様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月　１日現在

機関番号：６２６１５
研究種目：基盤研究（C）
研究期間：2009～2012
課題番号：２１５００１３０
研究課題名（和文）　自然言語処理特化型の視覚的・対話的な半自動エラー解析のできる統合機械学習システム
研究課題名（英文）　An integrated machine learning system that allows visual and interactive semi-automated error analysis specialized in natural language processing
研究代表者
狩野　芳伸（KANO YOSHINOBU）
国立情報学研究所・情報社会相関研究系・外来研究員
研究者番号：20506729

研究成果の概要（和文）：自然言語処理の研究開発において頻用される、教師つき機械学習手法をより容易に、かつ効果的に利用可能にするために、作業を省力化する自動化機能と、結果の解析を効率的・効果的に行う機能のプロトタイプ的な設計と実装を行った。主に、英語の生物学論文テキストマイニングにおける応用を行い、雑誌論文や学会等での発表や研究協力を通じユーザフィードバックを得て、システムの妥当性や将来研究に向けての課題点を検証した。

研究成果の概要（英文）：Supervised machine learning methods are very often used in research and development of natural language processing technologies. We have designed and developed prototypical features that allow users to exploit the machine learning methods more easily and effectively, supporting automation and result analysis. We have applied these features to text mining from English biomedical literatures, obtaining user feedbacks for potential future works.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	500,000	150,000	650,000
２０１０年度	1,300,000	390,000	1,690,000
２０１１年度	1,500,000	450,000	1,950,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

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：　自然言語処理、機械学習、エラー解析、視覚化、相互運用性

１．研究開始当初の背景

自然言語処理分野の研究開発において、教師つき機械学習手法の利用は標準的なものになり、自然言語処理のツールを作成する際にはほとんどの研究開発で教師つき機械学習手法が用いられるようになった。結果として多くの研究開発において類似のライブラリやタスクフローが用いられている。しかし、開発者は一からプログラミングを行うことが多く、本来核心的ではない部分の作業に労力を取られてしまいがちであった。

２．研究の目的

本研究の目的は、自然言語処理の研究開発において頻用される、教師つき機械学習手法をより容易に、かつ効果的に利用可能にすることである。そのために、ユーザの作業を省力化する自動化機能と、結果の解析を効率的・効果的にする補助機能のプロトタイプ的な設計と実装を行う。

３．研究の方法
研究目的の達成のためには、大きく分けて 1.互換コンポーネント(コーパスリーダー・ツール)の作成とデータ型階層の設計 2.基盤的な実行システム 3.ユーザ・開発者のからのフィードバックが必要である。

(1) まず、国際標準に対応したコンポーネントを充実させるため、各国の研究機関と協力し作業を行うと同時に、データ型階層の定義を行い新たなデータタイプをカバーした。基盤システムおよびコンポーネントは、国際標準である UIMA(Unstructured Information Management Architecture)に準拠した実装とすることで、標準化と相互運用性の向上を図った。

(2)基盤システムの拡張については、機械学習との接続関係、特に視覚化部分の改良と発展の実装作業を行った。また、並列化によりワークフロー実行時のパフォーマンスを向上させるため、任意の UIMA コンポーネントをクラスタシステムに自動分散展開して SOAP ウェブサービス化し、外部からは単一サービスとして実行できる機構を実装した。

(3)関連学会への出席や講演等を積極的に行い、ユーザ数の増加を目指すとともにフィードバックを受けシステムの改良を行った。

４．研究成果
システムの実行の流れとしてはおおまかに、入力データの読み込み、学習素性の生成、学習素性の抽出、機械学習、結果の解析、となる。これらの各要素を、コーパスリーダーや言語処理ツールといったコンポーネントに分割かつ標準化・互換化し、現実的なタスクにおける応用を行った。応用の対象は主に、英語で記述された生物学論文からのテキストマイニングであった。応用における具体的な事例について、データ型階層やコンポーネント化の体系化を行い、システムの応用を含めた成果をまとめ、雑誌論文や学会等において発表した。

(1)応用の一つとして、BioNLP 2009 Shared Task on Event Extraction に organizer として参加し、基盤システムを公式サポートシステムとして提供した。BioNLP 2009 Shared Task は、英語の生物学論文から、タンパク質の記述、およびタンパク質間の相互作用の記述を自動抽出する性能を競う国際コンテストである。

(2)さらに、基盤システムの機能を用いて、BioNLP Shared Task 2009 の参加者の結果の混合を行い、どの参加者の結果よりもよりよい結果を得ることに成功した。発展として、参加者から数グループと共同作業を行い、参加者のツールを基盤システムに対応した互換コンポーネントとした。

(3)また別の応用として、Bio Creative II.5 に参加し、基盤システムを用いてワークフローの生成とサービス化を行った。CoNLL 2010 Shared Task でも公式ツールとして対応互換コンポーネントが提供されるなど、ユーザ層を確実に広げることができた。

(4)U-Compare 互換の UIMA コンポーネントについては、英語の言語資源に限られていたが、他の研究機関と協力して日本語の主立った言語資源の互換化作業を行なった。具体的には type system(データ型定義)の拡張を行った上で、国立国語研究所の「日本語コーパス」を中心に形態素解析器や係り受け解析器などを互換化した上でリポジトリに追加した。この作業を通じて、基盤機能が言語に依存せず動作することも確認できた。

(5)基盤システムの拡張については、機械学習 API の統合と素性重みの解析機能について、プロトタイプの実装作業を行い動作を確認した。その結果を踏まえて、SVM・MEMM・CRF といった異なる機械学習手法をある程度共通して扱えるよう、また、素性選択がより容易に行えるような機能を考慮したリリース版の設計を進めた。ワークフロー生成 GUI についても、全面的な改良版の実装を進めた。

(6)ユーザフィードバックからの調査では、実装した各種機能への期待と需要が高い一方で、現行の設計では機能を使いこなすためのハードルが未だ高く、ユーザの期待する多様な応用領域のデータにも必ずしも対応しきれていないことがわかった。そこで本研究をパイロットタスク的な試みと位置づけ、まずは一通りの設計と実装を行い、将来研究でのさらなる発展を念頭に不足点の洗い出しを行った。また、これらの研究の経験から、新規の設計および実装による汎用自動化プラットフォームを構築する、別途の発展的な研究課題につなげることができた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）
① Yoshinobu Kano, Jari Bjorne, Filip Ginter, Tapio Salakoski, Ekaterina Buyko, Udo Hahn, K Bretonnel Cohen, Karin Verspoor, Christophe Roeder, Lawrence E Hunter, Halil Kilicoglu, Sabine Bergler,

Sofie Van Landeghem, Thomas Van Parys, Yves Van de Peer, Makoto Miwa, Sophia Ananiadou, Mariana Neves, Alberto Pascual-Montano, Arzucan Ozgur, Dragomir R Radev, Sebastian Riedel, Rune Saetre, Hong-Woo Chun, Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta and Jun'ichi Tsujii: "U-Compare bio-event meta-service: compatible BioNLP event extraction services" BMC Bioinformatics, 12:481, 2011.
DOI:10.1186/1471-2105-12-481

②Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano and Jun'ichi Tsujii: "Extracting bio-molecular events from literature - the BioNLP'09 shared task" Computational Intelligence, 27(4): 513-540, 2011.
DOI: 10.1111/j.1467-8640.2011.00398.x

③Yoshinobu Kano, Makoto Miwa, Kevin Cohen, Larry Hunter, Sophia Ananiadou and Jun'ichi Tsujii: "U-Compare: a modular NLP workflow construction and evaluation system" IBM Journal of Research and Development, 55(3): 11:1-11:10, 2011.
DOI: 10.1147/JRD.2011.2105691

④ Yoshinobu Kano, Paul Dobson, Mio Nakanishi, Jun'ichi Tsujii and Sophia Ananiadou: "Text mining meets workflow: linking U-Compare with Taverna" Bioinformatics, Oxford Journals 26(19): 2486-2487, 2010.
DOI: 10.1093/bioinformatics/btq464

⑤Rune Saetre, Kazuhiro Yoshida, Makoto Miwa, Takuya Matsuzaki, Yoshinobu Kano and Junichi Tsujii: "Extracting Protein-Interactions from Text with the Unified AkaneRE Event Extraction System" Transactions on Computational Biology and Bioinformatics (TCBB) 7(3): 442-453, 2010.
DOI: 10.1109/TCBB.2010.46

⑥Yoshinobu Kano, William A. Baumgartner Jr, Luke McCrohon, Sophia Ananiadou, K. Bretonnel Cohen, Lawrence Hunter and Jun'ichi Tsujii: "U-Compare : share and compare text mining tools with UIMA" Bioinformatics 25(15): 1997-1998, 2009.
DOI: 10.1093/bioinformatics/btp289

〔学会発表〕(計７件)

①狩野芳伸, 橋田浩一: "BCCWJ と関連ツールの相互運用"『現代日本語書き言葉均衡コーパス』完成記念講演会（招待講演）. (20110802). JA 共済ビルカンファレンスホール（東京都）

②狩野芳伸, 橋田浩一: "日本語言語資源の統合的相互運用" 言語処理学会第 17 回年次大会. (20110310). 豊橋技術科学大学

③狩野芳伸: "自然言語処理プラットフォームの現状と利用" 英語コーパス学会第 36 回大会シンポジウム. (20101009). 東京大学駒場キャンパス 招待講演

④Yoshinobu Kano, Ruben Dorado, Luke McCrohon, Sophia Ananiadou, Jun'ichi Tsujii: "U-Compare: An integrated language resource evaluation platform including a comprehensive UIMA resource library" Seventh International Conference on Language Resources and Evaluation (LREC 2010). (20100519). Valletta, Malta

⑤Rune Saetre, Kazuhiro Yoshida, Makoto Miwa, Takuya Matsuzaki, Yoshinobu Kano, Jun'ichi Tsujii: "AkaneRE Relation Extraction : Protein Interaction and Normalization in the BioCreAtIvE II.5 Challenge" In the BioCreative II.5 Workshop 2009 special session | Digital Annotations. (20091008). Madrid, Spain

⑥Yoshinobu Kano, Luke McCrohon, Sophia Ananiadou, and Jun'ichi Tsujii: "Integrated NLP Evaluation System for Pluggable Evaluation Metrics with Extensive Interoperable Toolkit (査読有)" In the Software engineering, testing, and quality assurance for natural language processing workshop (SETQA-NLP), the 2009 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). (20090605). Boulder, Colorado, USA

⑦Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii: "Overview of BioNLP'09 Shared Task on Event Extraction" In the Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. (20090605). Boulder, Colorado, USA

６．研究組織

(1)研究代表者

狩野　芳伸（KANO YOSHINOBU）
国立情報学研究所・情報社会相関研究系・
外来研究員
研究者番号：20506729

(2)研究分担者

三輪　誠（MIWA MAKOTO）
東京大学大学院情報学環・特任研究員
研究者番号：00529646
（H21→H22）