様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　25　年　6　月　7　日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2011～2012
課題番号：２３７００１６２
研究課題名（和文）　解析アクションの先読みに基づく高速・高精度な自然言語文解析

研究課題名（英文）　Efficient and accurate natural language analysis with lookahead of analysis actions

研究代表者
鶴岡　慶雅（TSURUOKA　YOSHIMASA）
東京大学大学院工学系研究科・准教授
研究者番号：５０５６６３６２

研究成果の概要（和文）：

本研究では、品詞タグ付けや構文解析といった様々な自然言語処理タスクに適用可能な機械学習アルゴリズムの開発を行った。本アルゴリズムは、「履歴に基づくモデル」に先読み機構を導入することを可能にし、その解析精度を大幅に向上することが可能である。実験の結果、複数の自然言語処理において、本アルゴリズムによるアプローチは、自然言語処理分野で標準的に使われるモデルである「条件付き確率場」モデルよりも精度の点で優れていることが示された。

研究成果の概要（英文）：

We have developed a novel machine learning algorithm that can be used for various natural language processing tasks such as part-of-speech tagging and parsing. The algorithm enables us to incorporate a look-ahead mechanism into a history-based model and significantly improve its accuracy. Experimental results demonstrate that our approach outperforms conditional random field models, which are currently the standard approach in the field, in several natural language processing tasks.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	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、機械学習

１．研究開始当初の背景

テキスト情報を計算機で扱うための基盤技術である自然言語処理技術は、機械学習による手法をベースとして近年著しく発展し、品詞タグ付けや形態素解析などの比較的浅い構造の処理に関しては、その解析精度が人間と遜色のないレベルに達しつつある。しかし、構文解析をはじめとする複雑な構造を解析する処理に関しては、いまだ十分な精度に達していない。品詞タグ付けのような比較的シンプルな処理においては、高い精度（97 %以上）が達成されているが、これらの処理の出力が、固有表現認識や構文解析といった後段の処理での入力になることを考慮すると、現状の精度でも十分に高いとはいえない（図１）。

他方、情報抽出をはじめとする、自然言語処理の応用対象となるテキストの量は、飛躍期に増大し、テラバイトを超えるテキストの

処理が必要とされることは珍しくない。また、リアルタイムおよびインタラクティブな情報抽出を実現するためには、ユーザーから受け取ったテキストをその場で処理する必要があるため、言語処理アルゴリズムを開発する上では、処理速度も精度と同様に重要な要素となっている。

機械学習をベースとした、自然言語処理の場合、解析速度と解析精度は基本的にはトレードオフの関係にある。解析精度を高めるためには、一般に、非局所的な素性（予測の際に用いる特徴量）を導入する必要があるが、そのような素性を単純に導入すると、最適な構造を計算するプロセスの計算量が文長に対して指数オーダーになってしまうため、実用にならない。したがって、このようなタスクのための機械学習モデルを開発するにあたっては、計算量の指数爆発をさけつつ、有用な非局所素性を導入できる仕組みが重要となる。そのような仕組みとしては、リランキングに基づく手法やサンプリングに基づく手法がよく知られている。

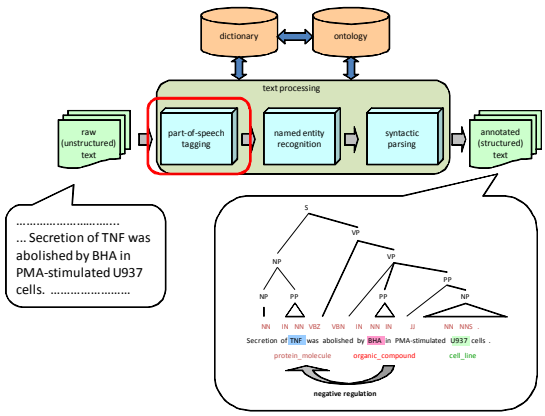

図１　自然言語処理パイプライン

２．研究の目的

本研究では、Maximum Entropy Markov Models (MEMMs) や Shift-Reduce Parsers に代表される、履歴にもどづくアプローチ（history-based models）を拡張し、解析アクションの先読み機構を統合した機械学習モデルを提案する。一般に、履歴にもどづくアプローチでは、解析アクションを決定する問題は、単純な多クラス分類の問題として定式化される。すなわち、文の表層的な情報とアクションの履歴とを素性として利用したシンプルな機械学習の問題となる。それに対して提案手法では、現時点で可能なそれぞれの解析アクションに対して先読みを行い、その後の解析アクションの系列によって達成される解析結果をもとに現時点での最適なアクションを選択する。このことにより、非局所的に整合的でない部分解析結果を検出することが可能になり、解析精度を向上させることができると考えられる。また、先読みの到達範囲まで自由に非局所素性を導入することが可能になるというメリットがある。

本研究プロジェクトでは、提案手法を自然言語処理における様々な構造予測の問題に適用し、その有効性を検証する。具体的には、自然言語処理における構造予測の代表的なタスクである、英語の品詞タグ付け、固有表現認識、構文解析において、デファクトスタンダードの機械学習モデルとして幅広く使われている条件付確率場（Conditional Random Fields）による手法などよりも精度、速度の面で優れていることを示す。

３．研究の方法

英語における品詞タグ付け、浅い構文解析、固有表現認識などの問題は、系列予測問題として定式化することが可能であり、自然言語処理における構造予測の問題としてはもっともシンプルな部類に属する。

品詞タグ付の問題を例にとると、Maximum Entropy Markov Models (MEMMs) のような履歴に基づく手法を用いた場合、予測器の役割は、文全体の単語情報と先行する単語へ付与した品詞の情報を入力として、現在注目している単語の品詞を正しく予測することである。このとき、予測器は単純な多クラス分類器として実装されるため、後続する単語にどのような品詞が付与されるかということは予測の時点では全く考慮されない。もちろん、MEMMs のように分類器として確率値を出力するモデルを利用する場合、Viterbi アルゴリズムなどの動的計画法を用いて確率最大のパスをにより見つけることで、個々の予測の誤りを事後的に修正することが可能ではあるが、後続する単語に付与される品詞の情報は、あくまでも間接的に利用されるにすぎない。

それに対して、本提案手法では、解析アクション（品詞タグ付の場合であれば付与する品詞そのもの）の先読みを行うことで、直接的に後続する品詞タグ列との整合性を考慮した予測を行う。図２に手法の概念図を示す。$m_1, \ldots, m_k$ が現時点で可能な解析アクション、$n_1, \ldots, n_k$ が、それらの解析アクションによって実現される状態、$t_1, \ldots t_k$ が、それぞれの状態から、先読みによって求まる最適パスの末端に対応する状態である。従来手法では、現時点での解析アクションによって直接実現される状態 $n_1, \ldots n_k$ を評価して最適なアクションを決定するのに対して、提案手法では、先読みの末端状態である $t_1, \ldots t_k$ を評価することで最適なアクションを決定

する。従来手法の場合、先読みのプロセスが入っていないため、将来的には行き詰る解析アクションであっても、現時点の見た目の評価が高ければそれが選択されてしまうのに対し、提案手法では、将来的に最もよい解析結果が得らるアクションを現時点で選択することが可能となっている。

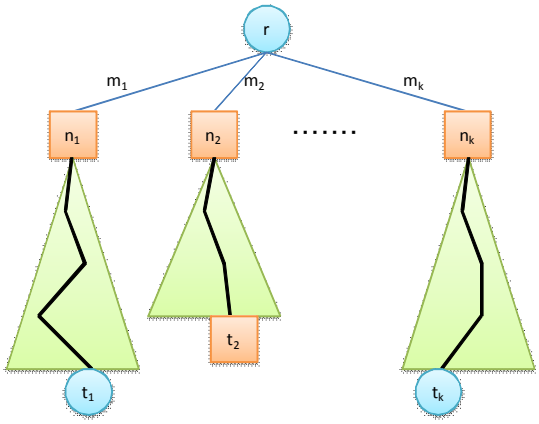

図 2: 先読みによる最適なアクションの決定

構文解析に関しても品詞タグ付けなどの系列予測の問題と同様に本アルゴリズムを適用することが可能である。たとえば、機械学習に基づく Shift-Reduce パージングにおいては、通常、パージングの各状態において、Shift するか Reduce するかを決定する分類問題として定式化されるが、提案手法では、これに先読み機構を導入し、Shift/Reduce アクションが n 回行われた未来の状態をもとに現時点での最適なアクションを決定する（n は先読みの深さ）。これによって、将来的に構築される部分構造の良さを考慮した予測が可能になり、係り受け解析の精度が向上することが期待される。

係り受け解析は、完全な句構造を解析する処理よりも計算量が小さく、構文解析への適用の最初のタスクとして適切であると考えられる。係り受け解析のアルゴリズムとしては、現在さまざまな手法が提案されているが、文全体を一度にモデル化する手法としては、Maximum Spanning Tree アルゴリズムを利用して、局所素性を用いたモデルを最適化するアルゴリズムがよく知られている。

提案手法を自然言語処理において実現するにあたって難しい問題は、状態を評価する関数のパラメータをどのようにして学習するかという問題である。自然言語処理用に学習コーパスには、当然のことながら、先読みにおける最適パスの情報は含まれていないため、直接的に最尤推定のような基準でパラメータの最適化を行うことはできない。そこで本研究では、学習時においては、現時点でのモデルパラメータを用いて計算された最適パスを用いて評価関数の最適化を行うというアプローチをとる。

４．研究成果

提案手法の性能評価は、英語の品詞タグ付け、固有表現認識、浅い構文解析、係り受け解析に関して、多くの研究で標準的に用いられているコーパスを利用して行った。

学習アルゴリズムとしては、平均化パーセプトロンをベースとした学習アルゴリズムを提案し、その収束性を理論的に示した。また、実験によって、実用上の収束性、予測性能の両面で優れた結果が得られた。

図３～図６に、それぞれの自然言語処理タスクにおける精度を示す。各グラフの最下段には、標準的な従来手法（条件付き確率場および構造化パーセプトロン）によって得られた精度を示す。

図から、先読みの深さを深くすることによって我々の提案手法の精度が向上していくことがわかる。我々の提案する先読みに基づくアプローチでは、先読みの深さと解析精度が深く関係する。もし、先読みの深さが０であれば、提案手法は、従来の履歴にもどづくアプローチと等価である。基本的には、先読みの深さを深くすればするほど、導入可能な素性の非局所性が高まり、さらに、解析の不整合を早期に発見できるようになるため、解析の精度は向上すると考えられる。しかし、先読みに必要な計算コストは、基本的に先読みの深さに対して指数的に増えていくため、現実的に可能な先読みの深さは限られていることに注意する必要がある。

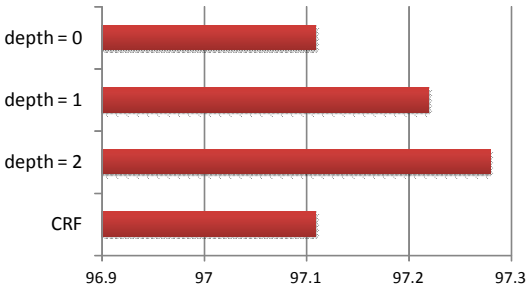

図３　品詞タグ付けの性能

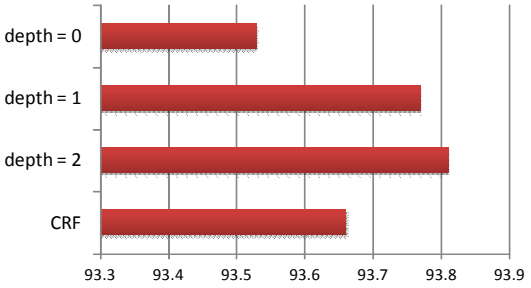

図４　チャンキングの性能

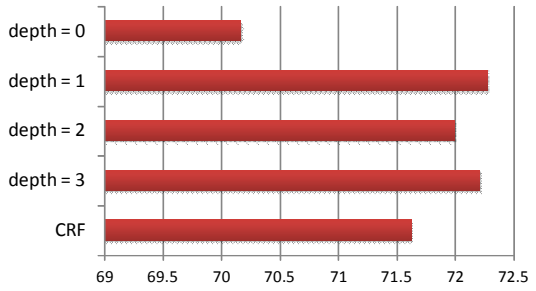

図５　固有表現認識の性能

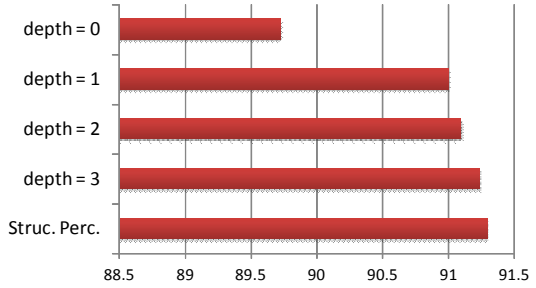

図６　係り受け解析の性能

本研究課題では、基盤的な自然言語処理技術である、品詞タグ付け、固有表現認識、構文解析などの、構造予測問題と呼ばれるタスクに対する新しい機械学習アプローチの提案を行った。提案手法は、解析アクションの履歴に基づくアプローチをベースとし、それに先読み機構を統合することで、解析精度の大幅な向上が達成された。具体的には、品詞タグ付けタスクであれば個々の単語に対するタグ付けのアクションを、構文解析であれば、Shift や Reduce といった解析アクションを先読みすることで、現時点でのスコアではなく、将来的な有望さに基づいてアクションを選択することを可能にしている。

提案手法を実現するにあたって最も難しい問題は、状態を評価する関数のパラメータをどのようにして学習するかという問題であるが、本研究では、学習時においても、先読みによって計算されたアクションの最適パスを用いて評価関数の最適化を行うというアプローチをとることでこの問題を解決した。具体的な最適化のアルゴリズムとしては、平均化パーセプトロンに基づく手法を提案し、その収束性を理論的に明らかにした。

提案手法を、英語の品詞タグ付け、固有表現認識、係り受け解析タスクに適用した結果、このようなタスクに対して幅広く使われている「条件付確率場」を用いた手法よりも、解析精度、計算量の点で優れた性能を達成できることを示した。特に、品詞タグ付けと固有表現認識については、計算コスト同等の条件で比較した場合、世界最高レベルの精度を実現している。また、提案手法を、近年注目を集めている「最易優先方策」と組み合わせることで、さらなる精度向上が可能であることを示した。

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０件）

〔学会発表〕（計２件）

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Kazama. 2011. Learning with Lookahead: Can History-Based Models Rival Globally Optimized Models? In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL), pp. 238-246

佐野峻平，三輪 誠，鶴岡慶雅，近山 隆，先読みを用いた単語系列ラベリングへの最易優先方策の適用，言語処理学会第 19 回年次大会，2013 年 3 月.

〔その他〕
ホームページ等
http://www.logos.ic.i.u-tokyo.ac.jp/~tsuruoka/conll111a/

６．研究組織
(1)研究代表者
鶴岡 慶雅（TSURUOKA YOSHIMASA）
東京大学・大学院工学系研究科・准教授
研究者番号：５０５６６３６２

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：