン大学（米国）による文脈構造アノテーションなどがその例である。これらにより、当初計画よりもはるかに早く、しかも、より豊かな意味リソースが，我々の GENIA コーパスを中核にして構築されつつある。

（３）生命科学分野で Semantic Web への興

味が高まり、ユニーク ID(URI)の設定の動きが国際的に加速、EU プロジェクトで構築された意味辞書(BioLexicon)など、国際協力で構築される大規模な意味リソースが使用可能となってきた。とくに、BioLexiocn は、緊密な共同研究パートナーのマンチェスター大学が構築したもので、本プロジェクトでの使用が始まっている。ソフトウェアの共有も、平成 20 年度より我々の主導で始まった U-Compare が普及し、他グループで開発されたソフトウェアが簡単に使用できるようになった。このような緊密なリソース共有の国際協力は，当初計画にはなかったものであり，研究の進展を加速している。

（４）計算環境の構築、パイプライン処理モデルの進展により、当初予期していた効率の問題が予想よりもはるかに早く解決できる見通しとなった。

これらは、当初の目的を超える研究の進展があり、予定以上の成果が見込まれる。

４．今後の研究の推進方策

（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。

５． 代表的な研究成果

（研究代表者、研究分担者及び連携研究者には下線）平成２０年９月～平成２１年３月

〔雑誌論文〕（計３件）

(1)Miyao, Yusuke, Kenji Sagae, Rune Sætre, Takuya Matsuzaki and **Jun'ichi Tsujii**. **Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction.** Bioinformatics. 25(3). pp. 394-400, Oxford University Press, 2009.

(2)Tsuruoka, Yoshimasa, **Jun'ichi Tsujii**, and Sophia Ananiadou. **Accelerating the annotation of sparse named entities by dynamic sentence selection.** BMC Bioinformatics. 9 (Suppl 11). pp. S8, Nov 2008.

(3)Tsuruoka, Yoshimasa, **Jun'ichi Tsujii**, and Sophia Ananiadou. **FACTA: a text search engine for finding associated biomedical concepts.** Bioinformatics. 24(21). pp. 2259-60, Nov 2008.

〔学会発表〕（計２６件）

(1)Yu, Kun and **Jun'ichi Tsujii**. Improving Long-length Dependency Parsing by Parser Ensemble. 言語処理学会第 15 回年次大会，2009.

(2)Wu, Xianchao, Naoaki Okazaki and **Jun'ichi Tsujii**. Self-Training for Mining Parenthetical Translations in Monolingual Web Pages. In the JaNLP (言語処理大会). 2009.

(3)Sun, Xu and **Jun'ichi Tsujii**. **Sequential Labeling with Latent Variables: An Exact Inference Algorithm and An Efficient Approximation.** In the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009). Athens, Greece, pp. 772-780, 2009 .

(4)岡野原大輔，**辻井潤一**．**階層木言語モデルの音声認識への適用．**日本音響学会 2009 年春季研究発表会．March 2009.

(5)Wang, Xiangli, Shunya Iwasawa, Yusuke Miyao, Takuya Matsuzaki and **Jun'ichi Tsujii**. **The Design of Chinese HPSG for Data-Oriented Parsing.** In the 言語処理学会第 15 回年次大会．2009.

(6)松林優一郎，**辻井 潤一**．**自動意味役割付与のための役割集合の設計**．言語処理学会第 15 回年次大会発表論文集．pp. 364-367, March 2009.

(7)綱川隆司，劉 瀟，岡崎 直観，**辻井 潤一**．**日中漢字の対応関係の自動獲得と中日語彙翻訳**．言語処理学会第 15 回年次大会発表論文集．pp. 857-860，2009.

(8)羽鳥潤，宮尾 祐介，**辻井 潤一**．**語義曖昧性解消における統語的依存関係の寄与について．**言語処理学会第 15 回年次大

会発表論文集 (NLP2009). pp. 658--661, March 2009.
(9)三輪誠, **辻井 潤一**. **蛋白質相互作用抽出への転移学習の応用**. 言語処理学会第15回年次大会発表論文集 (NLP2009). March 2009.
(10)大内田賢太, 金進東, **辻井潤一**. **GuideLink: ガイドラインの管理を同時に行う アノテーションツール**. 言語処理学会第 15 回年次大会発表論文集 (NLP2009). March 2009.
(11)岡野原大輔, **辻井潤一**. **ロジスティク回帰モデルを用いたラベル付文書クラスタリング**. 言語処理学会第 15 回年次大会発表論文集 (NLP2009). March 2009.
(12)柴田 剛志, **田浦 健次朗**. **ポロジ情報を用いた効率的かつ漸近安定な大容量ブロードキャスト**(SACSIS 2009)
(13)Nan Dun, Kenjiro Taura, and Akinori Yonezawa: **GMount: An Ad Hoc and Locality-Aware Distributed File System by Using SSH and FUSE** (CCGrid 2009)
(14)岡野原大輔, **辻井潤一**. **階層木 Logistic 回帰モデルによる多クラス分類**. IBIS. Oct 2008.
(15)Masuda, Katsuya and **Jun'ichi Tsujii**. **Nested Region Algebra Extended with Variables for Tag-Annotated Text Search.** In the Proceedings of CIKM 2008 Poster Sessions. pp. 1349-1350, October 2008.
(16)Kano, Yoshinobu, Ngan Nguyen, Rune Sætre, Kazuhiro Yoshida, Yusuke Miyao, Yoshimasa Tsuruoka, Yuichiro Matsubayashi, Sophia Ananiadou and **Jun'ichi Tsujii**. **Filling the Gaps Between Tools and Users: A Tool Comparator, Using Protein-Protein Interactions as an Example.** In the Proceedings of The Pacific Symposium on Biocomputing (PSB). (13). Hawaii, USA, pp. 616-627, January 2008.
(17)Okazaki, Naoaki, Yoshimasa Tsuruoka, Sophia Ananiadou and **Jun'ichi Tsujii**. **A Discriminative Candidate Generator for String Transformations.** In the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008). Hawai, USA, pp. 447-456, October 2008.
(18)Kim, Jin-Dong, Ohta, Tomoko, Oda, Kanae and Tsujii, **Jun'ichi Tsujii**. **From Text to Pathway: Corpus Annotation for Knowledge Acquisition from Biomedical Literature.** In Alvis Brazma, Satoru Miyano, Tatsuya Akutsu (Eds.), Proceedings of the 6th Asia Pacific Bioinformatics Conference. Series on Advances in Bioinformatics and Computational Biology6. pp. 165-176, Imperial College Press, 2008. ISSN 1751-6404.
(19)Wu, Xianchao, Naoaki Okazaki, Takashi Tsunakawa and **Jun'ichi Tsujii**. **Improving English-to-Chinese Translation for Technical Terms Using Morphological Information.** In the Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008). pp. 202-211, 2008.
(20)Miwa, Makoto, Rune Satre, Yusuke Miyao, Tomoko Ohta and **Jun'ichi Tsujii**. **Combining Multiple Layers of Syntactic Information for Protein-Protein Interaction Extraction.** In the Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008). Turku, Finland, pp. 101--108, September 2008.
(21)Wang, Yue, Jin-Dong Kim, Rune Sætre and **Jun'ichi Tsujii**. **Exploring the Compatibility of Heterogeneous Protein Annotations Toward Corpus Integration.** In the Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008). Turku, Finland, pp. 117--124, September 2008.
(22)Pyysalo, Sampo, Rune Sætre, **Jun'ichi Tsujii** and Tapio Salakoski. **Why Biomedical Relation Extraction Results are Incomparable and What to do about it.** In Tapio Salakoski, Dietrich Rebholz-Schuhmann and Sampo Pyysalo (Eds.), Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland. pp. 149--152, Turku Centre for Computer Science (TUCS), 2008.
(23)Sagae, Kenji, Yusuke Miyao, Takuya Matsuzaki, and **Jun'ichi Tsujii**. **Challenges in Mapping of Syntactic Representations for Framework-Independent Parser Evaluation.** In the Proceedings of the Workshop on Automated Syntatic Annotations for Interoperable Language Resources at the First International Conference on Global Interoperability for Language Resources (ICGL'08). 2008.

(24)岡野原大輔，辻井 潤一．全ての部分文字列を考慮した文書分類．情報処理学会研究会報告．NL(187)．September 2008．
(25)劉 瀟，綱川 隆司，岡崎 直観，辻井 潤一．アラインメントに基づいた日中漢字の対応関係における解析． 第 188 回自然言語処理研究発表会．2008．
(26)岡野原大輔，辻井 潤一．大規模コーパスを扱うためのツール群． 第 3 回言語処理若手の会．Sep 2008．

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）