様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　１４６０３
研究種目：　基盤研究(A)（一般）
研究期間：　2014～2016
課題番号：　２６２４００３５

研究課題名（和文）構文パターン獲得と並列構造解析による統語的依存構造解析の高精度化

研究課題名（英文）Improvement of Syntactic Dependency Analysis by Syntax Pattern Acquisition and Coordinate Structure Analysis

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）　　26,200,000円

研究成果の概要（和文）：自然言語の統語解析の高性能化を目指し，特に英語の複単語表現と複文パターンの収集とコーパスへのアノテーション，および，並列構造解析手法の開発とコーパスへのアノテーションを行った．
英語の複単語表現については，主として機能語相当の表現の収集とPenn Treebankへのアノテーションを行い，複単語表現と依存構造の同時解析手法を実装した．また，複文パターンの収集と例文へのアノテーションを行った．
並列構造解析については，並列構造の範囲同定と依存構造解析を同時に行うアルゴリズムを提案し，プロトタイプシステムの実装を行った．

研究成果の概要（英文）：Aiming at an improvement of syntactic dependency analysis of English sentences, we compiled an English multi-word expression (MWE) lexicon and an MWE-annotated corpus. We also developed an algorithm for coordination structure analysis on top of dependency structure analysis.
For English MWEs, we collected English functional multi-word expressions and performed MWE annotation on Penn Treebank. We also implemented an MWE-aware English dependency parsing algorithm.
  We collected English complex sentence patterns and made annotation on made annotation on complex sentence examples.
For coordinate structure analysis, we proposed a joint algorithm of dependency and coordinate structure analysis and implemented a prototype system.

研究分野：　自然言語処理

キーワード：　自然言語処理　統語解析　並列構造　複文構造　依存構造解析　機械学習　アノテーション　コーパス

１．研究開始当初の背景

自然言語文の統語解析は，自然言語文の基本的な解析技術であり，自然言語処理分野で最も多くの研究が行われてきた技術である．従来は，文を名詞句や動詞句など句(phrase)と呼ばれる文法的な塊を基本とし，それらがどのようにつながって言語的に正しい文が構成されるかを定義するいわゆる句構造文法を基礎とする構文解析アルゴリズムが研究されてきた．しかし，2000 年前後より，単語同士の係り受け（依存構造）関係を基本とする依存構造解析が次第に多く研究されるようになってきた．特に，国際会議 CoNLL(Conference on Natural Language Learning)の 2006 年および 2007 年の shared task として多言語の依存構造解析が取り上げられ, 10 か国語以上の学習データとテストデータが公開されることにより，急速に関心が高まるようになった．依存構造解析は，文を構成する単語間の係り受け（修飾）関係の解析であり，言語によらず統一的なアノテーションと解析手法を用いることができるのが特徴であり，上記の shared task でも，一つのプログラムを全言語のデータに適用してその精度を競うというものであった．

日本語の統語構造は，文節係り受けで表現することが最も自然であり，対象の単位が単語ではなく文節であることを除けば，依存構造解析そのものである．我々のグループでは，機械学習に基づく依存構造解析にいち早く取り組み，当時まだ広く使われていなかったサポートベクターマシンを学習アルゴリズムとして利用した文節係り受けシステムを開発し，その発展版を CaboCha というフリープログラムとして公開してきた．同様の手法を英語に対して適用した我々の研究は，機械学習に基づく依存構造解析の先駆的な研究の一つとして，この分野で広く引用される論文になっている．

依存構造解析の基本は，文中の 2 つの単語の間に依存関係があるかないかを判定することだが，2 つの単語の情報だけではそれを判定することはできず，如何に周囲の情報を考慮して，依存関係にあるかどうかの強さを判定する必要がある．2 つの単語間の 1 関係だけを見る初期の 1 次の（first-order）手法から始まり，2 つ以上の関係を同時に考慮する 2 次，さらに 3 次，場合によっては 4 次の情報を考慮しつつ効率の劣化を防ぐ方法が提案された．また，弱いモデルによって生成された複数解の比較を行う手法，あるいは，全体的な最適化を行うため整数線形計画法や双対分解法などを用いるなど，より広い文脈を考慮した解析法が提案され，当初の約 90%の係り受け精度が, 10 年をかけて最新の手法では 93%を越える精度を達成している．

一方で，過去 10 年の依存構造解析では，長文にかかわる 2 の大きな問題がほとんど解決されていない状態と言える．一つは，並列構造の問題，もう一つは複文構造の問題である．並列構造とは，同じ文法機能をもつ句同士が“ and, or ”などの接続詞によって並置される構造であり，短い名詞句や動詞句等の並列構造は問題ないが，例えば，“ Mary saw John yesterday and Bill today. ”の “ John yesterday ”と“ Bill today ”のように，句とは呼ぶことができない単語列の並列化や，「おじいさんは山へ柴刈に、おばあさんは川へ洗濯に行きました」の「おじいさんは...、」と「おばあさんは... 洗濯に」のような長い文節列の並列構造（部分並列と呼ばれる）の解析はほとんど未解決の問題であった．並列構造の範囲同定については，黒橋らの先駆的な研究があるが，並列構造の類似度は人手記述の規則により計算されていた．我々は類似度計算をデータからの学習によって行い，句構造解析アルゴリズムと組み合わせることにより，英語の並列構造の範囲同定において最高の性能を達成したが，それでも精度は 60%程度にとどまっていた．

もう一方の，複文構造に関しては，依存構造解析研究においては，これを明確に意識した研究は行われていない．依存構造解析は単語間の係り受け関係という一つの概念（関係にラベルを考えることはある）に基づくという意味でシンプルであり，言語に依存しない様々な一般的な解析手法が提案されてきた．しかし，単文内の比較的短い距離にある単語間と複文のように節を越えた長い距離にある単語間の係り受け関係を同等に扱うことには無理がある．複文を構成する構造は言語により異なると考えられるが，依存構造解析の精度を今後飛躍的に高めるためには，節と節を結びつける複文構造に関する文法的な知見を導入することは大きな意義がある．

２．研究の目的

並列構造と依存構造は，互いに影響をもつ現象であり，どちらか一方を先に処理するという単純な方法では高い精度を期待できない．我々は，両者を双対分解法によって融合することを試みたが，単純な適用ではわずかな精度向上しか達成できないことがわかった．一方，上に示したような複雑な並列構造を文法的に記述する理論として CCG(Combinatory Categorial Grammar)がある．CCG では，各単語に複雑な複合カテゴリを割り当てることができ，カテゴリ間の様々な演算を用いてほとんどすべての単語列に対して複雑なカテゴリを与える能力をもっている．これにより，CCG は並列構造を説明する能力を持つが，精度よく解析する能力をもつ訳ではない．

本提案では，係り受け解析と並列構造解析を同時に行う手法について検討する．その知見を経て，並列構造内の単語列がもつ依存関係の多様性を学習し，依存構造解析と以前我々が行った並列構造の範囲同定の手法との融合を実現する．

これとは平行に，英語の大規模コーパスに

現れる複文の構造を解析し，節と節や不定詞句など動詞を中心とする構造間の関係を網羅的に抽出し，共通の構造を抽出することによって複文の構文パターンの収集を行う．得られた構文パターンを用いて新たな長文を解析する際には，パターンの適用箇所の曖昧性や，複数の構文パターンの競合など解決すべき問題が予想され，これらの問題を解決しつつ，長文に対する高い解析精度をもった依存構造解析手法の研究を行う．複文パターンは複数の語の関連を記述したものと言え，いわゆる複単語表現の拡張と考えることができる．複文パターンの収集と並行して，英語の複単語表現，特に，統語解析に影響が大きい機能的な複単語表現の収集とコーパスへのアノテーションを行う．

複単語表現については，固定的な表現からはじめ，句動詞のように動詞と前置詞が分離して現れうる表現の抽出と，それらを格納する辞書管理ツールの開発を行う．

３．研究の方法

本提案は，自然言語文の統語解析のうち特に依存構造解析に焦点を当て，長文の解析を困難にしている並列構造と複雑な複文構造の問題を解決することを目標にしている．そのために，次の研究項目を想定する．

１）従来の句構造解析では説明できなかった部分並列などの複雑な並列構造の範囲同定と依存構造解析を同時に最適化する手法の研究

２）英語と日本語の長文に現れる複文構造を記述する構文パターンを大規模コーパスから半自動抽出し，管理可能な程度（数百程度）の構文パターンとしてまとめる．これを用いたトップダウンとボトムアップと融合した依存構造解析アルゴリズムの研究

３）単語や複単語表現の品詞や文法機能，構文パターンなどをデータベース化して格納する辞書管理ツール（これまで単語や固定的な複単語表現を格納するツールを開発している）の機能拡張

これら以外に，並列構造や複文構造をもつ英語コーパスへの依存構造のアノテーション作業（機械学習用データとして用いるため）を行う．

４．研究成果

自然言語の統語的依存構造解析の高性能化を目指し，特に英語の複単語表現と複文パターンの収集とコーパスへのアノテーション，および，並列構造解析手法の開発とコーパスへのアノテーションを行った．

英語の依存構造解析そのものの性能向上を目指す研究を行い，品詞よりも詳細な統語構造タグを各単語に対して予測することにより，解析性能向上を達成した[1)2)]．

長文等の複雑な文の解析に影響を及ぼす問題点について整理し[12)]，長文をトップダウンに分割することにより，統語解析の高速化を達成する手法を提案した[7)]．

英語の複単語表現については，種々の複単語表現の収集と Penn Treebank へのアノテーションを行った．具体的には，英語句構造の収集とコーパスへのアノテーション[4)]，英語の機能的複単語表現の収集と依存構造としてのアノテーション[11)]，修飾語を含みうる柔軟な複単語表現の収集と依存構造コーパスへのアノテーション[14)18)]，機能的複単語表現や固有表現を考慮した統語解析手法の提案とその評価[15)16)]を行った，また，日本語 Universal Dependency データへの複合辞情報付与に関する検討を行った[17)]．

これらとは別に，文の構文パターンを階層的に獲得する方法と獲得したパターンを言語モデルとして利用する方法について研究した[2)9)19)13)]．また，英語の複文例文に対し，複文パターンをアノテーションしたコーパスの構築を行った．

並列構造解析については，並列構造の範囲同定と依存構造解析を同時に行うアルゴリズムを提案した[5)6)]．また，Penn Treebank における並列構造の範囲を見直し，部分並列構造を考慮したアノテーションの修正作業を行った．これを利用して，並列構造の範囲同定を行う手法について検討を行った．

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　１件）

① Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto, Transition-based Dependency Parsing Exploiting Supertags, 査読あり, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.24, Issue 11, 2016, pp.2059-2068. (DOI: 10.1109/TASLP.2016.2598310)

〔学会発表〕（計１８件）

① 大内啓樹，Kevin Duh，進藤裕之，松本裕治，Supertag の曖昧性を考慮した依存構造解析，情報処理学会研究報告，第 218 回自然言語処理研究会，2014.9.1，首都大学東京南大沢キャンパス(東京都・八王子市)．

② Xiaoyi Wu and Yuji Matsumoto, A Hierarchical Word Sequence Language Model, 28th Pacific Asia Conference on Language, Information and Computation, 査読あり, 2014.12.12, Phuket(Thailand).

③ 俵雄貴，東藍，松本裕治，係り受け情報を利用した日本語形態素解析，情報処理学会研究報告，第 220 回自然言語処理研究会, 2015.1.19，九州大学医学部百年講堂(福岡県・福岡市)

④ 駒井雅之，進藤裕之，松本裕治，英語の

句動詞表現の同定とコーパス構築，言語処理学会第 21 回年次大会，2015.3.19，京都大学吉田キャンパス(京都府・京都市)．

⑤ 吉本暁文，新保仁，原一夫，松本裕治，並列構造解析に向けた依存構造解析アルゴリズムの拡張，情報処理学会研究報告，第 221 回自然言語処理研究会，2015.5.25，東北大学片平さくらホール(宮城県・仙台市)

⑥ Akifumi Yoshimoto, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto, Coordinate-aware Dependency Parsing, The 14th International Conference on Parsing Technology, 査読あり, 2015.7.22, Bilbao(the Basque Country)

⑦ Joseph Irwin and Yuji Matsumoto, CKY Parsing with Independence Constraints, The 14th International Conference on Parsing Technology, 査読あり, 2015.7.24, Bilbao(the Basque Country)

⑧ Masayuki Komai, Hiroyuki Shindo, Yuji Matsumoto, An Efficient Annotation for Phrasal Verbs using Dependency Information, The 29th Pacific Asia Conference on Language, Information and Computation, 査読あり, 2015.10.31, Shanghai(China).

⑨ Xiaoyi Wu and Yuji Matsumoto, An Improved Hierarchical Word Sequence Model Using Directional Information, The 29th Pacific Asia Conference on Language, Information and Computation, 査読あり, 2015.11.1, Shanghai(China).

⑩ Xiaoyi Wu and Yuji Matsumoto, An Improved Hierarchical Word Sequence Model Using Word Association, Statistical Language and Speech Processing, 査読あり, 2015.11.24, Budapest(Hungary).

⑪ Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto, Construction of an English Dependency Corpus incorporating Compound Function Words, the Tenth International Conference on Language Resources and Evaluation, 査読あり, 2016.5.26, Portorož(Slovenia)

⑫ Yuji Matsumoto, Parsing Complex Linguistic Constructions, The Eighth International Conference on Knowledge and Systems Engineering, Invited talk, 2016.10,7, Hanoi(Vietnam)

⑬ Xiaoyi Wu, Kevin Duh and Yuji Matsumoto, A Generalized Framework for Hierarchical Word Sequence Language Model, The 30th Pacific Asia Conference on Language, Information and Computation, 査読あり, 2016.10.29, Seoul(Korea)

⑭ Ayaka Morimoto, Akifumi Yoshimoto, Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto, Identification of Flexible Multiword Expressions with the Help of Dependency Structure Annotation, the Workshop on Grammar and Lexicon: interactions and interfaces, 査読あり, 2016.12.11, 大阪府立国際会議場(大阪府・大阪市)

⑮ 加藤明彦，進藤裕之，松本裕治，複単語表現を考慮した英語の依存構造解析モデリング，情報処理学会 第 229 回自然言語処理研究会，2016.12.22，NT 武蔵野研究開発センタ(東京都・武蔵野市)

⑯ 加藤明彦，進藤裕之，松本裕治，固有表現と複合機能語を考慮した MWE ベースの依存構造コーパス構築と解析，言語処理学会第 23 回年次大会，2017.3.14，筑波大学筑波キャンパス春日エリア(茨城県・つくば市)

⑰ 久保大輝，田中貴秋，進藤裕之，松本裕治，永田昌明，日本語 Universal Dependencies への複合辞情報付加の試み，日本語 Universal Dependencies への複合辞情報付加の試み，2017.3.14，筑波大学筑波キャンパス春日エリア(茨城県・つくば市)

⑱ 森元彩華，吉本暁文，加藤明彦，進藤裕之，松本裕治，依存構造情報を用いた柔軟な複単語表現の同定，言語処理学会第 23 回年次大会，2017.3.16，筑波大学筑波キャンパス春日エリア(茨城県・つくば市)

〔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〕
ホームページ等
http://cl.naist.jp/

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

進藤　裕之（SHINDO Hiroyuki）
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：２０７３４７８４

能地　宏（NOJI Hiroshi）
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：００７８２５４１

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

(3)連携研究者

（　　）

研究者番号：

(4)研究協力者

（　　）