

深層学習による CCG 構文解析と文生成の同時学習

理学専攻・情報科学コース 1840670 馬目華奈

1 はじめに

近年、LSTM 等のニューラルネットが自然言語処理の様々なタスクに用いられている。しかし、LSTM が文法を学習しているのかは明らかでない。一方、文法を学習するニューラルネットのアーキテクチャである Recurrent Neural Network Grammars (RNNG) [1] が研究されている。本研究では、構文解析と文生成の同時学習を Combinatory Categorical Grammar (CCG) [2] に基づいて行うことで高精度な解析と文生成を目指す。

2 構文解析と文生成の同時学習

RNNG は Shift-Reduce 法を用いて、句構造文法に基づいたアクション系列の予測することで解析・生成を行う。アクションの種類は、非終端記号を導入する NT(<category>)、スタックに記号を移動させる Shift、括弧を閉じる（句の成立を認める）Reduce である。句構造文法のラベルをスタックへと移動し、入力文の単語が入っているバッファを考慮しながら、アクションを適用させていくことで、スタック上に文全体の句構造が構築されていく。Shift 操作の際に単語予測を行うことで、文生成のモデルとして用いることができる。

入力文に対する正解の木を与え、木の予測を行う Discriminative モデルと、文と木の同時確率を求める Generative モデルがある。

3 提案手法

本研究では CCG に基づく RNNG アクション予測のモデルを提案する。

3.1 CCG

CCG は語彙化文法の一つで、近年はニューラルネットを用いた CCG 構文解析の技術が発展し、高精度な解析を行うことができる [3]。各語には統語範疇が割り当てられ、語と語の統語的・意味的な関係を関数適用や関数合成などの組合せ規則により計算する。統語範疇は句構造文法のラベルに比べ、豊富な情報を持ち、その種類も多い。CCG は、並列句などのより複雑な文の統語解析することができる。

3.2 CCG 構文解析と文生成の同時学習モデル

CCG 木の導出を手手でアノテーションしたデータセットである CCGBank[4] の正解の木から、組み合わせ規則を適用できるカテゴリのルールテーブルを作成する。ルールテーブルを参照し、Reduce 操作と NT(<category>) 操作を許可するか否かの制限を加える。しかし、CCGBank のカテゴリ数は 1639 個と膨大なため、NT(<category>) のカテゴリを当てるのが困難である。そこで、高精度な CCG 構文解析器である depccg[3] が用いるカテゴリに限定する。これにより、depccg で使用しているカテゴリ 526 個で NT(<category>) のアクション予測を行う。

4 実験

4.1 実験設定

CCGBank[4] の WSJ §2-21・§24・§23 をそれぞれ教師 (39604 文)・開発 (1338 文)・テスト (2407 文) データとする。教師データにより与えられた、アクション数は 528 個、終端記号数（語彙数）は 54012 個、非終端記号数は、Shift・Reduce アクションを除いた 526 個である。

4.2 評価方法

評価には構文解析の結果で用いられる、ラベルなし F 値を用いる。句構造文法では、ラベルを当てるタスク (POS tagging) は別のタスクとして扱うため、CCG 構文解析の評価もラベルなしで行う。

4.3 実験結果・考察

解析の結果を表 1 に示す。(D) は Discriminative モデル、(G) は Generative モデルを用いた結果である。上界は depccg によって作成したアクション系列である。深い構文解析が行える CCG に基づいたアクション予測モデルに加え、カテゴリを高頻度に出現するものに限定することで、Discriminative モデル、Generative モデルともに RNNG モデルを上回る結果となった。

表 1: WSJ §23 の評価

手法	F 値
RNNG(D)	80.7
RNNG(G)	82.7
depccg	94.0
提案手法 (D)	87.4
提案手法 (G)	89.7

5 おわりに

本研究では、CCG に基づく構文解析と文生成の同時学習モデルを提案した。CCG 構文解析器である depccg によって作成したアクション系列を用いて学習を行うことで、句構造文法に基づく RNNG の精度を上回った。今後は生成を行う言語モデルについても研究を進めたい。

参考文献

- [1] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proc. of ACL*, pp. 199–209, 2016.
- [2] Mark Steedman. *Surface Structure and Interpretation*. In *The MIT Press*, 1996.
- [3] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG Parsing with a Supertag and Dependency Factored Model. In *Proc. of ACL*, pp. 277–287, 2017.
- [4] Julia Hockenmaier and Mark Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics*, Vol. 33, No. 3, pp. 355–396, 2007.