科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　５　日現在

機関番号：　１４６０３
研究種目：　基盤研究(A)
研究期間：　2011～2013
課題番号：　２３２４００２０

研究課題名（和文）大域情報を利用した同時処理による自然言語解析手法の研究

研究課題名（英文）Joint Natural Language Processing with Global Information

研究代表者
松本　裕治（Matsumoto, Yuji）

奈良先端科学技術大学院大学・情報科学研究科・教授

研究者番号：１０２１１５７５

交付決定額（研究期間全体）：（直接経費）　　29,600,000円、（間接経費）　　8,880,000円

研究成果の概要（和文）：自然言語処理の基本的な解析法である単語分かち書き、品詞解析、統語解析、述語項構造解析に対して、広い情報および複数の解析の同時処理に関する研究を行った。
英語に関して、品詞解析と統語解析をつなぐ情報として機能として複単語表現の収集とそれを用いた品詞解析および統語解析の性能向上を実現した。統語解析について、トップダウンとボトムアップの情報の同時利用が可能な解析手法や複数の解候補の大域的な違いを同時に考慮するため統語森から最適解を選択する手法を提案した。
述語項構造解析について、様々な素性とそれらを同時利用する方法の提案、および、Markov Logic Networksを利用する手法を提案した。

研究成果の概要（英文）：We investigated the usage of global information for joint processing of natural language processing such as word segmentation, part-of-speech tagging, parsing and predicate argument structure analysis.
For English, we collected multi-word expressions that act as functional words and implemented part-of-speech tagging and syntactic analysis using them. We also proposed various syntactic processing methods, one that utilizes both top-down and bottom-up information jointly and another that keeps multiple candidates in the form of parse forest and uses re-ranking algorithm to select the best answer.
For predicate argument structure analysis, we proposed a method that utilizes various types of features in a joint framework and another method that used Markov Logic Networks to incorporate various constraints jointly.

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　形態素解析　統語解析　述語項構造解析　同時処理　大域情報　言語資源

１．研究開始当初の背景

自然言語解析は，単語分かち書きと品詞付与を行う形態素解析，文節・固有表現・基本句などのチャンキング，統語解析，述語項構造解析，共参照解析など，多段階の処理があり，それぞれ独立に解析アルゴリズムが研究されてきた．個々の処理システムはある程度の精度を達成しているものの，大きく２の問題点があった．一つは，形態素解析や統語解析など文に対して構造（品詞列，木構造など）を出力するタスクにおいて，内部の部分構造を決定するためにわずか数単語以内の狭い範囲の情報しか利用しない場合が多いこと，そして，二つめは，全体的な最適化が個別のタスクに限定されており，品詞の決定や統語構造の決定を個別の処理として行うことである．例えば，品詞情報は統語解析には重要な情報だが，その反面，品詞タグ付けのエラーが統語解析には悪い影響を与える．逆に，距離の離れた単語や表現の間の係り受けなどの統語情報がもし入手可能であれば，より精密に品詞情報を決定できる可能性がある．このような広い大域的な情報（素性）を利用すること，そして，解析方向を固定せず，異なる階層の処理を同時に解くことが重要であると考えられるようになってきた．

大域的な情報の利用や同時処理の重要性は認識されており，様々な試みが提案されてきた．局所的な素性を用いた学習システムの上位 N 解を出力し，出力全体の大域的な素性を利用して re-ranking を行うが一つの手法であった．その他，サンプリングを繰り返すことによって，大域的な傾向を学習し，統語解析の性能向上を実現する方法も提案された．また，Markov Logic Networks(MLN)や整数線形計画法(ILP)を利用して大域的な制約を記述する試みが行われ，その効果が示されてきた．N-best に基づく手法は，N-best 解に正しい解が含まれる保証がないこと，MLN や ILP に基づく手法は，制約を記述する述語や変数の選択，制約としての論理式や一次式の設計が必ずしも自明でないこと，また，処理効率の問題もあった．また，具体的なタスクについては，依存構造解析における依存関係を表わす枝の重みの学習が，単純な枠組みから２次さらに３次というように複数の枝の素性の組み合わせを見ることのできるモデルへと拡張が進んできた．

同時処理については，Sutton らによる Factorial CRF が品詞タグ付けと固有表現認識の２つの線形ラベリング問題を同時に解くのに適用できることがよく知られていた．しかし，このモデルは循環構造をもつグラフモデルであり，近似的な最適化を行わざるを得ない．また効率的に大規模データへの適用に困難がある．系列ラベリングをパイプラインとして重ねた形での同時処理の提案がいくつか見られるが，その中でも，Bunescu らは，前段の全解から素性の出現確率を計算し，確率値付きの素性を後段の処理に渡すことにより，前段の最適解だけを後段に渡す単純なパイプラインよりも性能が向上することを示した．このように，様々な処理を独立した処理として個別に行うことの限界が意識されてきた状況が背景にあった．

２．研究の目的

自然言語解析は，形態素解析，統語解析，意味解析など多数の段階を経て行われるが，これらを段階的に積み重ねた実システムでは，前段の誤りが後段に悪影響を起こすことが知られている．これを回避するため，前段と後段の処理を同時に行うことが考えられる．その際に，局所的な狭い情報のみを用いるのではなく，大域的情報の利用しつつ同時処理を可能にする自然言語解析法を探求し，より高精度な自然言語解析システムを実現することを本課題の目的とした．また，自然言語解析手法の研究と並行して，学習に用いるために種々の情報が付与されたタグ付きコーパスの構築と，タグ付きコーパスの管理ツールの開発を行うことを目的とした．

３．研究の方法

従来個別の処理として取り組まれることが多かった形態素解析，統語解析，述語項構造解析などの自然言語解析のタスクを互いの情報を利用する同時処理へ，そして，その拡張を通じてより大域的な素性を利用する枠組みへ拡張する．まずは，同時処理，大域素性の利用を目指した従来研究をベースラインとして，我々が取り組む種々の自然言語処理タスクに適用する．次に，個々のタスクについて，それ以外のタスクの正しい結果が得られているという前提で，上位タスクの情報（特に大域的素性）がいかに利用可能か検討する．これらが，本研究の目標の下限，上限としてのベースラインとなるため，その上限を目指した同時処理の実現を各タスクの目標とする．これと並行して，アノテーションの順序を固定せず，他のレベルのアノテーション結果を制約として有効に利用するアノテーション支援ツールおよびアノテーションが施されたコーパスの管理ツールを構築する．

４．研究成果

形態素解析およびその上位の固有表現認識や統語解析との同時最適化を実現するため，機械学習に基づく形態素解析と浅い統語解析のパラメータを同時学習する手法を提案した．

係り受け解析と並列構造解析の同時処理については，並列構造を知ることが係り受け解析の誤り訂正にどの程度貢献できるかを実験によって確認し，効果的なアノテーション作業への道筋を明らかにした．また，係り受け解析に基づく統語解析の高性能化のため，ボトムアップ的な情報とトップダウン的情報の同時利用に関する研究を行い，両者の

情報を融合できる新しい手法の提案を行った．また，遷移に基づく決定的な係り受け解析の誤りの伝搬を除去するため，後戻りなしに解析誤りを修正しつつ解析を継続する手法を，グラフデータマイニングのアルゴリズムを利用して実現する方法について研究し，従来手法の欠点を補う手法を提案した．

述語項構造解析については，大規模コーパスから得られる述語の項に関する情報の利用について研究を行った．また，複数の項の値の同定を同時に行なう手法について研究を行った．中国語の述語項構造解析を対象に，適用可能な様々な素性，特に，チャンキングや依存構造解析によって得られる素性を同時利用し，

単語の使用文脈の関係に基づく意味的類似度の計算については，文脈ベクトルに基づく類似度として類似度行列上のカーネルを用いているが，ベクトルが高次元の場合に生じる問題としてハブとなる点の存在が問題になることが明らかになった．その性質の解明について研究を行った．

これらの研究の基本データとなるタグ付きコーパスを構築するため，日本語係り受けと述語項構造のアノテーション作業を行った．また，タグ付きコーパスを格納するコーパス管理ツールに次のような機能拡張を行った．一つは，係り受け解析と並列構造解析を重ねて表示するインタフェースの構築，もう一つは，このツールから係り受け解析システムを呼び出して文あるいは文の一部を再解析させる機能である．

言語学習者による誤りを含む可能性のある文の誤り検出・修正を可能にする言語解析手法に関する研究を行った．英語の綴り誤り訂正と品詞解析を同時に行う手法を提案し，これらの処理を逐次的に行う従来手法より誤り訂正効率を向上できることを示した．また，学習者の作文に生じる動詞のテンス・アスペクトに関する誤りを修正するため，文書全体の大域的な情報を利用する方法を提案し，個々の文に現れる動詞の誤りを個別に検出・訂正するよりも良好な結果が得られることを示した．

単語を言語解析の基本要素とする従来の文解析法の性能を向上させるため，複数の語がまとまって役割や意味をもついわゆる多単語表現の英語辞書の構築と，それを用いた英語の品詞解析の方法を提案した．機能語として働く多単語表現辞書，および，英語の共通データである Penn Treebank への多単語表現のアノテーションコーパスの構築を行った．同時に，この辞書とコーパスを利用して，英語の品詞解析システムの試作を行い，解析精度の向上を確認した．

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 5 件）

① 林部祐太，小町守，松本裕治，述語と項の位置関係ごとの候補比較による日本語述語項構造解析，自然言語処理，査読有，Vol.21，No.1，pp.3-25，2014.
② 吉川克正，浅原正幸，松本裕治，Markov Logic による日本語述語項構造解析，自然言語処理，査読有，Vol.20，No.2，pp.251-271，2013.
③ 原一夫，鈴木郁美，新保仁，松本裕治，文法的・意味的共起を利用した単語類似度の計算，人工知能学会論文誌，査読有，Vol. 28，No. 4，pp.379-390，2013.
④ Katsuhiko Hayashi，Shuhei Kondo，and Yuji Matsumoto，Efficient Stacked Dependency Parsing by Forest Reranking，Transactions of the Association for Computational Linguistics，査読有，Vo.1，pp.139-150，May 2013.
⑤ Katsumasa Yoshikawa，Sebastian Riedel，Tsutomu Hirao，Masayuki Asahara，and Yuji Matsumoto，Coreference Based Event-Argument Relation Extraction on Biomedical Text，Journal of Biomedical Semantics，査読有，Vol.2，2011．DOI: 10.1186/2041-1480-2-S5-S6

〔学会発表〕（計 15 件）

① Masashi Tsubaki，Kevin Duh，Masashi Shimbo，and Yuji Matsumoto，Modeling and Learning Semantic Co-Compositionality through Prototype Projections and Neural Networks，Proceedings of the Conference on Empirical Methods in Natural Language Processing，査読有，2013 年 10 月 19 日，Seattle，USA.
② Ikumi Suzuki，Kazuo Hara，Masashi Shimbo，Marco Saerens，and Kenji Fukumizu，Centering Similarity Measures to Reduce Hubs，Proceedings of the Conference on Empirical Methods in Natural Language Processing，査読有，2013 年 10 月 19 日，Seattle，USA.
③ Yanyan Luo，Kevin Duh，and Yuji Matsumoto，What information is helpful for dependency based semantic role labeling，Proceedings of the International Joint Conference on Natural Language Processing，査読有，2013 年 10 月 16 日，Nagoya，Japan.
④ Xiaodong Liu，Fei Cheng，Yanyan Luo，Kevin Duh，and Yuji Matsumoto，A Hybrid Chinese Spelling Correction Using Language Model and Statistical Machine Translation with Reranking，Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing，査読

有，2013年10月14日，Nagoya，Japan.

⑤ Yutaro Shigeto，Ai Azuma，Sorami Hisamoto，Shuhei Kondo，Tomoya Kose，Keisuke Sakaguchi，Akifumi Yoshimoto，Frances Yung，Yuji Matsumoto，Construction of English MWE Dictionary and its Application to POS Tagging，Proceedings of the 9th Workshop on Multiword Expressions，査読有，2013年6月13日，Atlanta，USA.

⑥ Keisuke Sakaguchi，Tomoya Mizumoto，Mamoru Komachi，and Yuji Matsumoto，Joint English Spelling Error Correction and POS Tagging for Language Learners Writing，Proceedings of International Conference on Computational Linguistics，査読有，pp.2357-2374，2012年12月10日，Mumbai，India.

⑦ Yuji Matsumoto，Things between Lexicon and Grammar，invited talk，26th Pacific Asia Conference on Language Information and Computation，査読無，2012年11月8日，Legian，Indonesia.

⑧ Akihiro Inokuchi，Ayumu Yamaoka，Takashi Washio，Yuji Matsumoto，Masayuki Asahara，Masakazu Iwatate，and Hideto Kazawa，Mining Rules for Rewriting States in a Transition-based Dependency Parser，Proceedings of the 12th Pacific Rim International Conference on Artificial Intelligence，査読有，2012年9月5日，Sarawak，Malaysia.

⑨ Katsuhiko Hayashi，Taro Watanabe，Masayuki Asahara，and Yuji Matsumoto，Head-driven Transition-based Parsing with Top-down Prediction，Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics，査読有，pp.657-665，2012年7月10日，Jeju，Korea.

⑩ Toshikazu Tajiri，Mamoru Komachi，and Yuji Matsumoto，Tense and Aspect Error Correction for ESL Learners Using Global Context，Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics，査読有，pp.198-202，2012年7月11日，Jeju，Korea.

⑪ 林部祐太，小町守，松本裕治，文脈情報と格構造の類似度を用いた日本語文間述語項構造解析，情報処理学会研究報告 第201回自然言語処理研究会，査読無，2011年5月17日，東京.

⑫ Yanyan Luo，Masayuki Asahara and Yuji Matsumoto，Dual Decomposition for Predicate-Argument Structure Analysis，Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering，査読有，2011年11月28日，Tokushima，Japan.

⑬ Katsumasa Yoshikawa，Masayuki Asahara and Yuji Matsumoto，Jointly Extracting Japanese Predicate-Argument Relation with Markov Logic，Proceedings of the 5th International Joint Conference on Natural Language Processing，査読有，2011年11月9日，Chiang Mai，Thailand.

⑭ Yuta Hayashibe，Mamoru Komachi and Yuji Matsumoto，Japanese Predicate Argument Structure Analysis Exploiting Argument Position and Type，Proceedings of the 5th International Joint Conference on Natural Language Processing，査読有，2011年11月9日，Chiang Mai，Thailand.

⑮ Ai Azuma，Yuji Matsumoto，Multilayer sequence labeling，Conference on Empirical Methods in Natural Language Processing，査読有，2011年7月28日，Edinburgh，Scotland，UK.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
なし

６．研究組織
(1)研究代表者
松本 裕治（Matsumoto，Yuji）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：１０２１１５７５

(2)研究分担者
新保　仁（Shimbo, Masashi）
奈良先端科学技術大学院大学・情報科学研究科・准教授
研究者番号：９０３１１５８９

Kevin Duh（Duh, Kevin）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：８０６３７３２２

小町　守（Komachi, Mamoru）
首都大学東京・システムデザイン研究科・准教授
研究者番号：６０５８１３２９