DEEP LEARNING JP
[DL Papers]

Tree-Averaging Algorithms for Ensemble-Based Unsupervised Discontinuous Constituency Parsing

Presenter: Masaki Sashida, Matsuo-Iwasawa lab, M1

書誌情報

紹介論文

タイトル:"Tree-Averaging Algorithms for Ensemble-Based Unsupervised Discontinuous Constituency Parsing"

出典: ACL 2024, SAC Awards

著者:Behzad Shayegh, Yuqiao Wen, Lili Mou

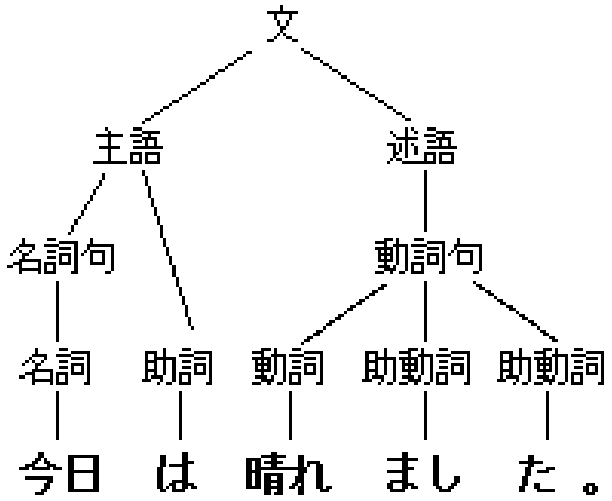
概要

教師なしの離散的構造句解析を行う、アンサンブル手法を用いたモデルを開発

※画像は出典記載のないものは、本論文から引用

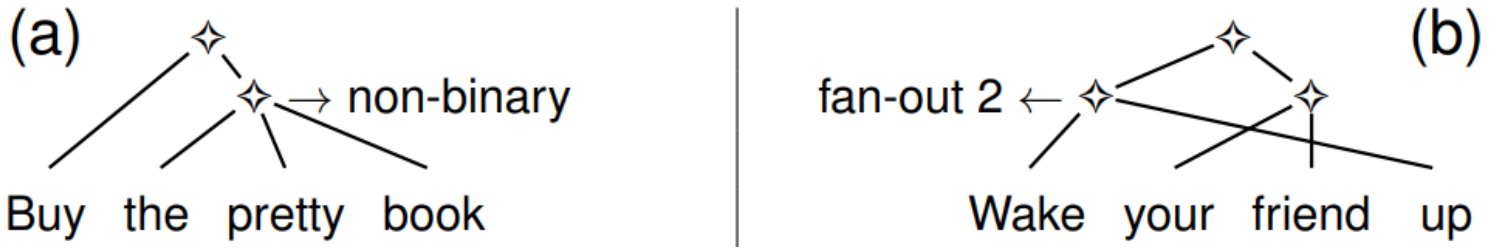
句構造解析とは

句構造解析とは・・・文の構造を解析し、単語間の文法的関係を明示すること。



教師あり　vs 教師なし

連続句構造解析　vs　非連続句構造解析



研究背景(1/2)

1. 教師なし構文解析の意義
   - 低リソースの言語やドメインに有益
   - 言語学や認知科学に有用
   - モーションセンサー信号などの他のタスクにも適用可能

2. 先行研究
   - 潜在変数手法（Clark, 2001; Petrov and Klein, 2007; Kim et al., 2019a）
   - ルールベースのシステム（Cao et al., 2020; Li and Lu, 2023）

研究背景(2/2)

3. 先行研究のリミテーション
   - ほとんどの手法では、連続構成句解析を行っている（実際はそうでないケースがある）
   - 非連続の構成句を扱った、Yang et al. (2023) の手法は、下記の課題があった
     - 性能が低い
     - ノイズが多く、シード別の実行間での相関が低い
   - CYKに類似した動的計画法を用いた、連続構成句解析のアンサンブリング（Shayegh et al., 2024）では、非連続構成句解析に適応できない

4. 目的
   - 前研究（Shayegh et al., 2024）の非連続構成句解析への応用

提案手法(1/4) ベースとなる文法

- LCFRS-2文法を採用
- LCFRS-2 は、 G = (S, N_1, N_2, P, Σ, R) で表される。
- Sは開始記号、N_1はファンアウト1の非終端記号の有限集合、N_2はファンアウト2の非終端記号の有限集合、Pは前終端記号、Σは終端記号
- Rは以下の形式のいずれかに従う有限個の規則からなる
- $B \in N_2$、$U, U' \in N_1 \cup P$

$$
\begin{aligned}
&R_1 : S(x) \rightarrow A(x) && A \in \mathcal{N}_1 \\
&R_2 : A(xy) \rightarrow U(x)U'(y) && A \in \mathcal{N}_1 \\
&R_3 : A(xyz) \rightarrow U(y)B(x,z) && A \in \mathcal{N}_1 \\
&R_4 : A(x,y) \rightarrow U(x)U'(y) && A \in \mathcal{N}_2 \\
&R_5 : A(xy,z) \rightarrow U(x)B(y,z) && A \in \mathcal{N}_2 \\
&R_6 : A(xy,z) \rightarrow U(y)B(x,z) && A \in \mathcal{N}_2 \\
&R_7 : A(x,yz) \rightarrow U(y)B(x,z) && A \in \mathcal{N}_2 \\
&R_8 : A(x,yz) \rightarrow U(z)B(x,y) && A \in \mathcal{N}_2 \\
&R_9 : T(w) \rightarrow w && T \in \mathcal{P}, w \in \Sigma
\end{aligned}
$$

非隣接構成句を扱うことができる！

提案手法(2/4) 精度向上に向けた工夫

- Yang et al. (2023) に倣い、テンソル分解に基づくニューラルネットワーク（TN-LCFRS）でパラメータ化された確率的LCFRS-2を学習
- ただ先行研究には下記の弱点
  - 性能が低い
  - ノイズが多く、シード別の実行間での相関が低い

- アンサンブル学習による改善を行った
- アンサンブルの出力は、単純に平均F1スコアが高い木を出力した

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \sum_{k=1}^{K} F_1(T, T_k)$$

提案手法(3/4) 問題の理論的な属性

- 「数 z と、同一の葉ノードを持つ構成句木 $T_1 \ldots T_k$ において、下記それぞれの条件で、$\sum_{k=1}^{K} F_1(T, T_k) \geq z$ となる構成句木 T が存在するか？」という問題を考える

1. 木はバイナリであり、ファンアウトが高々 F である。
2. ファンアウトは高々Fであるが、木は非バイナリであってもよい。
3. ファンアウトは無界であり、木は非バイナリであってもよい。

- １、２の条件が付いているときはPに属する
- ３の条件が付いているときはNP完全に属する

Binary individuals? \ Bounded fan-out?	Bounded	Unbounded
Binary	**P**	Unknown
Non-binary	**P**	**NP**-complete

提案手法(4/4) 計算量を落とす工夫

- バイナリかつファンアウトが有界な非連続ツリーのアンサンブル構築したいが、計算量を落とす必要
- 候補となる構成句をごく一部のみ考慮するような強力な枝刈りを行う
- これにより、$O(2^{2^n})$ の計算量を$O(2^n)$ に低減
- さらに、meet in the middle techniqueを用いることで指数を半分に削減し、最終的な計算量は$O(2^{\frac{n}{2}}n^2)$となる

実験設定

データセット

- 非連続構成句が比較的多いオランダ語およびドイツ語のデータセットで評価
- オランダ語についてはLASSY treebank (Van Noord et al., 2013)
- ドイツ語についてはNEGRA (Skut et al., 1997) と TIGER (Brants et al., 2002) のツリーバンクの合成データ

評価指標

- コーパス全体における全構成句、連続構成句、非連続構成句それぞれのF1 スコア

結果

	Method (# preterminal symbols)	NEGRA F_1^{overall}	NEGRA F_1^{cont}	NEGRA F_1^{disco}	TIGER F_1^{overall}	TIGER F_1^{cont}	TIGER F_1^{disco}	LASSY F_1^{overall}	LASSY F_1^{cont}	LASSY F_1^{disco}
Baselines (four runs each)†										
1	Left branching	7.8	–	0.0	7.9	–	0.0	7.2	–	0.0
2	Right branching	12.9	–	0.0	14.5	–	0.0	24.1	–	0.0
3	N-PCFG$^{(45)}$	$40.8_{\pm 0.5}$	–	0.0	$39.5_{\pm 0.4}$	–	0.0	$40.1_{\pm 3.9}$	–	0.0
4	C-PCFG$^{(45)}$	$39.1_{\pm 1.9}$	–	0.0	$38.8_{\pm 1.3}$	–	0.0	$37.9_{\pm 3.4}$	–	0.0
5	TN-PCFG$^{(4500)}$	$45.4_{\pm 0.5}$	–	0.0	$44.7_{\pm 0.6}$	–	0.0	$44.3_{\pm 6.4}$	–	0.0
6	N-LCFRS$^{(45)}$	$33.7_{\pm 2.8}$	–	$2.0_{\pm 0.8}$	$32.7_{\pm 1.8}$	–	$1.2_{\pm 0.8}$	$36.9_{\pm 1.5}$	–	$0.9_{\pm 0.8}$
7	C-LCFRS$^{(45)}$	$36.7_{\pm 1.5}$	–	$2.7_{\pm 1.4}$	$35.2_{\pm 1.2}$	–	$1.7_{\pm 1.1}$	$36.9_{\pm 3.7}$	–	$2.2_{\pm 1.0}$
8	TN-LCFRS$^{(4500)}$	$46.1_{\pm 1.1}$	–	$8.0_{\pm 1.1}$	$45.4_{\pm 0.9}$	–	$6.1_{\pm 0.8}$	$45.6_{\pm 2.3}$	–	$8.9_{\pm 1.5}$
Individuals: TN-LCFRS$^{(4500)}$										
9	Five runs	$46.4_{\pm 0.5}$	$49.8_{\pm 1.3}$	$6.0_{\pm 4.0}$	$45.8_{\pm 1.3}$	$49.9_{\pm 1.1}$	$4.0_{\pm 3.2}$	$46.7_{\pm 2.0}$	$50.9_{\pm 1.7}$	$6.2_{\pm 1.9}$
10	F_1^{overall}-best run	46.9	50.2	1.3	47.2	51.1	5.9	48.2	52.4	5.8
11	F_1^{cont}-best run	46.7	51.3	7.3	47.2	51.1	5.9	48.2	52.4	5.8
12	F_1^{disco}-best run	46.0	48.3	10.4	45.4	48.8	6.6	48.0	52.1	8.6
13	Binary ensemble	47.6*	50.1	9.9	47.8**	51.5	6.5	50.9**	54.6**	9.7**
14	Non-binary ensemble	**49.1****	**51.5**	**10.6**	**48.7****	**52.4****	**6.6**	**51.4****	**55.0****	**10.2***

Table 2: Main results. †Quoted from Yang et al. (2023). *p-value < 0.05 in an Improved Nonrandomized Sign test (Starks, 1979) against the best ensemble individual in each metric, indicated by underline. **p-value < 0.01.

提案手法がどの指標でも優れていることが示された

感想

強い点

- ニューラルネットワークだけではなく、LCFRS-2文法に着目し形式文法の観点からもアルゴリズムを構築している
- 計算時間を削減するとともにアンサンブルを行うことで、精度を保ちつつ総計算時間も実用的なものにしている
- アルゴリズムに対する理論的な証明も行っている

弱い点

- アンサンブル学習で何が行われているのか、なぜ必要なのか理論的な言及が少なかった
- DPアルゴリズムと比較して、計算量が軽いことを強調しているが、ほかの手法と比べてどうなのか示されていなかった
- 異なるアルゴリズムを用いた結果が一致していることから、アルゴリズムの推論の正当性を主張しているが、妥当性に疑問