対称的な系列集合を用いた教師なし構文解析モデルの分岐バイアスの検証

石井太河 [1]　宮尾祐介 [1]
[1] 東京大学
{taigarana,yusuke}@is.s.u-tokyo.ac.jp

概要

本研究は，教師なし構文解析モデルの潜在的な分岐バイアスを分析することを目的とする．分岐バイアスとは，構文解析モデルが学習・出力しやすい木構造の偏りであり，分析する上では，まず学習データの分岐情報の偏りを明確にすることが重要である．しかしながら，教師なし学習の設定では，学習データは系列の集合のみであり，木構造は直接与えられないため，分岐情報の制御が困難である．そこで，本研究では，分岐情報の偏りの無い対称的な系列集合を用いて学習を行い，モデルの出力木構造の偏りを分析する．人工データを用いた実験の結果，モデルによって異なる分岐バイアスが確認された．

1　はじめに

本研究は，教師なし構文解析モデルが「潜在的な分岐バイアスを持たない」ための必要条件を検証することを目的とする．

教師なし構文解析とは，テキスト（系列の集合）のみから，その背後にある木構造を出力するというタスクであり，低資源言語への対応や認知的・言語学的な視点から研究されてきた [1, 2, 3, 4, 5]．

分岐バイアスとは，構文解析モデルが学習・出力しやすい木構造に偏りがあることを指す．特に，木構造が右に深いときを**右分岐**，左に深いときを**左分岐**と言う（図 1下部）．ときに，構文解析モデルが分岐バイアスを持つことは望ましくない．というのも，モデルがある種の構造のみを学習しやすいということは，他の構造を持つ言語への適用性が下がることを意味するからである．例えば，英語は右分岐な言語として知られるが，日本語は左分岐とされる．右分岐バイアスを持つモデルは英語に対する精度が高くなりやすい一方で，日本語に関しては精度が低くなりやすい．実際に，Li ら [6] は，言語により高精度となるモデルが異なることを報告しており，モデルが分岐バイアスを持つことが示唆されている．以上のように，潜在的な分岐バイアスはモデルのある種の「アドホックさ」として考えられ，モデルの分岐バイアスを分析することは重要である．

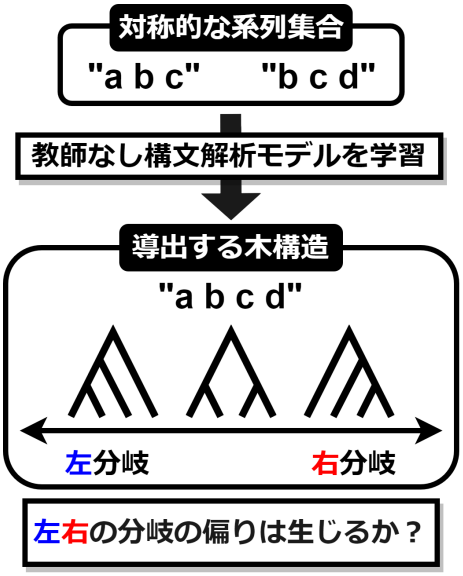

図 1　本研究の概要図

そこで，本研究では，分岐の偏りの無い対称的な系列集合を学習データに用いる分析手法を提案する（図 1）．これにより，学習の結果，モデルの導出する構造に偏りが生じる場合は，モデルそのものに分岐バイアスが内在することを示すことが可能となる．また，データセット中の系列の頻度を調整することで，対称性がくずれた場合のモデルの挙動についても分析を行う．

2　既存の分岐バイアス分析の問題点

教師なし構文解析において，モデルの潜在的な分岐バイアスを分析することは簡単ではない．というのも，分析を行う上では，「モデルそのもののバイアス」と「モデルがデータから学習したバイアス」

の 2 つを分離する必要があり，このためには学習データに含まれる情報を制御することが重要である．しかし，教師なし構文解析において学習データは系列の集合のみであり，木構造を明示的に制御することができない．

既存のバイアス分析は 3 つのアプローチに大別されるが，それぞれに問題点がある．

2.1 アーキテクチャ固有の分析

例えば，Dyer ら [7] により，PRPN [8] というモデルが右分岐バイアスを持つことが数学的に証明されている．しかしながら，Dyer ら [7] の手法は PRPN モデルに特化したものであり，任意のモデルに対して理論的分析を行うことは困難である．

2.2 自然言語を用いた分析

構文解析においては，モデルを評価する際に複数の異なる言語のコーパス [9, 10] を用いることが一般的であるが，言語間の差異は木構造の分岐の偏りに限らないため，分岐バイアスのみを評価することは難しい．そこで，Li ら [11] は，自然言語コーパス D と D 中の各系列を反転させたコーパス D^{-1} のそれぞれでモデルを学習させ，その精度の差を見ることで分岐バイアスの分析を行った．これにより，語順のみが異なり木構造の分岐が逆であるデータセットによる評価が可能となる．しかし，依然として，自然言語データは意味的にも複雑であり，語順の異なる元の系列と反転した系列の差が厳密に木構造の差だけであるかは明らかではない．

2.3 形式言語を用いた分析

Jin ら [12] は，左・右分岐な文脈自由文法で生成された言語を用いて分岐バイアスを分析している．しかし，左分岐な文法で生成される言語を右分岐の異なる文法で生成することも可能であるため，文法そのものの分岐によって系列の集合の分岐情報を制御することが可能かどうかは自明ではない．[1)]

3 対称的な系列集合を用いた分析

上述の問題点に対応するため，本研究では，分岐情報の偏りのない対称的な系列集合を人工的に構成し，それを学習データとすることを提案する．これにより，「モデルがデータから学習するバイアス」を制御し，「モデルそのものの分岐バイアス」のみを分析できることが期待される．

データセットを構成するにあたり，本研究では，コーパスレベルと系列レベルの 2 つのレベルの対称性について着目する．V を有限な語彙として，これらは以下のように定義される．

定義 1 系列集合 $D \subseteq V^*$ が**コーパスレベル対称**であるとは，語彙上の全単射 $\phi : V \to V$ が存在して，$\phi(D^{-1}) = D$ を満たすことを言う．ここで，D^{-1} は系列集合 D 中の系列を全て反転させたものであり，$\phi(D^{-1})$ は系列集合中の語彙を全て ϕ によって置換したものを表す．

定義 2 系列集合 $D \subseteq V^*$ が**系列レベル対称**であるとは，$\forall s \in D.s = s^{-1}$ を満たすことである．ここで，s^{-1} は系列 s の反転である．

系列レベル対称性は文字通りコーパス中の各系列が逆から読んでも同一であることを意味する．一方で，コーパスレベル対称性はコーパス中の語彙が反転に対して対称的であることを意味する．また，系列レベル対称性よりも条件が緩く，コーパス中の各系列が対称である必要はない．[2)] 例えば，以下の 3 つの系列集合 $x_0 \equiv \{\text{“}a\ b\ b\ a\text{”}, \text{“}b\ c\ c\ b\text{”}\}$, $x_1 \equiv \{\text{“}a\ b\ c\text{”}, \text{“}b\ c\ d\text{”}\}, x_2 \equiv \{\text{“}a\ b\text{”}, \text{“}a\ c\text{”}\}$ に関して，x_0 は系列レベル対称である．x_1 は系列レベル対称ではないが，$\phi : a \mapsto d, b \mapsto c, c \mapsto b, d \mapsto a$ の対応によりコーパスレベル対称である．一方，x_2 はコーパスレベル対称でもない．

4 実験設定

4.1 データセットの構成

本研究では，実験にあたり 2 種の対称的な系列集合を構成する．さらに，系列の出現頻度による影響も分析するため，頻度分布の異なるデータセットも生成する．

4.1.1 コーパスレベル対称なデータセット

まず，系列レベルでは対称でないがコーパスレベルで対称なデータセット D_{corpus} を構成する．できるだけシンプルな構成にするために，テストデータは L 個の異なる要素からなる 1 つの系列 s^{test} とし，

1) 付録 A に具体例を付す．

2) 例えば，$D^{-1} = D$ や $\phi(s^{-1}) = s$ のように，今回扱った 2 つのレベルの対称性の間に中間的な強さの対称性を考えることもできるが，今回は簡単のため扱わないこととする．

学習データはテストデータの M-gram 系列の集合とする．具体的には，以下のように定式化される：

$$s^{\text{test}} \equiv \text{“}v_0\, v_1 \ldots v_{L-1}\text{”}\ (\forall i, j. v_i \neq v_j) \tag{1}$$

$$S_M \equiv \{s^{\text{test}}_{i:i+M} \mid i = 0, \ldots, L-M\} \tag{2}$$

$$D_{\text{corpus}} \equiv \text{upsample}(S_M, N, w) \tag{3}$$

ここで，S_M は置換 $\phi(v_i) = v_{L-1-i}$ により対称であり，upsample は全体が N 個のデータになるように重み w を元にアップサンプルする処理である．

学習データセット中の系列の頻度制御のため，アップサンプルの際に各 M-gram 系列 $s^{\text{test}}_{i:i+M}$ に対し，以下のように i に関して線型な重み付けを行う：[3)]

$$w_i \equiv 1 + \alpha \cdot \left(i - \frac{L-M}{2}\right). \tag{4}$$

例えば，$\alpha = 0$ のときは頻度分布は一様であるが，α が大きくなると，右側の M-gram 系列は増加し，左側のものは減少する．

実験においては，$N = 3000$ とし，$L \in \{10, 20\}$，$M \in \{2, 5, 8\}$，右端の M-gram の重みが $w_{L-M} \in \{0.2, 0.6, 1.0, 1.4, 1.8\}$ となるような α の全ての組み合わせについて実験を行った．

4.1.2 系列レベル対称なデータセット

次に系列対称なデータセット D_{seq} を構成する．具体的には，上で定義した D_{corpus} の各系列 s に反転 s^{-1} を結合することによって生成する：$D_{\text{seq}} \equiv \{s \cdot s^{-1} \mid s \in D_{\text{corpus}}\}$．[4)]

なお，D_{seq} は元となる D_{corpus} に対して系列長が 2 倍になるため，実験においては $L \in \{5, 10\}$，$M \in \{1, 2, 4\}$ として元の D_{corpus} を生成することで系列長を同程度にした．

4.2 分析対象のモデル

本研究では，教師なし構文解析において代表的である，DIORA [13]，PRPN [8]，URNNG [14] の 3 つのモデルを分析対象とする．DIORA はチャートベースの再帰的なオートエンコーダであり，PRPN はゲート機構を用いることで言語モデリングの際に構造を明示的に学習するモデルである．URNNG は遷移型のモデルであり，言語モデリングの際に明示的に再帰的な木構造をモデル化する RNNG [15] の教師なし版である．なお，これら 3 つのモデルの出力は二分木となっている．

3) 頻度分布が一様でない場合，D_{corpus} の対称性は崩れる．
4) D_{corpus} と異なり，D_{seq} は頻度分布に偏りがあっても対称性は崩れない．

本研究で使用するデータセットの語彙が小さいことと計算量の問題から，学習にあたりモデルのハイパーパラメータはデフォルトのものより次元数を小さいものを使用した．[5)] 実験においては，各モデルを 30 個の異なるランダムシードで最大 30 エポックずつ学習を行い，平均的な振る舞いを分析する．

4.3 評価指標

本研究では，モデルの精度ではなく分岐バイアスの分析を目的とするため，系列の正解構造を仮定せずにモデルの導出する木構造の形を直接評価する．木構造の評価指標は様々あるが [16]，既存の指標である Corrected Colles index に加え，本研究で提案する Left leaf proportion の 2 つの指標を用いる．

Colles index Corrected Colles index [17] は，左右の部分木の葉数の差で二分木の均衡さを表す指標であり，以下のように定義される：

$$\text{Colles}(T) \equiv \frac{2}{(n-1)(n-2)} \sum_{v \in T} |n_{v_{\text{L}}} - n_{v_{\text{R}}}|. \tag{5}$$

ここで，$v \in T$ は T のノードであり，$n_{v_{\text{L}}}, n_{v_{\text{R}}}$ はそれぞれ v の左右の部分木の葉数である．Colles(T) は木が不均衡であるほど 0 から 1 に近づく．

Left leaf proportion Colles(T) では，木が具体的に左右のどちらに分岐しているのかは明らかでない．そこで，本研究では，「左の子となる葉が多いほど，木は右に分岐が深い」という直感の元，以下のような指標 Left leaf proportion（LLP）を用いる：

$$\text{LLP}(T) \equiv \frac{l_T - 1}{n_T - 2}. \tag{6}$$

ここで，n_T は T の葉数，l_T は葉のうち左の子になっているものの数であり，LLP(T) は右分岐であるほど 0 から 1 に近づく．また，系列の両端はかならず左・右の子になるためカウントから除外している．

5 結果と議論

データセットによってやや異なる結果が観察されたが，ここでは代表的な傾向の分かる D_{corpus}（$L = 20, M = 8$）と D_{seq}（$L = 5, M = 2$）の結果をそれぞれ図 2 と図 3 に示す．[6)]

5.1 分岐バイアスの検出

まず，$w_{L-M} = 1$ の頻度分布の偏りの無い設定での結果を分析する．

5) 詳細は付録 B に記載した．
6) その他の結果の抜粋を付録 C に記載する．

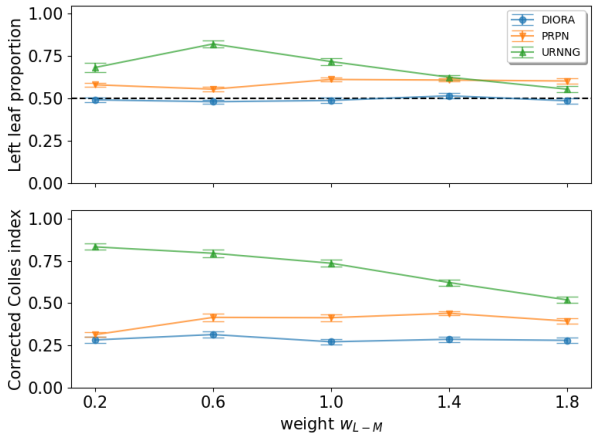

図 2 コーパスレベル対象なデータセット D_{corpus}（$L = 20, M = 8$）での結果．エラーバーは標準誤差を示す．

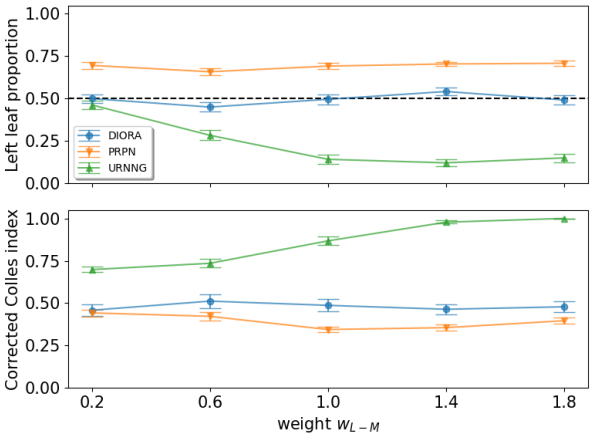

図 3 系列レベル対象なデータセット D_{seq}（$L = 5, M = 2$）での結果．エラーバーは標準誤差を示す．

LLP PRPN はどの設定でも LLP > 0.5 となり，右分岐バイアスが確認される．DIORA に関しては，どの設定でも LLP ≈ 0.5 であり，分岐バイアスを持たない必要条件を概ね満たしていると言える．一方，URNNG はデータセットによって異なる傾向が観察された．例えば，図 2上部にあるように，D_{corpus} ではどの設定でも右分岐な傾向が見られるが，一部の D_{seq} ではバイアス無し，あるいは左分岐（図 3下部）な傾向も見られ，URNNG が分岐バイアスを持つものの，左右の偏りはデータセットによって変化しうるということが明らかになった．

Colles index Colles index に関しては，URNNG は他のモデルに比べ，どのデータセットでも相対的に高く木構造の偏りが強いことが確認される．一方で，DIORA と PRPN に関しては，相対的な差がデータセットによって異なることが観察された．LLP での相対的な差と異なっているのは，LLP と違い Colles index は木の根に近い部分の分岐ほど強い影響を受ける [18] ことが関係していると考えられる．

5.2 頻度分布の影響

PRPN と DIORA は頻度分布の偏りによって対称性が崩れる D_{corpus} においても，結果に大きな影響が見られない（図 2）．一方で，URNNG は系列の頻度に強く影響されることがわかる．例えば，図 2では LLP，Colles index ともに右肩下がりになっている．ただし，同じ D_{corpus} であっても，設定によっては，右肩上がりな傾向も見られ，変化の方向性に関しては統一的な結論を下せない．また，頻度分布に偏りがあっても対称性が保持される D_{seq} であっても，URNNG の挙動が大きく変わること（図 3）は，URNNG がデータセットの分岐情報の偏りではない要素に対して不安定なのではないかと推測される．

5.3 先行研究の結果との比較

PRPN の右分岐バイアスは Dyer ら [7] によって数学的に証明されており，今回の実験ではこれと無矛盾な結果が得られたため，本研究の分析法の妥当性が支持される．また，Li ら [6] は，URNNG，PRPN は英語データセットで高精度である一方，日本語データセットでは DIORA が高精度であることが報告している．多くの設定で右分岐バイアスを持つ URNNG や PRPN と比較して，DIORA は分岐バイアスを持たない傾向があることが精度の差の要因ではないかと推測される．

6 結論と今後の展望

本研究では，分岐情報の偏りの無い対称的な系列集合を構成することで，PRPN や URNNG といったモデルが分岐バイアスを持つことを検証できた．そして，いずれのデータセットでも目立った分岐バイアスを示さなかった DIORA は「分岐バイアスを持たない」ことの必要条件を満たしていると言える．より詳細に分岐バイアスを分析するには，性質の異なる系列集合を網羅的に構成したり，分析をモデル固有のものへまで拡張することが必要になる．

また，本研究では，頻度分布の偏りによって対称性を崩した場合の実験も行ったが，この操作により分岐情報にどの程度偏りが生じたのかは明らかではない．今後，教師なし構文解析モデルを分析する上では，系列集合そのものに対して分岐の偏りを定義・測定することが重要な課題になる．

謝辞

本研究は，JST，CREST，JPMJCR2114 の支援を受けたものです．

参考文献

[1] Menno van Zaanen. ABL: Alignment-Based Learning. In **COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics**, 2000.

[2] Yoav Seginer. Fast Unsupervised Incremental Parsing. In **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics**, pp. 384–391, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] Dan Klein and Christopher D. Manning. A Generative Constituent-Context Model for Improved Grammar Induction. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 128–135, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[4] Steven Cao, Nikita Kitaev, and Dan Klein. Unsupervised Parsing via Constituency Tests. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4798–4808, Online, November 2020. Association for Computational Linguistics.

[5] Kewei Tu, Yong Jiang, Wenjuan Han, and Yanpeng Zhao. Unsupervised Natural Language Parsing (Introductory Tutorial). In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts**, pp. 1–5, online, April 2021. Association for Computational Linguistics.

[6] Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. An Empirical Comparison of Unsupervised Constituency Parsing Methods. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3278–3283, Online, July 2020. Association for Computational Linguistics.

[7] Chris Dyer, Gábor Melis, and Phil Blunsom. A Critical Analysis of Biased Parsers in Unsupervised Parsing. **arXiv:1909.09428 [cs]**, September 2019.

[8] Yikang Shen, Zhouhan Lin, Chin-wei Huang, and Aaron Courville. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In **International Conference on Learning Representations**, February 2018.

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[10] Alastair Butler, Zhu Hong, Tomoko Hotta, Ruriko Otomo, Kei Yoshimoto, and Zhen Zhou. Keyaki treebank: phrase structure with functional information for japanese. In **Proceedings of Text Annotation Workshop**, p. 41, 2012.

[11] Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. On the Branching Bias of Syntax Extracted from Pre-trained Language Models. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 4473–4478, Online, November 2020. Association for Computational Linguistics.

[12] Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. Unsupervised Grammar Induction with Depth-bounded PCFG. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 211–224, 2018.

[13] Andrew Drozdov, Patrick Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive Auto-Encoders. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1129–1141, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised Recurrent Neural Network Grammars. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1105–1117, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[16] Mareike Fischer, Lina Herbst, Sophie Kersting, Luise Kühn, and Kristina Wicke. Tree balance indices: A comprehensive survey, September 2021.

[17] Stephen B. Heard. Patterns in Tree Balance Among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. **Evolution**, Vol. 46, No. 6, pp. 1818–1826, 1992.

[18] Mark Kirxpatrick and Montgomery Slatkin. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. **Evolution**, Vol. 47, No. 4, pp. 1171–1181, 1993.

A　文脈自由文法による系列の構造の特徴づけの難しさ

ここでは，Jin ら [12] が分岐バイアスを分析するのに使用した左分岐な文脈自由文法で定まる言語が右分岐な文脈自由文法で表現できることを示す．Jin ら [12] の例は以下のような左分岐の確率文脈自由文法で記述される：

$$S \xrightarrow{1} X,\ X \xrightarrow{p} X\,Y,\ X \xrightarrow{1-p} a,\ Y \xrightarrow{1} b. \tag{7}$$

これに対し，以下のような右分岐の確率文脈自由文法を考える：

$$S \xrightarrow{1} X,\ X \xrightarrow{q} a\,Y,\ X \xrightarrow{1-q} a,\ Y \xrightarrow{r} b\,Y,\ Y \xrightarrow{1-r} b. \tag{8}$$

これらは両者とも言語 $\{a \cdot b^n \mid n \in \mathbb{N}\}$ を表現する．さらに，$q = r = p$ とすると，系列 $a \cdot b^n$ を生成する確率は両者とも $(1-p) \cdot p^n$ となる．

B　モデルの学習設定

ここでは，本研究の分析対象となるモデルのハイパーパラメータや学習設定について述べる．

DIORA　max_epoch = 30, hidden_dim = 50 とし，それ以外は著者実装[7]のデフォルト値のままである．また，DIORA はデフォルトで事前学習済み単語埋め込みを使用する設定であるので，本研究で人工データで実験する際には one-hot 埋め込みを使用した．

PRPN　epochs = 30, emsize = 25, nhid = 50 とし，それ以外は著者実装[8]の教師なし構文解析用のデフォルト値のままである．

URNNG　num_epochs = 30, w_dim = 50, h_dim = 50, q_dim = 50 とし，それ以外は著者実装[9]の教師なし構文解析用のデフォルト値のままである．また，著者実装において URNNG は validation データセットを使用しているが，本研究では学習データの情報をコントロールするため，validation データセットを使用しないように修正を加えた．

さらに，上記のモデルのいずれにも，backpropagation を行う際の損失のエポック間の差が 1.0×10^{-5} 以下になった時点で収束したと判定し，学習を停止した．

C　その他の実験結果

ここでは，本文に記載していない他の設定での結果を抜粋して掲載する．

7）https://github.com/iesl/diora
8）https://github.com/yikangshen/PRPN
9）https://github.com/harvardnlp/urnng

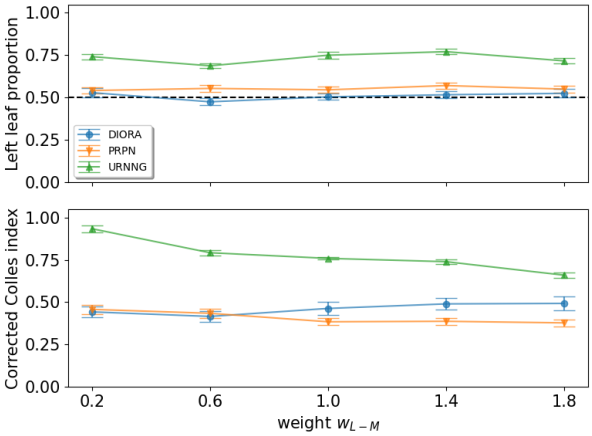

図 4　コーパスレベル対象なデータセット D_{corpus}（$L = 10, M = 8$）での結果．エラーバーは標準誤差を示す．

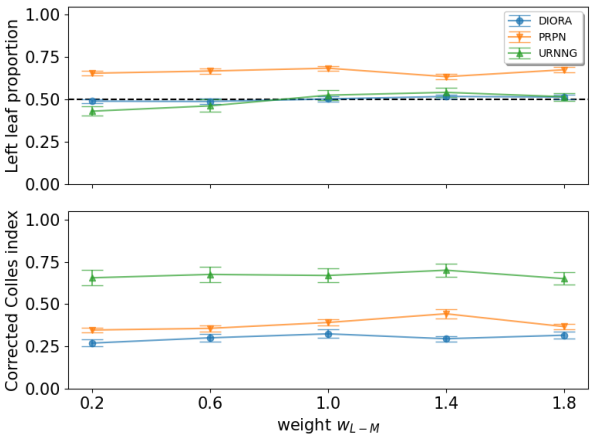

図 5　コーパスレベル対象なデータセット D_{corpus}（$L = 20, M = 2$）での結果．エラーバーは標準誤差を示す．

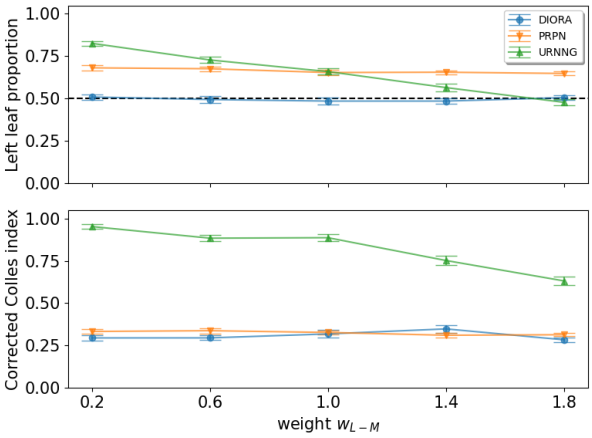

図 6　系列レベル対象なデータセット D_{seq}（$L = 10, M = 4$）での結果．エラーバーは標準誤差を示す．