様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 6 月 9 日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18700148
研究課題名（和文）逆アラインメント問題の条件付確率場による解法と情報抽出への応用
研究課題名（英文）Approach to inverse sequence alignment with conditional random fields and its application to information extraction

研究代表者
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・助教
研究者番号：90311589

研究成果の概要：

条件付き確率場の近似法である平均化パーセプトロン学習を用いて逆系列アラインメント (inverse parametric sequence alignment) 問題を解くための手法を提案した．逆系列アラインメントとはすなわち，与えられた訓練データから編集コストモデルを学習する問題である．応用として生物学文書（英語）および日本語（百科辞典および新聞記事）からの並列句検出および範囲同定に適用し，既存法に比べて高い精度を得た．タグ付け（教師データ）が不十分な場合でも対処するために 2 種類のヒューリスティックを提案し，その有効性を実証した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,300,000	0	1,300,000
2007 年度	1,200,000	0	1,200,000
2008 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,500,000	300,000	3,800,000

研究分野：　総合情報
科研費の分科・細目：情報学・知能情報学
キーワード：テキストマイニング，情報抽出，並列句解析，系列アラインメント

１．研究開始当初の背景

系列アラインメントは，計算生物学において盛んに用いられ，その有用性は広く知られている．そこでは，もっぱら既定のパラメタ (編集操作コスト) 設定の下での相同性検出問題が扱われている．生物学分野では，PAM や BLOSSUM といった編集コストモデルが存在するため，この制約は大きな問題とはならない．一方，自然言語処理や情報抽出などの新しい分野にこの手法を応用する場合，妥当なコストモデルが存在せず，専門家によってモデルを作成する手間が必要となる．その費用は，問題に関する専門知識が必要とされることから高価であり，そもそもコストモデルが人手によって調整できるほど単純なものかすら，わからない場合もある．このため，専門家に頼らずとも，訓練データ (系列間の

相当部分の対応情報）から自動的に最適な編集コストモデルを学習する手法が必要とされていた.

近年，機械学習および自然言語分野においては，系列タギングの手法として条件付き確率場が提案され，品詞タグ付けや，固有表現抽出などにおいてきわめて高い精度が得られることが報告されていた．この手法は，多くの素性を導入でき，素性間の確率的な独立性を仮定する必要がない，といった，隠れマルコフモデルなどの既存法とは違った好ましい性質を持つ．我々はこの性質に着目し，逆系列アラインメント問題への応用を試みた．さらに，自然言語処理の困難な問題の一つである，並列句解析への応用を目指した．並列句は特に臨床試験文献に頻出することから，高精度の並列句解析技術は医学系文書からの情報抽出には欠かせない.

２．研究の目的

本研究では，専門家が経験的に決めてきた編集操作コストを，与えられたアラインメント(相同箇所が明示された訓練事例）から学習し自動設定する「逆」系列アラインメント問題に取り組み，現実の文書の解析・情報抽出技術に役立てることを目指した．具体的には次の 2 点の開発を行った.

(1) 逆系列アラインメント問題，特にローカルアラインメントの，条件付確率場による解法.
(2) 上記 (1) を応用した，文書からの情報抽出法.

３．研究の方法

文書からの情報抽出（特に並列句解析）に役立てることを最終的な目的とするため，医学生物学文献コーパスである GENIA（東京大学辻井研究室; MEDLINE アブストラクトの一部に対して構文情報などを付与したもの）や，日本語百科事典コーパスを対象から並列句範囲を抽出し，人手による修正などを行った．複数の言語コーパスを用いたのは，学習に基づく手法の柔軟性を確認するためである．この際，人手による修正のためのタグ付け支援ツールをあわせて作成した．その後，モデル化と実装とを並行して行い，相互に改良していく手法をとった．当初は既存の条件付き確率場の実装をそのまま利用することを考えたが，改良の容易さから系列学習パーセプトロンを用いて新たに実装を行った．系列学習パーセプトロンは，条件付き確率場の近似版実装とみなせ，近似なしの場合と同等の性能が得られることが知られている.

４．研究成果

(1) 系列アラインメントに基づく並列句同定の基礎技術を開発した.類似の方法で並列句同定を行う手法はすでに存在するが,既存の手法がパラメタを人手で調整する必要があったのに対し,われわれが開発した手法は,学習機能を有しており，パラメタ調整が自動で行える，という大きな違いがある.このため我々の手法では,既存法では不可能であった,多数のパラメタを導入することが可能になる.学習には系列学習パーセプトロンをアラインメント学習用に拡張し用いた.この手法は条件付確率場の近似を用いた実装法の一種とみなせ,実装が比較的容易でありながら,条件付確率場と同等の精度が得られると報告されている.近年はパーセプトロン同様のオンライン学習法に関する研究が盛んに行われており,それらのより新しい手法に切り替えることも容易である.

(2) 開発した手法を実装し,この実装を用いて実験を行った.既存の英語医学生物学コーパスを対象にその精度を測定した.結果として，必要最低限の素性のみを用いたにもかかわらず,句構造文法パーザおよび系列チャンキングによる手法を上回る精度を得た.成果は人工知能学会論文誌にて公表した.

(3) 学習用訓練データに対するタグ付け作業負担を軽減するため,並列句両端のみのタグ付けデータを用いても十分な解析精度が得られる手法を考案した.このような不確定なデータに対処する手法を新たに考案し,英語医学論文アブストラクト集に適用して新手法の有効性を確認した.

(4) 並列句コーパスを作成するための 2 種類のタグ付けシステムを作成した.これらを用いて訓練データを作成することで,学習に用いるデータ量を効率的に増やすことが可能となる.

(5) 開発した英語並列句解析技術を日本語に対して適用した．対象文書は，医療文献ではなく,一般の新聞記事や百科事典である．日本語については，英語の並列句と異なり,並列句が文中に含まれているかいないかの判別自体が問題となり，これが精度向上の妨げとなることがわかった．英語の場合には少数の接続詞

“and” “or” などの手がかり表現が文内に含まれていれば,ほぼ間違いなく並列句がその周辺にある．これに対し,日本語では,「と」「も」といった助詞が,下の例のように並列句を導くとは限らない.

高台寺と清水寺に行った
（「高台寺」と「清水寺」の並列）
友達と清水寺に行った.
（「友達」と「清水寺」は非並列）

このため,並列解析モデルを改良し,（並列句範囲の同定に加えて）並列句の存在判定も同時に行う手法を提案した．この改良はアラインメント計算に用いるグラフに一本の辺を追加するだけの簡単な変更である．しかしながら，EDR コーパスの平凡社百科事典セクションを用いて評価したところ，改良前と比べて大きな性能の向上が見られ，また，既存の規則ベース並列解析器を上回る性能が得られた.

(6) 並列句間の距離に応じて素性を分解する（ことなる素性として扱う）ことでさらに性能が向上することがわかった.

(7) 本研究を通じて，素性としては，単純に文中から抽出できるもの（単語そのもの,品詞，単語の前後数文字の一致，など）をもっぱら用いている．外部のリソース（シソーラスや大規模コーパス中の単語共起頻度など）は大規模には用いなかった．この方策は，手法そのものの有効性を実証するには有効であったが，実用上はこれら外部リソースに基づく情報を素性として導入することで，よりよい編集コストモデルが得られる可能性がある．この点で，外部リソースの有効利用は今後の課題として残っており，リンク解析技術などを応用した類似度測定法を応用し今後研究を継続する予定である．この方針にそって，機械学習分野で注目されているカーネル法をリンク解析に適用する際の問題点について調査を行い，複数のトピック（コミュニティ）が存在するグラフにおける，ある種のカーネルの問題点を指摘し，そのための解決法を提案した．成果は PKDD, KDD, IJCNLP といった国際学会にて公表した．この知見を生かして，医学生物学コーパス中の単語のネットワークを作成し，開発したリンク解析技術を適用して得られた情報を（医療文献の）並列句解析の際に素性として用いる計画である.

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 3 件）

(1) 原 一夫, 新保 仁, 松本 裕治. アラインメントと機械学習を応用した並列句解析: 医学生物学論文からの情報抽出に向けて. 人工知能学会論文誌 Vol. 22, No. 3, pp. 248–255, 2007. 査読あり.

(2) 伊藤 敬彦, 新保 仁, 持橋 大地, 松本 裕治. コミュニティを考慮したカーネル引用解析. 電気電子通信情報学会論文誌 D, Vol. J90-D, No. 2, pp. 233–244, 2007. 査読あり.

(3) 新保 仁, 伊藤 敬彦, 松本 裕治. カーネルリンク解析におけるパラメタ依存性と近似計算について. 日本データベース学会 Letters (DBSJ Letters), Vol. 5, No. 2, pp. 101–104, 2006. 査読あり.

〔学会発表〕（計 9 件）

(1) 大熊 秀治, 新保 仁, 原 一夫, 松本 裕治. バイパス付き編集グラフを用いた日本語並列構造解析. 情報処理学会研究報告 自然言語処理研究会 2009-NL-190, pp.111-118. 2009 年 3 月 東京. 査読なし.

(2) L. イェン, M. サレンス, A. マントラック, 新保 仁. A family of dissimilarity measures between nodes generalizing both the shortest path and the commute-time distances. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 785–793. August 2008, Las Vegas, USA. 査読あり.

(3) 大熊 秀治, 原 一夫, 新保 仁, 松本 裕治. 機械学習と系列アラインメントを応用した日本語並列句解析. 2008 年度人工知能学会全国大会（第 22 回）論文集, 1H1-03, 2008 年 6 月 旭川. 査読なし.

(4) 小町 守, 工藤 拓, 新保 仁, 松本 裕治. カーネル法を用いた意味的類似度の定義とブートストラップの一般化.言語処理学会第 14 回年次大会論文集, pp. 825–828, 2008 年 3 月 東京. 査読なし.

(5) H. バンダリ, 新保 仁, 伊藤 敬彦, 松本 裕治. Generic Text Summarization Using Probabilistic Latent Semantic Indexing. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pp. 133–140. Hyderabad, India, January 2008. 査読あり.

(6) 新保 仁. カーネル法によるリンク・引用解析. 人工知能学会基本問題研究会(第 66 回)予稿集, p. 65, 2007 年 7 月 湯布院. 査読なし.

(7) 原 一夫, 新保 仁, 松本 裕治. 部分アラインメント同定を応用した並列句解析. In *Proceedings of the First International Workshop on Data-Mining and Statistical Science (DMSS 2006)*, pp. 167–172. 2006 年 9 月 札幌. 査読なし.

(8) 伊藤 敬彦, 新保 仁, 持橋 大地, 松本 裕治. Exploring multiple communities with kernel-based citation analysis. In *Proceedings of the Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pp. 235–246. *Lecture Notes in Artificial Intelligence* 4213, Springer. September 2006, Berlin, Germany. 査読あり.

(9) 伊藤 敬彦, 新保 仁, 持橋 大地, 松本 裕治. Investigating the effect of multiple communities on kernel-based citation analysis. In *Proceedings of the Second International Special Workshop on Databases for Next-Generation Researchers (SWOD 2006).* Atlanta, USA, April 2006. 査読あり.

〔図書〕(計 1 件)

(1) 新保 仁, 伊藤 敬彦. Kernels for link analysis. In *Mining Graph Data*, Chapter 12, pp. 283–310. John Wiley & Sons, December 2006.

〔その他〕

成果ページ
http://cl.naist.jp/project/coordination

６．研究組織
(1)研究代表者
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・助教
研究者番号：90311589

(2)研究分担者
該当なし

(3)連携研究者
該当なし