様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 3 月 31 日現在

機関番号：１２６０１
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（JUNICHI TSUJII）
東京大学・大学院情報理工学系研究科・教授
研究者番号： ２００２６３１３

研究成果の概要（和文）：本研究は、文解析研究で成功してきた手法、すなわち、巨大な文書集合を使った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して研究を遂行した。この結果、(1)言語理論に基づく深い文解析の高速で高耐性なシステムの開発、(2)意味・知識処理のための大規模付記コーパス(GENIA コーパス)の構築と公開、(3)深い文解析の結果を用いた固有名、事象認識などの意味・知識処理手法の開発、(4)大規模なテキスト集合の意味・知識処理を行うためのクラウド処理用ソフトウェアシステムの開発、において世界水準の成果を上げた。

(2) で構築された GENIA コーパスは、生命科学分野でのテキストマイニング研究のための標準データ (Gold Standard) として、国際コンペティション (BioNLP09、BioNLP11) の訓練・テスト用のデータとして、採用された。また、(1)の研究成果と機械学習とを組み合わせた(3)の成果は、これらのコンペティションで高い成績を収めている。また、(1)と(4)の成果により、Medline の論文抄録データベース(２千万件、２億超の文)からの事象認識と固有名認識を数日で完了できることを実証した。その成果は、意味処理に基づく知的な文献検索システム(MEDIE)として公開されている。

研究成果の概要（英文）：The objective of the project was to apply the methodology of combining statistical modeling with structure-based symbolic processing, which had proven successful in sentence parsing, to more challenging tasks such as deep semantic processing, knowledge-based information extraction and contextual processing. We have achieved significant results in (1) efficient and robust deep parsing based on a linguistically sound formalism, (2) a large scale semantically annotated corpus for the biology domain (GENIA corpus), (3) information extraction programs (named entity recognizers and event recognizers) for the biology domain which combine the deep parsing in (1) and structural machine learning algorithms, and (4) Workflow software for data-centered parallel processing.

The GENIA corpus in (2) has been recognized as the gold standard corpus for research of text mining for biology and has been used by many groups in the world. It was adopted as the training and test corpus for international shared task competition twice (BioNLP 09 and BioNLP 11). The extraction programs developed in (3) successfully showed the state of the art performance in these international shared task competitions. The system based on (1) and (4) showed that the technology developed by this project was practical for processing the real world text. We successfully processed the whole of MEDLINE (20 million abstracts, more than 2 billion sentences) and indexed them semantically in less than a week. The processing results of MEDLINE has been made publicly available through an intelligent document retrieval system (MEDIE)

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	73,200,000	21,960,000	95,160,000
２００７年度	80,700,000	24,210,000	104,910,000
２００８年度	77,700,000	23,310,000	101,010,000
２００９年度	79,400,000	23,820,000	103,220,000
２０１０年度	73,100,000	21,930,000	95,030,000
総　計	384,100,000	115,230,000	499,330,000

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：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。浅い文解析を使った情報抽出の研究は行われつつあったが、文の深い意味構造に関する吟味は行われておらず、大量のデータの存在のみに依存した、理論の裏づけのとぼしいものであった。

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には、自然言語ワークフロー中で生成される大量の中間処理の自動的な蓄積と後続処理への自動引渡しなど不可欠な機能を備えたシステム・ソフトウェアは存在しなかった。

実際のユーザに供される情報検索サービスや機械翻訳は、単語 n-字組による索引付けやその統計的な性質を利用したものにとどまり、言語の意味や構造を利用したものなかった。

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４つの基盤、すなわち、（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・知識リソースの構築、（３）深い文解析の成果を活用した意味・知識処理技術の確立、（４）大規模テキスト集合を処理するためのソフトウェア環境の開発、を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である。このために、研究成果の有効性を実証するシステムの開発を同時並行的に行なった。実証システムとしては、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をとりあげ、これら２つの応用システムにおいて、従来システムの性能や機能を格段に向上させることを目的とした。

３．研究の方法

以下では、研究目的の項にしたがって、それぞれの方法を記述する。

（１）　確率・機械学習と構造処理の統合: 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(HPSG 文法)を計算言語学の「深い文解析」に適用する研究を行う。HPSG に代表される現代的な文法理論は、単純な非終端記号や終端記号による形式言語理論からの文法枠組みとは異なり、これらの記号を素性の束と捉える。従って、従来の文脈自由型文法の確率モデルは適用できず、このための新たな確率モデルを開発する。また、この確率モデルにしたがって計算される確率を利用することで、非決定的な処理である文解析を効率的に行うための探索手法を確立する。この探索手法は、コンピュータ・ゲームにおけるビーム探索、反復ビーム探索を参考に開発する。と同様な枠組みに基づく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。

（２）　意味・文脈処理のためのリソース構築: 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、処理の対象分野を生命科学（とくに、分子生物学）とすることで、この分野の代表的なオントロジーである GO(Gene Ontology)をもとにして言語の意味処理用のオントロジーを設計し、このオントロジーに基づいてテキストに意味の付記（アノテーション）を行う。アノテーションの対象は、遺伝子・たんぱく質・セルラインなどの固有名（Named Entities）だけでなく、GO で機能・事象に分類される事象についてもテキストに対する意味付与を行う。これは、プロジェクト前より構築を開始していたした生命科学分野の意味コーパス（GENIA）の更なる拡充と利便性をあげるものである。

（３）　意味・知識にもとづく情報抽出手法の研究：言語処理における中間的な構造（品詞並び、句構造、深層格構造、依存構造など）を機械学習による分類器の入力として活用する手法について研究を行う。特に、その計算効率や精度の高さから研究が急速に進展している SVM、CRF の手法と構造的なカーネルを組み合わせる方向で研究を行い、意味・知識処理の代表的な課題である固有名・事象認識の手法を確立する。また、テキストに意味付記を行ったコーパスは、分子生物学に限っても数多く作成されつつある。これらの中には、機械学習の訓練データとしては量的に不十分なものもあるが、これらの多様な意味付記コーパスを材料にして、今後の言語処理における不可欠な技術になると考えられる分野適応、タスク適応に関する手法を確立する。

（４）　大規模テキスト処理のためのソフトウェアと計算環境: 複雑な意味知識処理を大規模に実行するためには、多様な処理モジュールが処理結果を交換する必要がある。しかも、処理が必要なテキスト集合の爆発的な増大に対処するためには、GRID 環境のような並列・分散的な計算環境の上で、このような複雑なワークフローを簡単に実行できる必要がある。また、研究グループによって使用する計算環境が変化する場合にも、この言語処理のためのワークフローがそのまま実行できることが必要となる。このためワークフローソフトウェアを、プロジェクト前から開発してきた仮想 GRID 環境のためのソフトウェア（GXP）をデータ中心的なワークフローに対処し、最適化する形で開発する。

４．研究成果

１．深い文解析と意味知識処理：深い文解析は、言語処理の基礎技術として研究されてきたが、これまでその効率と処理範囲の限定のために実世界のテキスト処理への適用は行われてこなかった。特に後者は、記号表現された文法が精密化するにともない、現実に使われる多様な言語表現が取り扱えなくなるという矛盾をしめすものであった。

これに対して、本研究では、記号表現される文法の役割を統計モデルの定義域を定めるものと捉え、言語表現が満足すべき制約条件ではなく、優先解釈を選択するための確率モデルの中で捉えることで、非常に高い耐性をもった深い解析器の開発に成功した。これにより、生命科学・医学分野の抄録を集めた、大きなテキスト集合（2 億文超）を処理することに成功し、深い文解析器が現実の言語処理応用システムの基盤技術となりえることを実証した。また、確率モデルの整備により、係り受け関係を単位とする精度評価で92%(F-値)を達成し、精度の点でも実用レベルの性能を得た。

この深い文解析器は、本プロジェクトにおいて、生命科学分野での本格的な情報抽出（タンパク質相互作用の抽出、事象認識)、および、高品質の機械翻訳システムに適用し、それぞれの分野で State-of-the-Arts の性能を上げた。

２．発見的探索手法、系列 tagging の文解析への適用：現代言語理論を言語処理に適用する際のもうひとつの問題は、処理速度にある。言語処理が高度化し、意味処理・文脈処理などより洗練された処理を組み込めば組み込むほど、処理速度は低下する。これに対して、ウェッブ規模の巨大なテキスト集合を処理するためには、大規模な計算機並列環境を整えたとしても、基本の文解析器の処理速度が全体の性能を大きく左右する。

上記の考察から、本プロジェクトでは、処理精度とともに、文解析器の処理速度の向上に関する研究を積極的に推進した。特に、文構造に確率が付与されることから、これを解探索に反映させ、反復的にビーム幅を広げる反復ビーム探索が文解析に有効であることを示した。これにより、探索誤りを最小化し、かつ、処理効率を画期的に改善できることを示した。プロジェクト終了時では、文解析器（Enju）は、平均文長が３０単語と長い MEDLINE の文も、平均 600ms で処理が可能であった。

反復ビーム法では、解析初期にボトムアップに組み立てられる小さな解析木の中からどれだけ正しいものがビーム幅内に入るかが探索精度の性能、処理速度の双方に大きな影響を持つ。このことから、解析初期、特に、各単語に品詞割り当てを行う POS Tagger の性能が全体の性能を大きく左右することが研究の結果、明らかとなった。このために、部分解析木のボトムアップな確率付与とは

別に、単語の出現する周辺環境を考慮する系列 Tagging のモデルに研究を集中し、この処理の初期段階により豊かな統語情報（Argument 構造、受身や関係節などの文法変形の有無）を推定する行う Super tagging 処理を提案、これを処理の中心にすえる文解析器（Mogura）は、Enju と同じ文法を用いながら、処理速度を 15 倍程度向上させることに成功した（平均 40ms)。また、浅い依存構造解析器をこの段階で併用することで、探索誤りを減少させ、Enju の統合モデルよりも精度を向上させえることを実証した。このことは、深い解析器が、精緻な確率モデルを用いるが、文法自体は現代言語学の文法理論には依拠しない浅い解析器よりも、処理速度の点でも優れた性能を示す、画期的な研究成果である。

３．GENIA コーパス：プロジェクト開始前に構築していた GENIA コーパスは、すでに生命科学のテキストマイニングの標準データとして国際的に多くの研究グループによって使われていた。この時点での GENIA コーパスは、品詞・句構造などの言語アノテーションを中心に、意味アノテーションとしては、遺伝子・たんぱく質名などの固有名アノテーションのみであった。

本プロジェクトでは、生命科学研究にとってより重要な生命機能、生命プロセスに対応する事象のアノテーションを意味アノテーションに追加し、また、それまで、生命科学文献では皆無であった同一参照表現などの文脈アノテーションを行った。

また、固有名アノテーションでも、生命科学のデータベースとのリンクを重要視し、データベースのエントリーとして重要な役割を果たしている GGP(Gene and Gene Product)のカテゴリを新たに追加することで、生命科学のテキストマイニング用の標準データとしての価値を向上させた。

事象アノテーションでは、生命科学の標準オントロジーである GO（Gene Ontology）から機能、生命プロセスのオントロジーを取り上げ、その中間階層を言語処理のための意味クラス（３６クラス）として設定し、アノテーション作業を行った。プロジェクト終了時には、１０００抄録、1 万文のアノテーションを完了した。また、より現実の応用を目指した意味アノテーションでは、米国ヴァージニア工科大学の生命科学者と共同して、かれらが興味をもつ伝染病(Infectious Disease)の分野の意味アノテーションをわれわれの事象クラスを拡張することで実行し、われわれの事象アノテーションの応用面での拡張可能性を実証した。プロジェクト終了後も、英国マンチェスター大学と国際的な製薬会社（アストラ・ジェネカ）と共同で英国の BBSRC のプロジェクトを推進し、ガン研究に必要な事象クラスの設定とアノテーション作業を推進している。

文脈アノテーションの同一参照表現アノテーションは、国立シンガポール大学のグループと共同して行ったものであるが、GENIA コーパスへの文脈アノテーションは、ハンガリーや米国ウィスコンシン大学、英国・マンチェスター大学のグループによっても実施、公開され GENIA コーパスの国際標準データとしての地位は、今回のプロジェクトにより飛躍的に向上した。GENIA の事象アノテーションは、2 回の国際ワークショップ(BioNLP09、BioNLP11)の訓練・テストデータとしてつかわれた。

４．文解析に基づく情報抽出技術：　複数個の固有名の間の関係を認識する研究、あるいは、それらが関与する事象認識は、新聞記事からの情報抽出研究で米国を中心に活発に研究されてきた。ただ、認識すべき関係が、会社とその社長といった比較的局所的な言語表現（同格表現、前置詞句など）に現れる関係とは別に、関係が動詞として明示的に現れるような事象の認識、あるいは、出来事が複数文で記述されるようなプロセスとその関与者間の関係の認識技術の研究は、その技術の難しさだけでなく、この種の関係の定義自体の難しさから、あまり取り組まれてこなった。これに対して、本プロジェクトでは、この種の関係概念や事象概念の把握と理解が、言語を理解するシステムにとって不可欠であると考え、関係や事象が言語外の対象分野で明確に定義できる生命科学の分野で、この種の技術の系統的な研究をおこなった。

分子生物学の分野では、テキストから認識すべき事象や生命プロセスは、この分野の知識、オントロジーにより明確に定義される。また、関係や事象がどのように言語表現されるかは、項目３の GENIA コーパスの事象アノテーションにより与えられる。

この種の関係が動詞を通して表現されることが多く、同格など名詞句内での局所的表現を超えた、大域的な表現をとることから、従来の BOW(Bag of Words)の手法の限界はあきらかである。このことから、われわれは項目１、項目 2 で開発した深い文解析の手法を基盤とした技術を開発した。ただ、関係が動詞中心に表現される場合であっても、名詞句内の構造、非明示的な表現からの推論といった人間の事象認識と同じ能力を持たせるためには、深い文構造だけでなく、これらの表現の変容を取り扱い、呼び出し率（Recall）を向上させるためには、機械学習手法を持ち込む必要がある。これらのことから、本プロジェクトでは、深い文解析結果を入力素性とし、BOW からのほかの組成と組み合わせて認識を行う手法を開発してきた。このような手

法は、単なる理論モデルにとどまらず、膨大な素性（個々の単語、その構造的な組み合わせなど）によって張られる、疎な空間を取り扱うための効率的な計算手法が伴わなければならない。われわれの開発した SVM や CRF などのプログラムは、世界でも有数の処理効率を持つものであり、プロジェクト終了後にも多くの研究グループにより使われている。

われわれの開発した事象認識システムのひとつであるたんぱく質相互作用の認識タスクで、５つの代表的なテストデータいずれもに対して、世界最高の認識精度を達成し、もっとも代表的な AIMed コーパスでは F 値 64%を達成した。また、GENIA コーパスを使った事象認識（BioNLP09）でも世界 3 位（24 チーム中)、その後の改良により世界最高の性能を示すシステムを開発した。これらの研究は、単に性能的に世界最高を達成したことにとどまらず、事象認識と文解析、事象の種類とそれに適合した処理モデル、事象認識と推論(TE, text entailment)の相互関係など、今後の研究を行うための貴重な知見をもたらした。

４．大規模なテキスト処理のための計算機環境の構築： これまでのシュミレーションを中核にした計算科学での並列処理では、大規模な計算環境による計算処理に重点が置かれていた。これに対して、テキスト処理では、処理単位（文、論文など）ごとに処理が可能であり、個々の処理単位ごとの計算は独立度が高く個々に実行可能であり、それほど大きな計算とはならない。たとえば、文単位の処理では、文解析が比較的重い処理であるが 1 秒未満の処理であり、固有名認識・事象認識などの処理を加えても 2 秒を超えることはなく、並列の粒度は粗い。しかしながら、言語処理では、テキストの文単位への分割、文の単語分割、品詞付け（POS Tagging)、文解析、固有名認識、事象認識といった多くのモジュールがパイプライン的に結合され、それぞれの処理が、たとえば、MEDLINE テキストベースの処理では、2 千万抄録（2 億超文）という膨大な量の処理単位に対して実行される必要がある。

このため、言語処理ではワークフローの管理と個々の処理単位が作り出す中間的な処理結果を蓄積するファイルシステムの効率的な構築が、処理資源の有効活用の鍵となる。また、研究集団が効率よく研究を進めていくためには、実験段階での比較的規模の小さな処理から本格処理の大規模テキスト集合の処理段階までのスムーズな移行、他グループとのプログラムや言語資源の共用など、多様な計算環境の使用が最小の労力で可能なことが不可欠である。

このようなことから、本プロジェクトでは、大規模なテキスト集合処理のための処理系として，(1) ワークフローシステム，(2)拠点をまたがった計算機間でも直ちにファイルを共有できる,アドホックな並列ファイルシステム，および，(3) ワークフローを構成するタスクのファイルアクセスを抽出し,タスク間の依存関係や通信量を解析・可視化するツールを設計し、実装した。

このワークフローシステム(GXP)は、make という,Unix 上で広く使われているシステムをそのまま大規模な分散環境で実行できるもので，e-Science シンポジウムで発表し、その後委員会より論文誌への投稿を推薦を受けるなど(142 件中 14 件)、国際的にも高い評価を得た。 また、GXP は、本プロジェクトのための単なる研究用のシステムとして使われただけでなく、英国マンチェスター大学の国立テキストマイニングセンター(NaCTeM)が運用するサービスのインデキシング処理にも使われている。本プロジェクトでは、Medline アブストラクト全件(2010 年 5 月当時)から、生命科学にとって重要な事象を認識し、それを索引構造に反映するシステムに用いられた。これらの処理は、 いずれも 8000 以上の並列度(CPU コア数)を用いた大規模なもので、複雑なワークフローが安定して高速な処理に使えることを実証した。

テキスト処理のようなデータ中心の大規模ワークフローの実行においては、ファイルアクセスの局所性を保ちつつ負荷分散を達成するためのスケジューリングが重要である。このため、GXP という基盤ソフトウェアを開発するだけでなく、ファイルアクセスパターンを抽出するツールを設計・実装し，実際のワークフローアプリケーションから実データ処理から収集し、ファイルアクセスの実態を可視化することで、最適なストラテジーをワークフローに反映する総合システムを開発、これを用いて、 ファイルアクセスの局所性を考慮したワークフロースケジューラを開発，信頼性の高いワークフローの実行系を構築することに成功した。

５．実応用システムの開発：言語の意味・知識は、それ自体は各種の応用システムを開発するための基盤技術であり、その効果は、応用システムを構築することにより、実証されなくてはならない。本プロジェクトでは、研究が抽象的になり、技術としての有効性を失うことを避けるために、常に応用システムの中で技術の有効性を確認することを行ってきた。このために、意味に基づく知的検索システム（MEDIE)、高品質な機械翻訳システムを構築してきた。知的検索システム MEDIE は事実検索という、これまでのシステムにはなかった機能を持つことで、製薬会社や科学出版を行う出版社から注目された。また、深い

解析器の出力を使った機械翻訳は、英日機械翻訳に適用され、これまでのシステムのパフォーマンスを大幅に向上させることに成功した。この成果は、プロジェクト終了後、民間会社との共同研究に引き継がれることとなった。

５．主な発表論文等

以下のリストには、プロジェクト成果の論文であるが、プロジェクト終了後に出版されたものは含めていない。

〔雑誌論文〕
（国際論文誌 28 件、国際会議 84 件）

[国際論文誌]

[1]Okazaki, Naoaki, Sophia Ananiadou, Jun'ichi Tsujii. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. Bioinformatics. Oxford University Press, 2010.

[2]Riedel, Sebastian, Rune Sætre, Hong-Woo Chun, Toshihisa Takagi, Jun'ichi Tsujii. Bio-Molecular Event Extraction with Markov Logic. In Jin-Dong Kim (Eds.), Computational Intelligence. Special Issue. Edmonton, Alberta, Canada T6G 2E8, 2010.

[3]Miwa, Makoto, Rune Sætre, Jin-Dong Kim, Jun'ichi Tsujii. Event Extraction with Complex Event Classification Using Rich Features. Journal of Bioinformatics and Computational Biology (JBCB). 8(1). pp. 131-146, February 2010.

[4]Yu, Kun, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, Yaozhong Zhang, Kiyotaka Uchimoto, Junichi Tsujii. Comparison of Chinese Treebanks for Corpus-oriented HPSG Grammar Development. Journal of Natural Language Processing (Special Issue on Empirical Methods for Asian Language Processing). April 2010.

[5]Wu, Xianchao, Takuya Matsuzaki, Jun'ichi Tsujii. Improve Syntax-based Translation Using Deep Syntactic Structures. Journal of Machine Translation (Special Issue: Pushing the frontiers of SMT). 24(2). pp. 141-157, Springer, 2010.

[6]Sætre, Rune, Kazuhiro Yoshida, Makoto Miwa, Takuya Matsuzaki, Yoshinobu Kano, Junichi Tsujii. Extracting Protein-Interactions from Text with the Unified AkaneRE Event Extraction System. Transactions on Computational Biology and Bioinformatics (TCBB), BioCreative II.5 Special Issue. 7. pp. 46pp, IEEE/ACM, 2010.

[7]Tsunakawa, Takashi, Naoaki Okazaki, Xiao Liu, Jun'ichi Tsujii. A Chinese-Japanese Lexical Machine Translation through a Pivot Language. ACM Transactions on Asian Language Information Processing. 8(2). pp. 9:1-9:21, May 2009. ISSN: 1530-0226.

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

[9]Miwa, Makoto, Rune Sætre, Yusuke Miyao, Jun'ichi Tsujii. Protein-Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers. International Journal of Medical Informatics. 78(12). pp. e39-e46, April 2009. Mining of Clinical and Biomedical Text and Data Special Issue.

[10]Kim, Jin-Dong, Tomoko Ohta, Jun'ichi Tsujii. Corpus annotation for mining biomedical events from lterature. BMC Bioinformatics. 9(1). pp. 10, BioMed Central, January 2008. ISSN 1471-2105.

[11]Oda, Kanae, Jin-Dong Kim, Tomoko Ohta, Daisuke Okanohara, Takuya Matsuzaki, Yuka Tateisi, Jun'ichi Tsujii. New challenges for text mining: Mapping between text and manually curated pathways. BMC Bioinformatics. 9(Suppl 3). pp. S5, BioMed Central, April 2008. ISSN 1471-2105.

[12]Tsuruoka, Yoshimasa, Jun'ichi Tsujii. Sophia Ananiadou. FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics. 24(21). pp. 2259-2260, November 2008.

[13] Miyao, Yusuke, Jun'ichi Tsujii. Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics. 34(1). pp. 35-80, MIT Press, March 2008.

[国際会議での発表論文]

[1] Taura, Kenjiro, Takuya Matsuzaki, Makoto Miwa, Yoshikazu Kamoshida, Daisaku Yokoyama, Nan Dun, Takeshi Shibata, Choi Sung Jun, Jun'ichi Tsujii. Design and Implementation of GXP make---a Workflow System Based on Make. 2010 IEEE 6th International Conference on e-Science, pp. 214-221, December 2010.

[2]Miwa, Makoto, Sampo Pyysalo, Tadayoshi

Hara, Jun'ichi Tsujii. Evaluating Dependency Representation for Event Extraction. 23rd COLING. pp. 779-787, August 2010.
[3]Miwa, Makoto, Yusuke Miyao, Rune Sætre, Jun'ichi Tsujii. Entity-Focused Sentence Simplification for Relation Extraction. 23rd COLING. pp. 788-796, August 2010.
[4]Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. A Simple Approach for HPSG Supertagging Using Dependency Information. 11th NAACL-HLT'10. pp. 645-648. June 2010.
[5]Wu, Xianchao, Takuya Matsuzaki, Jun'ichi Tsujii. Fine-Grained Tree-to-String Translation Rule Extraction. 48th ACL. pp. 325-334. July 2010.
[6] Dun, Nan, Kenjiro Taura, Akinori Yonezawa. ParaTrac: A Fine-Grained Profiler for Data-Intensive Workflows. 19th ACM HPDC 2010, pp. 37-48, June 2010.
[7] Shibata, Takeshi, SungJun Choi, Kenjiro Taura. File-Access Patterns of Data-Intensive Workflow Applications and their Implications to Distributed Filesystems. 3rd DIDC 2010, pp. 746-755, June 2010.
[8]Hanaoka, Hiroki, Hideki Mima, Jun'ichi Tsujii. A Japanese Particle Corpus Built by Example-Based Annotation. LREC2010. pp. 1876-1880, May 2010.
[9]Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. Forest-guided Supertagger Training. 23rd COLING. pp. 1281-1289, 2010.
[10]Yu, Kun, Junichi Tsujii. Bilingual Dictionary Extraction from Wikipedia. Proceedings of Machine Translation Summit XII. 2009.
[11]Yu, Kun, Junichi Tsujii. Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. NAACL HLT 2009. pp. 121-124, 2009.
[12]Miwa, Makoto, Rune Sætre, Jin-Dong Kim, Jun'ichi Tsujii. Event Extraction with Complex Event Classification using Rich Features. In the 3rd International Symposium on Languages in Biology and Medicine (LBM 2009). pp. 11--19, November 2009. (Honorable Mention Award).
[13]Matsubayashi, Yuichiroh, Naoaki Okazaki, Jun'ichi Tsujii. A Comparative Study on Generalization of Semantic Roles in FrameNet. ACL-IJCNLP2009. pp. 19-27, August 2009.
[14]Wu, Xianchao, Takuya Matsuzaki, Naoaki Okazaki, Yusuke Miyao, Jun'ichi Tsujii. The UOT System: Improve String-to-Tree Translation Using Head-Driven Phrase Structure Grammar and Predicate-Argument Structures. IWSLT 2009. pp. 99-106, December 2009.
[15]Sun, Xu, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, Jun'ichi Tsujii. A Discriminative Latent Variable Chinese Segmenter with Hybrid Word/Character Information. NAACL-HLT'09. Boulder, Colorado, pp. 56-64, 2009.
[16]Yao-zhong Zhang, Takuya Matsuzaki, Jun'ichi Tsujii. HPSG Supertagging: A Sequence Labeling View. 11th IWPT'09. pp. 210-213, 2009.
[17]Sun, Xu, Naoaki Okazaki, Jun'ichi Tsujii. Robust Approach to Abbreviating Terms: A Discriminative Latent Variable Model with Global Information. ACL. Singapore, pp. 905-913, 2009.
[18]Tsuruoka, Yoshimasa, Jun'ichi Tsujii, Sophia Ananiadou. Fast Full Parsing by Linear-Chain Conditional Random, EACL. pp. 790-798, April 2009.
[19]Miyao, Yusuke, Jun'ichi Tsujii. Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes. EMNLP 2009. Singapore, pp. 1328-1337, August 2009.
[20]Wu, Xianchao , Okazaki, Naoaki, Tsujii, Jun'ichi. Semi-Supervised Lexicon Mining from Parenthetical Expressions in Monolingual Web Pages. Human Language Technologies: NAACL. Boulder, Colorado, pp. 424-432, 2009.
[21]Sun, Xu, Takuya Matsuzaki, Daisuke Okanohara, Jun'ichi Tsujii. Latent Variable Perceptron Algorithm for Structured Classification. IJCAI. Los Angeles, pp. 1236-1242, 2009.
[22] Hara, Tadayoshi, Yusuke Miyao, Jun'ichi Tsujii. Effective Analysis of Causes and Inter-dependencies of Parsing Errors. IWPT-09 Paris, France, pp. 180-191, October 2009.
[23]Uematsu, Sumire, Jun'ichi Tsujii. Evaluating Contribution of Deep Syntactic Information to Shallow Semantic Analysis. IWPT' 09. pp. 85-88, 2009.
[24]Sun, Xu, Jun'ichi Tsujii. Sequential Labeling with Latent Variables: An Exact Inference Algorithm and An Efficient Approximation. 12th EACL 2009. Athens, Greece, pp. 772-780, 2009.

〔学会発表〕(計 48 件)

〔図書〕(計 3 件)

[1]Kim, Jin-Dong, Jun'ichi Tsujii. Corpora and their Annotation. In Sophia Ananiadou, John McNaught, (Eds.), Text Mining for Biology and Biomedicine. 46 Gillingham Street, London SW1V 1AH UK, Artech House, 2006. ISBN 1-58053-984-X.

[2] Hara, Tadayoshi, Yusuke Miyao, Jun'ichi Tsujii. Evaluating the Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. In Harry Bunt, Paola Merlo, Joakim Nivre (Eds.), Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing. Text, Speech and Language Technologypp. pp. 253-272, Springer, October 2010.

[3]Matsuzaki, Takuya, Yusuke Miyao, Jun'ichi Tsujii. Probabilistic Context-Free Grammars with Latent Annotations. In Srinivas Bangalore and Aravind K. Joshi (Eds.), Supertagging - Using Complex Lexical Descriptions in Natural Language Processing. pp. 337-354, MIT Press, March 2010.

〔その他〕

プロジェクト
http://www-tsujii.is.s.u-tokyo.ac.jp/aNT/

研究成果を反映したサービスを行う英国マンチェスター大学、国立テキストマイニングセンター
http://www.nactem.ac.uk/pathtext/

６．研究組織
(1)研究代表者
辻井潤一 (TSUJII JUNICHI)
東京大学・大学院情報理工学研究科・教授
研究者番号：20026313
(2)研究分担者
米澤明憲 (YONEZAWA AKINORI)
東京大学・大学院情報理工学研究科・教授
研究者番号：00133116
(H18、H19：研究分担者)
田浦健次朗 (TAURA KENJIRO)
東京大学・大学院情報理工学研究科・准教授
研究者番号：90282714
(H20→H22：連携研究者)
宮尾祐介 (MIYAO YUSUKE)
東京大学・大学院情報理工学研究科・助教
研究者番号：00343096
(H20→H22：連携研究者)
松崎拓也 (MATSUZAKI TAKUYA)
東京大学・大学院情報理工学研究科・助教
研究者番号：40463872
(H19：研究分担者)
(H20→H22：連携研究者)
(3)研究協力者
狩野芳伸 (KANO YOSHINOBU)
東京大学・大学院情報学環・特任研究員
大田朋子 (OHTA TOMOKO)
東京大学・大学院情報学環・特任研究員
Saetre, Rune (SAETRE RUNE)
東京大学・大学院情報学環・特任研究員
柴田剛志 (SHIBATA TAKESHI)
東京大学・大学院情報学環・特任研究員
三輪 誠 (MIWA MAKOTO)
東京大学・大学院情報学環・特任研究員
(H20→H22)
Pyysalo, Sampo Mikael (PYYSALO SAMPO MIKAEL)
東京大学・大学院情報学環・特任研究員
金 進東 (KIM JIN-DONG)
東京大学・大学院情報学環・特任講師
(H21)
Sagae, Kenji (SAGAE KENJI)
東京大学・大学院情報理工学系研究科・特任研究員
(H19→H21)
Sagae T., Alicia (SAGAE T. ALICIA)
東京大学・大学院情報理工学系研究科・リサーチアシスタント
(H19)
王 向莉 (WANG XIANGLI)
東京大学・大学院情報理工学系研究科・特任研究員
(H20→H21)
綱川隆司 (TSUNAKAWA TAKASHI)
東京大学・大学院情報理工学系研究科・特任研究員
(H20)
原 忠義 (HARA TADAYOSHI)
東京大学・大学院情報学環・特任研究員
(H22)