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　３０　年　　６　月　１１　日現在

機関番号：　１４６０３
研究種目：　研究活動スタート支援
研究期間：　2016～2017
課題番号：　１６Ｈ０６９８１

研究課題名（和文）単語のベクトル表現に基づく分野の変化に頑健な構文解析器に関する研究

研究課題名（英文）Studies on robust statistical parsing across different domains using word embeddings

研究代表者

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

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：現在の機械学習に基づく自然言語処理の一つの問題として、学習に用いた分野（ドメイン）と大きくかけ離れた文書の解析がうまく行えないという点が挙げられる。現在、構文解析を始めとした多くの解析手法は新聞記事をもとに学習を行っているため、他の種類の文書、例えば論文やWeb上に存在する文をうまく解析することができない。この点を解決するための新しい構文解析技術を目指し、まず、単純なモデルで、品詞などの前処理に依存しないながら、高精度を達成することのできる新しい組み合わせ範疇文法の解析法を提案した。また、構文解析のドメイン適応に適した新しいニューラルネットワークを提案し、有効性を示した。

研究成果の概要（英文）：A problem in statistical natural language processing based on machine learning is that a system performs poorly on texts, which come from a different domain than the one of the training data. Since most systems, such as parsers, are trained with annotated data in the newspaper domain, their performance significantly drops on other kinds of texts, e.g., web and scientific papers. Toward more robust parsing method across different domains, we first developed a new simple parser based on Combinatory Categorical Grammar (CCG), which has an advantage that it does not require preprocessing including POS tagging. We also designed a new neural network architecture for parser domain adaptation, and verified the effectiveness of the approach.

研究分野：　計算言語学

キーワード：　構文解析　組み合わせ範疇文法　ドメイン適応

１．研究開始当初の背景

現在の機械学習に基づく自然言語処理の一つの問題として、学習に用いた分野（ドメイン）と大きくかけ離れた文書の解析がうまく行えないという点が挙げられる。現在、構文解析を始めとした多くの解析手法は新聞記事をもとに学習を行っているため、他の種類の文書、例えば論文や Web 上に存在する文をうまく解析することができない。

機械学習分野の成功を受け、2013 年頃から、自然言語処理の分野でもニューラルネットワークに基づく手法が注目を集めるようになってきた。自然言語処理の基礎解析の多くは、一つの文を入力とし、各単語に対しラベルを付与することで行われる。この問題に対し、ニューラルネットワークに基づく手法では、各単語を数十〜数百次元の連続値ベクトルで表現し、入力となるこれらベクトル列の間の複雑な非線型変換により、各単語に相当するラベルを予測する。

構文解析器などのドメイン適応を目指す際も、ニューラルネットワークは強力な枠組みとなる可能性が高い。ドメイン適応の際に問題となるのは、一つは使われる単語が大きく異なり、元のドメインで出現しない単語の使われ方が予測できないこと、もう一つは、同じ単語であっても、ドメインが異なると単語の意味や用法が異なる場合があることである。これらの問題に対し、ニューラルネットワークでは、単語を離散的な一つのシンボルではなく、連続値ベクトルとして表現するため、単語の意味の変化なども、ベクトル演算などを通じ効率的に学習できる可能性がある。

一方、構文解析の研究自体は、ニューラルネットワーク登場後も様々に盛んであるものの、そのドメイン適応での可能性を追求したものはあまり見たらない。そこで本研究では、ニューラルネットワークをうまく活用した構文解析のドメイン適応手法や、関連する構文解析の問題について探求する。

２．研究の目的

ニューラルネットワークは自然言語処理にとって新しい道具であり、また単語を離散的なシンボルでなく連続値で扱うことから、単語の意味変化などが生じるドメイン適応についても、従来の方法よりも有効に対処できる可能性が高い。そのようなニューラルネットワークの可能性を探求するため、現代の時代に即した、新しい構文解析の手法を構築し、その有効性を検証する。

３．研究の方法

特に以下の 2 点に着目し、研究を進める。

(1) 構文解析器のモデルとしては様々な枠組みが考えられるが、出力となる文法も、選択の一つに含まれる。モデルの複雑さや文法の複雑さの間には密接な関係があるが、ドメイン適応を考えた場合、モデルとしてはなるべく単純なものの方が適していると考えられる。本研究では特に、対象とする文法に若干の複雑さを加えることで、単純なモデルにより高精度を達成できることの可能性を探求する。特に、近年自然言語処理で利用が広まっている、組み合わせ範疇文法 (CCG) の解析をより単純なモデルで解析することを目指す。

(2) 構文解析のための新しい一般的なドメイン適応の方法として、敵対的学習 (adversarial training) に基づく方法を試し、有効性の検証と、有効なニューラルネットワークの構成法についての知見を得る。敵対的学習は、機械学習や画像処理の分野で有効性が示されつつある方法であり、ドメイン適応においては、適応元を表すドメインのモデルでの、入力文のベクトルと、適応先のドメインのモデルでの入力文のベクトルを、なるべく区別できなくするように学習を進めることで、少量のデータしかない適応先のドメインにおいても、適応元のデータを最大限に利用できるようにするもの、と理解される。これは、ニューラルネットワークに基づく表現学習によって初めて可能になった新しい方法であり、構文解析に対して有効か、どのような時に有効かを検証することは重要である。

４．研究成果

まず CCG 構文解析に対して、CCG の文法の特性をうまく活用することにより、シンプルなモデルながら高精度を達成できることを示した。これは本質的には、CCG 構文解析をスーパータギングと呼ばれる、単語に対する統語範疇の割り当てに帰着させるものである。従来手法では、スーパータギングだけでは適切な曖昧性解消が行われないという問題が知られていたが、これに対し我々は、同じく単語レベルの単純な依存構造のモデル化を同時に行うことで、スーパータギングだけでは解消できない曖昧性を有効に解消できることを示した。この際に興味深い観察として、同時に学習する依存構造は、言語学的に望ましいと考えられるものでなくてもよく、むしろ CCG の構文木から単純なルールで抽出したものを同時に予測することで、精度の向上につながることを発見した。提案法は、日本語および英語の文に対し高速に動作し、またどちらも世界最高精度を達成した。本システムは一般に公開している。

敵対的学習に関する研究では、依存構造を対象に研究を行った。ここで用いた基本となるモデルは、上記 CCG 構文解析で利用したものと非常によく似た、両方向 LSTM からの特徴抽出結果を利用するものである。構文解析での依存構造に適したネットワーク構造として、ドメイン間での共通の表現に加え、

ドメイン固有の表現も抽出する新しいモデルを提案し、これが有効に働くことを検証した。また構築したシステムを利用し、国際会議 CoNLL での shared task に参加し、全 33 チーム中 6 位の成績を収めることができた。

提案した CCG 構文解析の重要な特性として、単語から直接、統語範疇（スーパータグ）を予測するため、品詞解析などの前処理を必要としない、という点が挙げられる。これは実用的には大きな強みであると考えられる。多くの構文解析器は、正しい品詞を想定しその上で解析を行うものが多く、品詞が間違ってしまうと、そのエラーが伝播し、構文解析も間違ってしまう可能性が高いからである。特にドメイン適応の際は、この問題が顕著となる。この点から、提案した手法は、ドメイン適応にも特に有効なモデルとなっている可能性が高く、上記敵対的学習などと組み合わせることで、より効率的なドメイン適応が行える可能性が高い。本研究の期間内にこの点を検証することはできなかったが、CCG 構文解析を様々なドメインで活用としたいという要望は耳にしており、現在、その点について研究を進めている。

別の関連する問題として、依存構造解析を行う際の適切な文法表現、及び、所定の表現を得たい場合であっても、解析しやすい表現を経由することで精度を向上させる方法について研究を行った。この研究を行った背景として、近年依存構造解析では Universal Dependencies と呼ばれる、多数の言語に渡って統一的な文法表現によるアノテーションの研究が進められていること、しかし構文解析器にとってみれば、Universal Dependencies が解析しやすい表現となっているかについては疑問が持たれていること、が挙げられる。つまり、最終的には Universal Dependencies を得たいというニーズがあるが、解析を全てその表現で行うべきか、または解析の際は別の表現で行い、その後後処理として、Universal Dependencies の表現に直すべきか、という点について、検証を行った。様々な実験を通し、 Universal Dependencies に、複雑すぎない適切な量の変換を施し、その上でモデルの学習、及び解析を行うという流れによって、一貫して、最終的に出力を精度の向上が可能となることを示した。これは特定の構文解析器の仕組みとは全く独立の外部から働く仕組みであるため、あらゆる構文解析器と組み合わせて使うことが可能である。

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　　5 件）

① Masashi Yoshikawa, Koji Mineshima, Hiroshi Noji, and Daisuke Bekki. Consistent CCG Parsing over Multiple Sentences for Improved Logical Reasoning. Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2018). Vol. 2. pp. 407-412. (査読あり)

② Ryosuke Kohita, Hiroshi Noji, and Yuji Matsumoto. Effective Online Reordering with Arc-Eager Transitions. Proceedings of the 15th International Conference on Parsing Technologies. 2017. Vol. 2. pp. 88-98. (査読あり)
https://aclweb.org/anthology/W/W17/W17-6313.pdf

③ Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. Adversarial Training for Cross-Domain Universal Dependency Parsing. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2017. Vol. 1. pp. 71-79. (査読あり)
http://www.aclweb.org/anthology/K17-3007

④ Masashi Yoshikawa, Hiroshi Noji, Yuji Matsumoto. A* CCG Parsing with a Supertag and Dependency Factored Model. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Vo. 1. pp. 277-287. (査読あり)
DOI: 10.18653/v1/P17-1026

⑤ Ryosuke Konita, Hiroshi Noji, and Yuji Matsumoto. Multilingual Back-and-Forth Conversion between Content and Function Head for Easy Dependency Parsing. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Vol. 2. pp. 1-7. (査読あり)
http://aclweb.org/anthology/E17-2001

〔学会発表〕（計　　4 件）

① 小比田涼介, 能地宏, 松本裕治. 逐次型依存構造解析における動的特徴量選択. 言語処理学会第 24 回年次大会. 2018 年 3 月 14 日. 岡山コンベンションセンター.

② 吉川将司, 能地宏, 松本裕治. 係り受け構造とスーパータグの同時予測による A* CCG 解析. 情報処理学会 第 231 回自然言語処理研究会 第 116 回音声言語情報処理研究会 合同研究発表会. 2017 年 5 月 15 日. 大阪大学中之島センター.

③ 吉川将司, 能地宏, 松本裕治. 係り受け構造との同時予測による A* CCG 解析. 言語処理学会第 23 回年次大会. 2017 年 3 月 14 日. 筑波大学 (茨城県つくば市).

④ 小比田涼介, 能地宏, 松本裕治. 依存構造解析のための内容語と機能語の多言語可逆変. 言語処理学会第 23 回年次大会. 2017 年 3 月 14 日. 筑波大学 (茨城県つくば市).

〔その他〕
ホームページ等
① 依存構造変換ソフトウェア
https://github.com/kohilin/MultiBFConv
② CCG 構文解析器 (depccg)
https://github.com/masashi-y/depccg

６．研究組織

(1)研究代表者
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