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 24 年 4 月 6 日現在

機関番号：63801
研究種目：若手研究（Ｂ）
研究期間：2009～2011
課題番号：21700170
研究課題名（和文）系列アラインメントに基づく並列構造解析と統語解析の統合に関する研究

研究課題名（英文）Parsing with Coordinate Structure Analysis based on Sequence Alignment

研究代表者
原　一夫（HARA　KAZUO）
国立遺伝学研究所・生命情報・DDBJ 研究センター・研究員
研究者番号：30467691

研究成果の概要（和文）：
並列構造解析は自然言語処理の基礎技術の一つであるにもかかわらず，既存解析器の精度は高くない（英語の並列構造解析の精度は約 50%であった）．そこで，並列構造解析に特化した文法および系列アラインメントによる類似度測定法を土台とし，（構文解析で用いられる）CKY アルゴリズムを応用する方法を開発した．並列構造を構成する単語系列の範囲と階層構造を同時に決定するこの方法により，並列構造解析の精度を大きく改善することに成功した．

研究成果の概要（英文）：
For a given sentence in which coordination conjunctions such as "and" or "or" occur, detecting the scope of word sequences that the coordination conjunctions join together, as well as distinguishing the grammatical category of it, are deemed fundamental in natural language processing. However, even the state-of-the-art parsers can identify only around fifty percent of such coordination scopes and their grammatical category. To cope with the problem, we developed a methodology that employs a CKY algorithm based on the grammar we customize for coordination as well as a tunable similarity measures we define via sequence alignment and averaged perceptron, which improves performance very well.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,700,000	510,000	2,210,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	500,000	150,000	650,000
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，構文解析，機械学習，並列句解析，アラインメント

１．研究開始当初の背景

自然言語処理研究において，並列構造解析は困難な課題の一つであり，既存の優れた句構造解析器を使用しても解析誤りが多く発生する（図 1）．特に，医学/生物学分野の学術論文テキストを対象にするとき，誤りが頻出する．なぜなら，そこで主に記されるのは生命科学実験の問題設定および実験結果であり，これらは並列構造を用いて記述されやすいからである（典型的には，新規療法/仮説に基づく手法と既存療法/コントロール手法

との対比）．実際，GENIA treebank（500 の MEDLINE アブストラクトからなる句構造木コーパス）では，1 文につき約 1 つの割合で並列構造が出現するが，既存の句構造解析器による並列構造範囲同定の精度は約 50%でしかない．以上を動機として，本研究課題では並列構造解析精度の向上を第一の目的とした．

並列構造解析は自然言語処理の基礎技術の一つであるが，自然言語処理の基礎技術は，文を単語あるいは句の系列として扱う技術（以下，シャロー解析）と，文を木構造の形で表現する技術（以下，構文解析）に大まかに分類できる．それらが取り扱う代表的なタスクは，シャロー解析では，文を単語に区切ってその品詞を推定するタスク（形態素解析），単語を句としてまとめるタスク（チャンキング）であり，構文解析では，文中に現れる句を階層的にまとめあげるタスク（句構造解析），単語間の関係（修飾・被修飾等）の有無を判定するタスク（係り受け解析）である．

本研究課題で取り組んだ並列構造解析は，シャロー解析と構文解析の中間に位置する．すなわち，並列構造解析は，（並列構造を構成する）単語系列の範囲を検出する側面をシャロー解析と共有し，他方，文中に現れる複数の並列構造の階層構造を決定する部分を構文解析と共有する．

研究代表者はすでに，英語文を対象に，系列アラインメントを応用した手法を用い，並列構造構成要素間の類似度を測ることで，並列構造を構成する単語系列の範囲を同定する機械学習（averaged perceptron を用いた識別）による手法を開発し，既存解析器を上回る精度を得ることに成功していた．しかし，シャロー解析に準じるこの方法は，複数の並列構造が文に含まれる場合に対応できないため実用的ではない．この問題を解決するために，複数の並列構造が形成する階層構造を表現できる（並列構造解析に特化した）文法を考案し，系列アラインメント手法に加えて構文解析で用いられる CKY アルゴリズムを応用して，それら並列構造を構成する単語系列の範囲と階層構造を同時に決定する方法の提案を考えるに至った．

２．研究の目的

本研究課題においては，以下の３つを目標とした．

（１）並列構造解析の精度向上のための，並列構造の範囲と階層構造を同時に決定する方法の優位性を，大規模なベンチマークデータ（コーパス）を使用する実験で確認する．

（２）日本語文に対する並列構造解析の精度を向上させる．

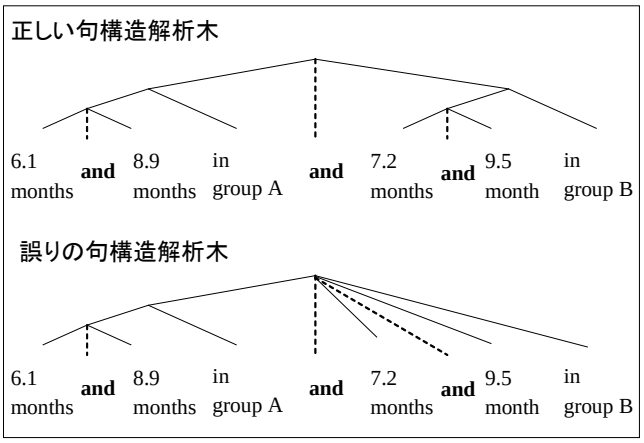

図 1: 並列構造を含む文は，既存の句構造解析器では正しく解析できない場合があることを示す例．

Median times to progression **and** median survival times were 6.1 months **and** 8.9 months in group A **and** 7.2 months **and** 9.5 months in group B" **の後半部分に対する正しい句構造解析木（上）と，句構造解析器（**[Charniak and Johnson, 2005; ACL]**）による出力結果（下）．**

（３）自然言語処理の他の基礎技術（チャンキング，係り受け解析，句構造解析）と並列構造解析の融合解析を行う．

並列構造の構成要素の多くは句を単位としていると考えられる．たとえば，“median times to progression and median survival times”では，“median times to progression” や “median survival times” は名詞句であるが，“median” や “median survival” は名詞句ではない．このことは，“median times to progression and median” や “median times to progression and median survival” よりも “median times to progression and median survival times” が並列構造として正しいことの証拠となる．このように，チャンキングは並列構造解析の精度向上に寄与すると考えられる．係り受け解析については，Eisner (1997; IWPT) のアルゴリズムに，並列構造構成要素間の類似性に関する（系列アラインメントをベースとする）素性の使用による類似度計算モジュールを追加することで，並列構造解析の誤りによる係り受け解析の精度低下を防ぐことが可能になると考えられる．さらに，句構造解析の精度向上は，並列構造構成要素間の類似性に関する（系列アラインメントをベースとする）素性を用い，句構造解析器のＮベスト出力のリランキングを行うことで実現できると見込まれる．

３．研究の方法

並列構造の範囲と階層構造を同時に決定する方法の優位性を，大規模なベンチマークコーパスを使用する実験で確認する．さらに，日本語並列構造解析を行う．その後，並列構

造解析とチャンキングとの融合解析を行う．実装には，C++言語を用いるが，単独の並列構造解析と，融合解析では，ソースコードの一部の共有を見込める．訓練/評価データとしては，英語では，GENIA treebank を使用する．このコーパスには，並列構造の範囲を明示的に示す“COOD”タグが存在するため，ほとんどそのまま訓練/評価データとして使用できる．日本語については，EDR コーパス（文法構造と意味構造が主に付与されたコーパス）を使用する．

４．研究成果

英語並列構造解析については，以前に提案した手法には入れ子となる並列構造の解析ができないという欠点があった．そこで，並列構造の範囲と階層構造を同時に決定する方法を提案し，その優位性を，ベンチマークデータを用いた実験で確認した．

日本語並列構造解析の精度向上に向けては，以前に英語に対して提案した手法を日本語に適用できるように改良した．英語と日本語の違いは，英語並列句の手がかり表現(″and″や″or″など)が常に並列句を導くのに対し，日本語並列句の手がかり表現(「と」や「,」など)は並列句を導くとは限らないことである(例:「二条城と清水寺に行った」に現れる「と」は並列助詞であるが，「友達と清水寺に行った」の「と」は並列関係を示さない格助詞である)．このことに対応するため，アラインメントグラフにバイパス経路を新たに追加し，手がかり表現が並列句を導くかどうかを判定できるようにした．この結果，日本語並列構造解析の精度向上に成功した．

並列構造解析と統語解析（構文解析，固有表現抽出等）の融合に向けては，ベンチマークデータではない一般のテキストデータを用いて，手法の開発および適用を行った．具体的な成果として，人手作業による専門用語抽出が困難となる膨大な数（1,000,000 件）の生命医学文書に対して，生物種名，病名，遺伝子・蛋白質名，化合物名を自動アノテーションするプロジェクト（欧州バイオインフォマティクス研究所(EMBL)が主催する CALBC プロジェクト，http://www.calbc.eu/）において，本研究課題により開発されたシステムは，参加 16 機関中で，病名，化合物名，生物種名について 1 位，遺伝子・蛋白質名について 3 位の精度を達成した．これは，本研究課題による開発手法の（ベンチマークデータではない）一般のテキストデータに対する有効性を示すものである．

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 4 件）

① Dietrich Rebholz-Schuhmann , Nigel Collier, Udo Hahn, Kazuo Hara ら 37 名（Kazuo Hara は 21 番目）, Assessment of NER solutions against the first and second CALBC Silver Standard Corpus, Journal of Biomedical Semantics, 2(Suppl 5), S11, 2011, doi:10.1186/2041-1480-2-S5-S11, 査読有

② 原 一夫, 新保 仁, 松本 裕治, 文法制約と系列アラインメントによる並列構造の解析, 人工知能学会論文誌, Vol. 25, No. 5, pp. 569-578, 2010 , doi:10.1527/tjsai.25.560, 査読有

③ 大熊 秀治, 原 一夫, 新保 仁, 松本 裕治, バイパス付きアラインメントグラフを用いた日本語並列句検出と範囲同定, 人工知能学会論文誌, Vol. 25, No. 1, pp. 206-214, 2010, dou:10.1527/tjsai.25.206, 査読有

④ Javier Tejada-Cárcamo, Hiram Calvo, Alexander Gelbukh, and Kazuo Hara, Unsupervised WSD by finding the predominant sense using context as a dynamic thesaurus, Journal of Computer Science and Technology, Vol. 25, No. 5, pp. 1030-1039, 2010, doi:10.1007/s11390-010-1081-8, 査読有

〔学会発表〕（計 7 件）

① 原 一夫, 新保 仁, 松本 裕治, バイオ医療テキストに対する並列構造解析と固有表現抽出の統合解析, 第 33 回日本分子生物学会年会・第 83 回日本生化学会大会合同大会（BMB2010）, 2010 年 12 月 10 日, 神戸ポートアイランド

② 鈴木 郁美, 原 一夫, 新保 仁, 松本 裕治, バイオ医療専門用語のシソーラス拡張のための分布類似度計算法の提案, 第 33 回日本分子生物学会年会・第 83 回日本生化学会大会合同大会（BMB2010）, 2010 年 12 月 10 日, 神戸ポートアイランド

③ 鈴木 郁美, 原 一夫, 新保 仁, 松本 裕治, 係り受け木を利用した単語類似度計算方法とそのシソーラス拡張への応用, 情報処理学会研究報告, 自然言語処理研究会, 2009-NL-199, No.1, 2010 年 11 月 18 日, 広島市立大学

④ Kazuo Hara , Towards automatic biomedical entity annotation by reducing error propagation, In Proc. of the 1st CALBC workshop, pp. 35-37, 2010 年 6 月 18 日, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK

⑤ Ikumi Suzuki, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto , A Graph-based Approach for Biomedical Thesaurus Expansion, In Proc. of the ACM 3rd International Workshop on Data and Text Mining in Bioinformatics (DTMBIO), pp. 79-82, 2009年11月6日, Hong Kong, China

⑥ Kazuo Hara, Masashi Shimbo, Hideharu Okuma and Yuji Matsumoto, Coordinate structure analysis with global structural constraints and alignment-based local features , In Proc. of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009), pp.967-975, 2009年8月5日, Singapore

⑦ Hideharu Okuma, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto, Bypassed alignment graph for learning coordination in Japanese sentences, In Proc. of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009): Short Papers, pp.5-8, 2009年8月4日, Singapore

〔その他〕
ホームページ：
http://cl.naist.jp/project/coordination/en/

６．研究組織
(1)研究代表者

研究者番号
原　一夫（HARA　KAZUO）
国立遺伝学研究所・生命情報・DDBJ研究センター・研究員
研究者番号：3046769