談話依存構造解析の教師なしドメイン適応

西田典起
理化学研究所 AIP
noriki.nishida@riken.jp

松本裕治
理化学研究所 AIP
yuji.matsumoto@riken.jp

1 はじめに

談話依存構造解析 (Discourse Dependency Parsing) は文書内にある節や文の間の係り受けと修辞関係を同定し、文書の談話依存構造を明らかにすることを目指す [1, 2]。本研究で解析対象にする論文アブストラクトの談話依存構造の例を図 1 に示す。談話構造はこれまで文書要約 [3, 1] や質問応答 [4]、極性分類 [5, 6] などのタスクでその有用性が確認されてきた。

談話構造解析の従来法は RST-DT [7] などのツリーバンクを使い、解析器を学習する [8, 9]。しかし、既存のツリーバンクのサイズおよびそれらがカバーするドメインは限定的なため、一般的には解析したい文書はそれらのドメインから外れ、解析精度は低下する。よって、利用できる談話構造ツリーバンクのドメインと解析器を適用したいドメインが異なるときに、どのように適用先ドメインに解析器を適応するかという「教師なしドメイン適応」の問題は、現実的に重要な課題となっている。

本稿では、談話依存構造解析における教師なしドメイン適応問題に対する疑似ラベリング法 (Self-Training [10, 11, 12], Co-Training [13, 14, 15], Tri-Training [16], Asymmetric Tri-Training [17] など) の効果を調べ、その結果を分析する。疑似ラベリングはラベル付きデータとラベルなしデータからモデルを訓練する半教師あり学習の代表的な方法論である。本研究では特に科学技術論文アブストラクトの談話依存構造解析を対象に、自然言語処理アブストラクト (ソースドメイン) のツリーバンクで訓練された解析器を、医学生物学アブストラクト (ターゲットドメイン) に適応することを考える。

実験の結果、疑似ラベリング、特に係り受けの一致率に基づいて合意基準を緩和化した Asymmetric Tri-Training が談話依存構造解析の教師なしドメイン適応に有効であることを確認した。解析精度は Labeled Attachment Score で 4.1 ポイント向上し、さらに人手による談話構造のアノテーションコストが大幅に (e.g., 約 60%) 削減できることがわかった。また、疑似ラベルの質と量の影響力は大きく、閾値による合意基準の調整によってそれらの最適なバランスを実現できることを示した。また、解析結果を分析し、解析エラーの主要因が (1) 係り受けパターンの頻度の偏り、(2) 談話関係クラスの曖昧性、(3) ドメインごとの論旨の展開の違いの 3 つである可能性があることを確認した。

2 方法

2.1 談話依存構造解析の疑似ラベリング

談話依存構造解析は、EDU の系列として表現された入力文書 $d = e_0, e_1, \ldots, e_n$ に対し、談話依存構造 $T = \{(h, m, l) \mid 0 \leq h \leq n, 1 \leq m \leq n, l \in \mathcal{L}\}$ を出力するタスクである。ここで e_0 は根ノードを表し、係り受け (h, m, l) は e_h (head) と e_m (modifier) が談話関係 $l \in \mathcal{L}$ によって結合することを表す。

本研究の疑似ラベリングは 4 ステップからなる。

1. 少量のラベル[1)]付きデータ (ゴールドデータ) の集合 $L = \{\langle d, T\rangle\}$ を使い、解析器 f を訓練する。
2. その解析器を使い、大量のラベルなしデータ $d' \in U$ に対して疑似ラベル $T' = f(d')$ を付与する。
3. 疑似ラベル付きデータの一部からシルバーデータセット $L' = \{\langle d', T'\rangle\}$ を構成する。
4. ゴールドデータとシルバーデータを合わせて、解析器の再訓練を行う。

教師なしドメイン適応の問題設定では、ラベル付きデータについてはソースドメインのみに存在し、ラベルなしデータはテストデータと同じターゲットドメインから収集したものとする。

疑似ラベリングの方法として Self-Training, Co-Training, Tri-Training, Asymmetric Tri-Training の 4 つの

1) 本研究では、「ラベル」は談話依存構造を指す。

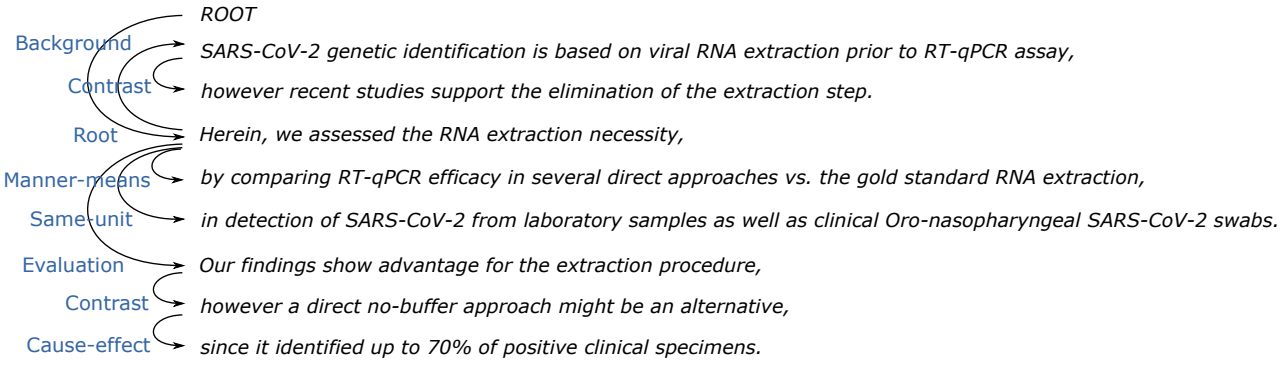

図 1　科学技術論文アブストラクトの談話依存構造の例。

アルゴリズムを実装し、比較する。

Self-Training (ST) [10, 11] は最も単純な疑似ラベリング法である。本研究では Reichart と Rappoport (2007) [12] に従い、すべての疑似ラベルをシルバーデータとして用いた。ST は単一の解析器を用いるため、自己の誤り傾向を軌道修正する仕組みがなく、学習過程で誤りが増幅する可能性がある。

Co-Training (CT) [13, 14, 15] は二つの異なる解析器を使い、一方の解析器の予測結果をもう一方の解析器のシルバーデータとして活用する。

Tri-Training (TT) [16] は三つの解析器を使い、二つの解析器の予測結果が一致 (合意) している場合のみ、それを残り一つの解析器のシルバーデータとして採用する。すなわち、上記の ST や CT と異なり、疑似ラベル T' のうち二つの解析器の間で合意が得られたもののみを再訓練で用いる。

Asymmetric Tri-Training (AT) は TT の拡張であり、斎藤ら (2017) [17] によって教師なしドメイン適応の方法として開発された。AT では三つの解析器のうち一つをターゲットドメイン専用の解析器 f_t として訓練する。具体的には、二つの解析器 f_1 と f_2 による予測結果 T'_1, T'_2 が一致しているときのみ、それをシルバーデータセット L' に追加し、f_1 と f_2 は $L \cup L'$、f_t は L' のみで再訓練する。

2.2　合意基準の緩和化

TT および AT では一般に予測した「ラベル」が完全一致しているかどうかで合意判定されるが、談話構造解析では「ラベル」は文書レベルの木構造であるため、完全一致に基づく合意基準は非常に厳しい。その結果、ほとんどの疑似「ラベル」はシルバーデータとして採用されず、捨てられてしまう。

そこで、本研究では局所的に一致している係り受けの割合 (一致率) がある閾値以上の場合には合意していると判定する。具体的には、$\frac{2|T'_i \cap T'_j|}{|T'_i|+|T'_j|} \geq \tau$ を満たす場合のみ、T'_i と T'_j の両方ともをシルバーデータセットに加える, i.e., $L'_k \rightarrow L'_k \cup \{\langle d', T'_i\rangle, \langle d', T'_j\rangle\}$. このような合意基準の緩和化によって、局所的に合意を得ている係り受け情報の活用を目指す。

2.3　談話依存構造解析器

Zhou と Goldman (2004) [15] によれば、CT においてデータの *view* を明示的に二つに分けない場合、異なる inductive bias をもつモデルを使う必要がある。また、モデル間の同意に基づく TT および AT でも、三つのモデルの多様性は重要であると考えられる。

本研究では、異なる解析戦略をもつ三つのニューラル談話依存構造解析器を実装する。具体的には、グラフ型解析器 (G)、遷移型解析器 (T_{LR})、逆向きの遷移型解析器 (T_{RL}) を実装した。紙面の都合上、各解析器の詳細は省略する。

3　実験結果と考察

3.1　データ

ソースドメインのラベル付き訓練データとして、SciDTB [18] に収録されている 742 件の NLP アブストラクトの談話依存構造を用いた。ターゲットドメインのラベルなし訓練データとして、CORD-19 [19] に収録されている 22,515 件の医学生物学アブストラクトを用いた。EDU 分割については Wang ら (2018) [20] の方法で自動的に行った。また、ターゲットドメインのテストデータとして、CORD-19 からランダムに 70 件のアブストラクトを選択し、それらに対して人手で談話依存構造をアノテーションした[2)]。以降ではこのデータセットを CORD19-DT

2)　アノテーションに関しては、著者自身が行った。今後、より公平で大規模な設定で評価を行うために、現在第三者によるアノテーション作業を行っている。

表 1　各システムの評価結果。評価尺度として Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS), Undirected UAS (UUAS) を用いる。括弧内に解析器を示す。τ は合意基準の閾値を表す。SciDTB の結果の括弧内の数値はラベルなしデータとして NLP アブストラクトを用いたときのスコアである。

	CORD19-DT			SciDTB
Method	**LAS**	**UAS**	**UUAS**	**LAS**
Seed data only (G)	53.3	67.0	71.9	61.7
Seed data only (T_{LR})	55.3	67.0	71.8	63.2
Seed data only (T_{RL})	52.4	64.2	69.8	63.4
ST (G)	54.7	67.7	72.4	61.8 (62.5)
ST (T_{LR})	57.4	69.2	74.2	62.9 (62.4)
ST (T_{RL})	55.3	66.5	71.1	60.9 (61.8)
CT (G, T_{LR})	57.8	69.9	73.6	62.9 (64.3)
TT (G, T_{LR}, T_{RL}), $\tau = 0.9$	57.0	68.9	72.1	63.2 (**65.0**)
TT (G, T_{LR}, T_{RL}), $\tau = 0.5$	58.4	**72.1**	**75.6**	**65.2** (64.2)
AT (G, T_{LR}, T_{RL}), $\tau = 0.9$	57.7	69.5	72.8	62.1 (64.0)
AT (G, T_{LR}, T_{RL}), $\tau = 0.5$	**59.4**	69.9	74.4	62.9 (64.8)

と表記する。70 件のアノテーションデータをテストセット (50 件) と検証セット (20 件) に分割した。CORD19-DT と SciDTB の特徴の比較については付録に載せる。

3.2　解析精度の比較

表 1 に各システムの評価スコアを載せる。比較のために、ターゲットドメインのラベルなしデータを活用しないベースラインシステム (Seed data only) のスコアも載せる。また、参考のために SciDTB のテストセットでの結果も載せる。

CORD19-DT での結果から、疑似ラベリングによるシステムはベースラインよりも一貫して高い解析精度であることがわかる。これは、疑似ラベリングによって、ターゲットドメインのラベルなしデータを活用して談話依存構造解析器のドメイン適応ができていることを示唆する。また、TT, AT は ST, CT よりも精度が高く、このことから複数の解析器の合意によってノイジーな疑似ラベルをフィルタリングすることが有効であるとわかる。また、本研究で導入した閾値による合意基準の緩和化も有効であった ($\tau = 0.9$ vs. $\tau = 0.5$)。

一方、SciDTB での結果では、TT ($\tau = 0.5$) を除いて、疑似ラベリングによる精度向上は見られなかった。しかし、ラベルなしデータとしてテストデータと同じドメインである NLP アブストラクトを用いた場合、ST を除いて疑似ラベリングによる精度向上が見られた。このことから、疑似ラベリングを用いる場合、テストデータと同じドメインのラベルなしデータを用いることが重要であると判断できる。

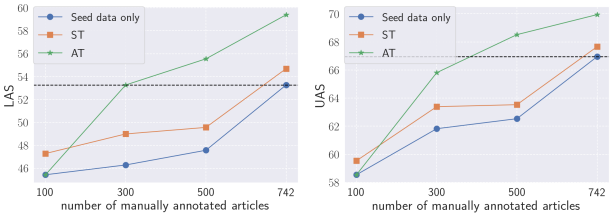

図 2　ラベル付きデータの個数と解析精度の関係。

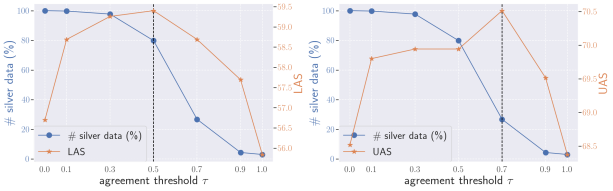

図 3　AT における合意基準の閾値と、シルバーデータの量および解析精度の関係。量は全ラベルなしデータ数に対する割合で表す。

3.3　アノテーションコストの削減率

図 2 にラベル付きデータの個数と解析精度 (LAS, UAS) の関係を示す。ラベル付きデータの個数の変化に対し、疑似ラベリングは一貫してベースラインを上回っている。また、AT の精度は ST よりもほぼ一貫して高い。以上から、人手による談話構造データが少ないときにも、談話依存構造解析の教師なしドメイン適応に対し疑似ラベリングは有効であることがわかる。

また、注目すべきことに、図 2 は AT によって大幅にアノテーションコスト減らせることを示唆する。例えば、ベースラインシステムが 54.7%の LAS を実現するためには 742 件のラベル付きデータが必要であるのに対し、AT を用いることで必要なラベル付きデータの個数を 300 件まで減らすことができる。これは、AT によって、$\frac{742-300}{742} \times 100 \fallingdotseq 60\%$のアノテーションコストの削減が可能であることを表す。

3.4　シルバーデータの量と質の重要性

AT における合意基準の閾値 τ と、シルバーデータの量および解析精度 (LAS, UAS) の関係を調べた結果を図 3 に示す。閾値が低すぎれば、解析器が予測する談話依存構造のうち一致率 (信頼性) が低いものまでシルバーデータに追加されてしまい、結果

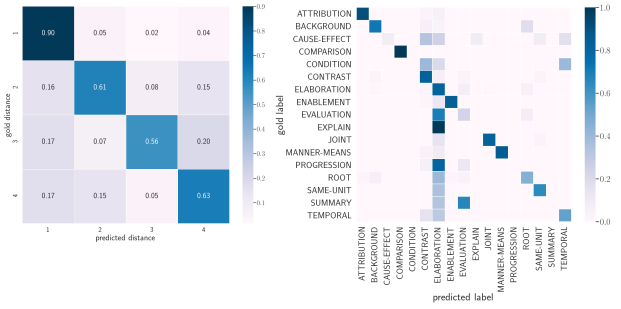

図 4 AT による解析結果の混合行列。縦軸および横軸はそれぞれ人手またはシステムによる係り受けに対応する。各行で正規化している。

として解析精度が低くなってしまっている。一方、閾値が高すぎれば、ほとんどの談話依存構造が捨てられてしまい、結果としてシルバーデータの量が少なくなりすぎてしまい、ドメイン適応を行うのが難しくなっている。すなわち、閾値によってシルバーデータの量と質が制御され、閾値をうまく調整することで量と質の最適なバランスを実現することができる。

3.5 エラー分析

AT を用いたシステムの解析エラーについて詳しく分析した結果、多くのエラーが以下の 3 つに起因することがわかった。

一つめの要因は、係り受けパターンの頻度の偏りである。低頻度のパターン (e.g., 長さ ≥2, ラベル=Evaluation) を誤って最頻出の係り受けパターン (i.e., 長さ=1, ラベル=Elaboration) として解析する傾向が見られた。図 4 は、CORD19-DT における解析結果の混合行列である。それぞれ、係り受けの長さと談話関係ラベルに対応する。長さ 1 の係り受けの 90%については同定できているのに対し、長さ 2 以上の係り受けについては約 60%しか同定できていない。また、談話関係の誤分類のほとんどは Elaboration として予測されており (Elaboration の列を参照)、実際に Elaboration は SciDTB 中で最頻出の談話関係クラスである (付録参照)。

二つめの要因は、談話関係クラスそのものの曖昧性である。例えば、Elaboration と Progression, Cause-effect と Explain の識別は常に自明であるとは限らず、これらはしばしば人間にとっても識別するのが難しい。図 5 は、CORD19-DT のデータに対するシステムの解析結果の例である。EDU 間の係り受け (エッジ) については正しく同定できているが、エッジ $(1, 2)$ についてその関係クラスを Cause-effect

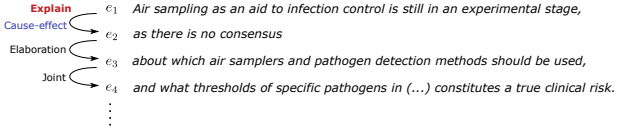

図 5 CORD-19-DT における解析結果の例。赤は解析エラー、青は正解を表す。

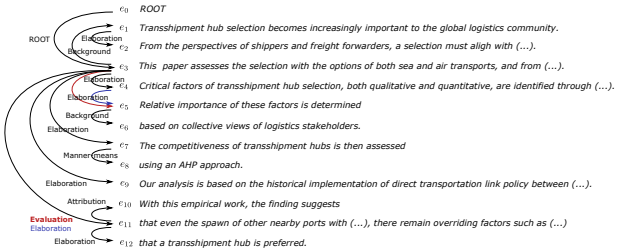

図 6 CORD19-DT における解析結果の例。

ではなく Explain と誤分類してしまっている。

最後の要因は、ドメイン (分野) ごとの論旨の展開傾向の違いである。具体例を図 6 に示す。システムは、e_3 (head) から e_{11} (modifier) への係り受けのラベルを、e_{11} の内容は e_3 の評価結果ではなく研究による知見であるにも関わらず、Evaluation と誤分類している。これは、Evaluation 関係の係り受けの頻度が SciDTB では CORD19-DT にくらべて非常に多く (7.4% vs. 1.9%)、SciDTB で訓練された解析器が「Evaluation 関係の係り受けは論文アブストラクトの終盤 EDU との間で起こる」と学習しているからだと考えられる。

4 おわりに

本稿では、談話依存構造解析の教師なしドメイン適応問題に対する疑似ラベリング法 (Self-Training, Co-Training, Tri-Training, Asymmetric Tri-Training) の検討を行った。閾値により同意基準を緩和化した Asymmetric Tri-Training が最も効果的であり、4.1 ポイントの解析精度の向上と 60%のアノテーションコストの削減が可能であることを確認した。また、閾値によって制御できるシルバーデータの量と質のバランスの重要性を示し、解析エラーの 3 つの主要因について分析した。

今後は、CORD19-DT の拡大および、論文全体の談話構造を解析する方法の開発に取り組んでいく。

謝辞

本研究は JST CREST(課題番号: JPMJCR1513) の支援を受けて行った。

参考文献

[1]Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[2]Mathieu Morey, Philippe Muller, and Nicholas Asher. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, Vol. 44, No. 2, pp. 197–235, 2018.

[3]Annie Louis, Aravind Joshi, and Ani Nenkova. Discourse indicators for content selection in summarization. In *SIGDIAL'10*, 2010.

[4]Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[5]Livia Polanyi and Martin Van den Berg. Discourse structure and sentiment. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011.

[6]Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[7]Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[8]Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[9]Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. CODRA a novel discriminative framework for rhetorical analysis. *Computational Linguistics*, Vol. 41, No. 3, pp. 385–435, 2015.

[10]David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

[11]David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2006.

[12]Roi Reichart and Ari Rappoport. Self-training for enhancement and domain adaptationa of statistical parsers train on small datasets. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[13]Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 1998.

[14]Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paula Ruhlen, Steven Baker, and Jeremiah Crim. Bootstrapping statistical parsers from small datasets. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

[15]Yan Zhou and Sally Goldman. Democratic co-learning. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, 2004.

[16]Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 11, 2005.

[17]Kuniaki Saito, Toshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of The 34th International Conference on Machine Learning*, 2017.

[18]An Yang and Sujian Li. SciDTB: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[19]Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

[20]Yizhong Wang, Sujian Li, and Jingfeng Yang. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

A　付録: CORD19-DT vs. SciDTB

本付録では、本研究で構築している CORD19-DT と、Yang ら (2018) によって構築された SciDTB の特徴の違いについて報告する。これら二つはそれぞれ医学生物学論文または自然言語処理論文のアブストラクトに対して人手で談話依存構造をアノテーションしており、ここで示す特徴の違いはドメイン間のギャップの大きさ、ドメイン適応の難しさを表す。

表 2 に CORD19-DT と SciDTB の特徴を載せる。医学生物学のアブストラクトの長さ (i.e., 文数、トークン数) は NLP アブストラクトよりも長い傾向にあることがわかる (7.2 文 vs. 5.3 文; 197.5 トークン vs. 130.5 トークン)。一方、平均 EDU 数については一致している (14.0 vs. 14.0)。このことから、医学生物学の EDU は平均的に NLP の EDU よりも長く (情報量が多く)、教師なしドメイン適応ではこの点が課題の一つになりうる。

係り受けの平均長 (談話関係によって結合される EDU 間の平均距離) についてはほぼ等しい (2.5 vs. 2.4)。これは、一般に長い係り受けは読み手への負荷が高いため自然言語文書において起こりにくく、この傾向は分野に依らないことを示唆する。

一方、文書 $d = e_1, \ldots, e_n$ 全体の親であり、根ノード e_0 の唯一の子である Root EDU の平均位置は、医学生物学では前から 5.9 番目であるのに対し、NLP では 3.9 番目であった。両分野ともにアブストラクトは大まかに研究背景 → 研究目的と方法 → 結果と考察という順番で記述される傾向にあると考えると、医学生物学では研究背景部分に関する記述が NLP よりも多く、対応して談話依存構造の傾向も分野間で異なることを示唆する。

各データセットにおける談話関係クラスごとの係り受けの分布 を表 3 に載せる。いくつかの談話関係については出現率が大きく異なることがわかる。Elaboration の出現率の違い (47.2 vs. 39.1) については、表 2 で見たように医学生物学アブストラクトは NLP アブストラクトよりも情報量が多く、順接で情報展開されることが多いことを反映していると考えらえる。Evaluation の出現率の違い (1.9 vs. 7.4) については、医学生物学論文の主軸は知見 (実験結果) の記述であることが多く、Root EDU そのものが知見に関するものであることが多いのに対し、NLP では主軸が開発した手法であることが多く、その評価結果についてアブストラクトの後半で別に記述する傾向にあることを反映していると考えられる。談話関係の分布は文書においてどのように論旨が展開されるのかという傾向を反映しており、このような論旨の展開傾向の違いはドメイン適応の難しさを表す。

表 2　CORD19-DT と SciDTB [18] の比較. 数値は各データセットのすべてのアブストラクト数を用いて平均化している。

	CORD19-DT	**SciDTB**
文数	7.2	5.3
トークン数	197.5	130.5
EDU 数	14.0	14.0
係り受けの長さ	2.5	2.4
Root EDU の位置	5.9	3.9

表 3　CORD19-DT と SciDTB における各談話関係クラスごとの係り受けの分布。

談話関係	**CORD19-DT**	**SciDTB**
Root	7.1	7.1
Attribution	5.1	5.9
Background	7.1	6.9
Cause-effect	1.7	2.1
Comparison	0.1	1.0
Condition	0.7	1.2
Contast	4.1	2.8
Elaboration	47.2	39.1
Enablement	5.6	7.3
Evaluation	1.9	7.4
Explain	0.1	1.3
Joint	8.8	6.8
Manner-means	3.0	5.3
Progression	2.3	1.9
Same-unit	2.4	2.5
Summary	0.9	0.2
Temporal	1.9	1.2
Total	100.0	100.0