医学生物学論文解析のための談話依存構造ツリーバンクの構築

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

概要

実用的な談話構造解析法の開発やその応用に関する研究には，信頼性の高い談話構造ツリーバンクの存在が不可欠である．しかし，既存の談話構造ツリーバンクのカバレッジはサイズやドメインの観点から十分とは言い難い．本プロジェクトでは，特に医学生物学論文の談話構造解析とそこからの知識獲得を念頭に，そのための談話依存構造ツリーバンクの構築を目指す．本稿では，GENIA コーパスに収録されている医学生物学論文要旨 1,999 件に対する談話依存構造のフレームワークとアノテーションプロセスの詳細および現在収集できているデータに基づく統計情報について，既存のコーパスと比較しながら説明する．

1 背景と目的

一貫性のある文章では，文章の構成要素 (節や文) は統語的，意味的，または論理的に孤立せず，各節は相互に様々な役割を果たしながら全体的な論旨展開を形成している．談話構造は，文書の構成要素間の関係性に基づく文章の構造的な表現である．特に本研究で扱う談話依存構造 [1, 2, 3, 4, 5, 6] は，Elementary Discourse Units (EDUs) と呼ばれる節レベルのテキストスパン (ノード) の間の係り受けと談話関係に基づいて文章をグラフ構造として表現する．談話依存構造の例を図 1 に示す．図 1 において，各矢印の始点は談話依存構造における「親」(中心部)，矢印の終点は「子」(周辺部) に対応し，談話関係 (Elaboration, Cause-Result 等) は親に対する子の働き・役割を表し，親から子へのラベル付き有向リンクとして表される．

自動解析された談話構造は文書要約 [7, 2, 8, 9, 10, 11], や文書分類 [12, 13]，質問応答 [14, 15]，関係抽出 [16] など様々な自然言語処理技術で有用であることが知られている．高精度で実用的な談話構造解析器の開発には，人手で構築された談話構造ツリーバンクが訓練データおよび評価用データとして不可欠である．しかし，既存の談話構造ツリーバンクのサイズおよびドメインに関するカバレッジは十分とは言い難い．例えば，談話構造解析の研究で最もよく用いられている RST-DT コーパス [17] には 385 文書の談話構造しかアノテーションされておらず，明らかにサイズが少ない．SciDTB コーパス [18] には自然言語処理分野の 798 論文要旨に対する談話依存構造がアノテーションされているが，分野によって語彙や論旨の展開傾向は大きく異なるため，SciDTB で訓練した解析システムによる他分野の論文要旨に対する解析精度は大きく低下してしまう．また，本プロジェクトの一環として著者らは COVID-19 に関連する医学生物学論文要旨 300 件に対して人手で談話依存構造を付与し，COVID-19 Discourse Dependency Treebank (COVID19-DTB) [19] として公開したが，そのサイズはまだまだ小さく，さらに拡張していく必要がある．

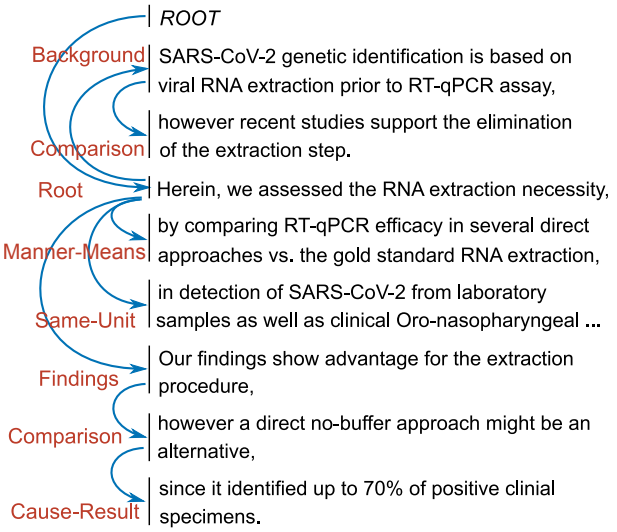

図 1 医学生物学論文要旨に対する談話依存構造の例．

本プロジェクトの目的は，医学生物学分野の論文に対する談話構造解析法およびその応用の研究開発に有用な，高品質で既存のコーパスに比べて大規模な談話依存構造ツリーバンクを構築，整備，公開す

ることである．医学生物学論文に焦点をあてることにはいくつかの利点がある．一つは，論文の文章はSNS やブログ等の文章よりも論旨が明確で論理的であることが期待されており，談話構造の研究対象として望ましい．また，膨大な医学生物学論文から有用な情報を抽出して体系化する医療知識獲得技術の開発は社会的に喫緊の課題であり，医学生物学論文を対象に談話構造解析システムを開発することの意義は大きい．

本稿では，GENIA コーパス [20] に収録されている論文要旨 1,999 件に対して談話依存構造を付与するための詳細について説明する．具体的には，談話依存構造のフレームワークとアノテーションプロセスの詳細，現状のツリーバンクに基づく統計情報について説明する．

2　談話依存構造のフレームワーク

2.1　EDU 分割ルール

本プロジェクトのアノテーションでは，まず論文要旨を Elementary Discourse Unit (EDU) と呼ばれる節 (clause) レベルのテキストスパンに分割するところから始める．各 EDU は連続した単語列からなっており，EDU 間にオーバーラップはない．あるテキストスパンが EDU の基準を満たすかどうかは，主に動詞 (述語) に基づいて判断する．ただし，“in spite of” や “due to” などのディスコースマーカーを伴う句については，独立の EDU として認める．結果的に，以下のようなケースでは EDU 分割を行う[1)]．

1. 主節，並列節
2. 接続詞で結合される従属節
3. 分詞構文 (participle clause)
4. 「目的」「結果」の意の to 不定詞，“in order to” 節，so that 節
5. 副詞的役割の「前置詞 + 動名詞」
6. 名詞を後置修飾する {分詞，to 不定詞，「前置詞 + 動名詞」}
7. 関係節
8. 同格の that 節
9. 動詞を含む相関従属節 (correlative subordinators)
10. ディスコースマーカーを伴う句

1)　各ケースの例については紙面のスペース状省略するが，公開しているガイドライン (+ アノテーションツール) で見ることができる: https://norikinishida.github.io/tools/discdep/Biomed-DTB-annotation-guideline.pdf

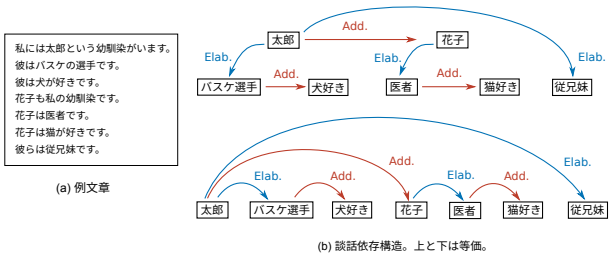

図 2　Elaboration と Addition による談話依存構造の例．

Carlson ら (2001) のマニュアル [17] に倣い，本プロジェクトでも以下のようなケースでは独立した EDU とは認めない．前節の基準を満たす場合でも，これらの例外に該当する場合は EDU 分割しない．

1. 動詞の主語・目的語・補語や，前置詞の目的語としての節 (clausal subject, clausal object, clausal complement)

   (1)　[Making computers smaller often means sacrificing memory.]

   (2)　[He is interested in climbing Everest.]

2. 分裂文・疑似分裂文 (強調構文)，外置構文など

   (3)　[It is sleep deprivation that exacerbates health problems.]

2.2　談話依存関係

本プロジェクトでは，談話関係を 13 クラスにカテゴライズする．各談話関係クラスの意味や対応するディスコースマーカーの例を表 1 に載せる．これらは，SciDTB や RST-DT, Penn Discourse Treebank [21], ISO 24617-8 [22] などをもとに，実際のアノテーション作業を通してアノテーション方針の一貫性が高くなるように設計した．

Root　論文要旨の談話依存構造では，Root 関係の子は研究目的や主要な報告内容について記述している EDU になる．すなわち，論文要旨のなかで最も中心的な EDU である．

Elaboration と Addition　Elaboration と Addition は最も一般的で頻度の高い談話関係であり，これらの違いは二つの EDU が従属的な関係にあるか，等位的な関係にあるかである．三つ以上の EDU が添加・累加・系列・同列の関係にある場合は，一つ前の EDU から直後の EDU に順番に接続していくことによってチェインを構成する．図 2 に Elaboration と Addition を用いた例を示す．

表 1　本コーパスで採用する談話関係とその意味，代表的なディスコースマーカー．

談話関係	意味	代表的なディスコースマーカー
0. Root	研究の目的，研究の主要内容	
1. Elaboration	詳細化，例示，定義	
2. Addition	添加，累加，系列，同列	also, as well as, moreover, furthermore, besides, in addition, next, then
3. Comparison	逆説，譲歩，対比，比較	but, however, although, yet, despite, whereas, instead of, alternatively, on the other hand
4. Cause-Result	原因，理由，結果	because, so, therefore, thus, due to, consequently, as a result, leading to
5. Condition	条件，仮定	if, as long as, unless, when
6. Temporal	時間，状況	when, before, after, while
7. Enablement	目的，可能化	in order to, for ...ing, so as to, so that, which enables to, which allows to
8. Manner-Means	方法，手段，手法セクション	by, using
9. Background	背景セクション	
10. Findings	実験結果セクション，結論セクション	
11. Textual-Organization	文書構造 (e.g., 見出し，タグ)	
12. Same-Unit	分離した疑似 EDU の結合	

基本談話関係　表中の Comparison から Manner-Means までは既存コーパスにおける談話関係の定義とほぼ同様である．

マクロ談話関係　Background と Findings はすこし特殊であり，論文要旨における研究背景と研究結果・結論に対応する領域 (セクション) を指し示すためのマクロな視点の談話関係とする．したがって，これらのクラスが一つの論文要旨中で複数回出現することは原則ない．

Textual-Organization と Same-Unit　Textual-Organization はタイトルやタグ等を指し示すためのものである．Same-Unit は EDU の埋め込みによって分離して非連続になってしまった EDU を結合するために用いる．

2.3　典型的な談話依存構造

医学生物学系の論文要旨に限らないが，談話依存構造には分野ごとに典型的な構造の傾向がある．一般的に論文要旨では大きくわけて (1) 背景 (Background), (2) 目的 (Objectives), (3) 手法, 実験設計 (Methods), (4) 結果，考察，結論 (Results, Discussion, Conclusions) を含んでいる．実際，論文要旨中でこれらの項目ごとに見出し (e.g., “Background:”, “Results:”) を設けている論文も数多く存在する．

大多数の論文要旨では，上記の順番通りに項目を説明している．それらは図 3 のような談話依存構造になる．また，工学系では少ないが，医学生物学などの分野では研究によって発見された知見や事例そ

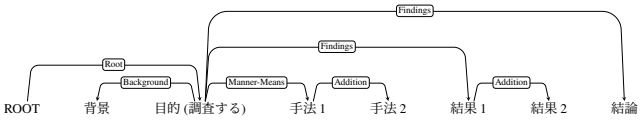

図 3　典型例 1

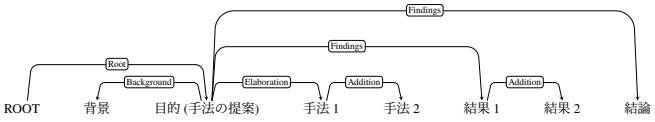

図 4　典型例 1 (b)

のものを論文 (論文要旨) の中心的な情報として置き，論文要旨はその知見と意義について記述していることがある．そのような場合は，談話依存構造は図 5 のようになる．

3　アノテーションプロセス

医学生物学論文要旨の読解および談話依存構造のアノテーションはそれぞれ高度な専門知識と技能を要求するため，信頼性の高いツリーバンクを作成するためにはアノテーションプロセスに工夫が必要になる．本プロジェクトでは，まずアノテーションガイドラインとブラウザ上で動くアノテーションツールを整備した[2)]．本ツールは SciDTB の著者ら [18] によるツールをもとに改変を行った．

本プロジェクトでは，アノテーションプロセスを

2)　https://norikinishida.github.io/tools/discdep

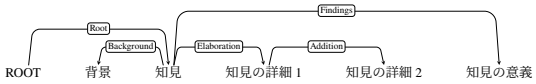

図 5　典型例 2

(1) 選別フェイズ，(2) 修練フェイズ，(3) 拡張フェイズの 3 段階に分けた．選別フェイズでは，アノテーション作業や当該分野について一定の習熟度がある 10 人のワーカーに対してガイドラインを配布し，GENIA コーパス中からランダムにサンプリングされた 20 件の論文要旨に対して 10 人のワーカーと著者によってアノテーションを行った．そして，著者による談話依存構造を正解データとし，各ワーカーの Labeled Attachment Score (LAS) を計算し，LAS が高い上位 5 名を選別した．本フェイズにおける LAS は 31.28%から 79.89%と開きが非常に大きく，上位 5 名の平均 LAS は 75.20%であった．

次の修練フェイズでは，GENIA から新たにランダムサンプリングした 20 件に対して，前フェイズで選別された 5 名のワーカーと著者によって再度アノテーションを行った．実際のアノテーションに入る前に，疑問点やアノテーションツールに関する要望を把握するために各ワーカーと会議を行い，ガイドラインやツールに対して適宜変更を行った．また，本フェイズでは Google Form と Google Spreadsheet を用いてワーカーが疑問点等を投稿し，著者らと迅速にディスカッションできるようにした．再度著者によるアノテーション結果を正解として各ワーカーのアノテーション結果の LAS と定性的な質が水準を満たすまで本フェイズを繰り返した．本フェイズの結果，最高 LAS は 82.13 に上昇し，5 名の平均 LAS も 76.28 に上昇した．

最後の拡張フェイズでは，修練フェイズの基準をクリアしたワーカーと著者によって，GENIA からランダムサンプリングされた数百件に対するアノテーションを行った．本フェイズでは，各ワーカーはそれぞれ異なる論文要旨集合に対してアノテーションを行う．アノテーション一致率の計算については，著者によってすべての論文要旨をアノテーションすることで行う予定である．

4　統計情報

本節では，作成中のツリーバンク (GENIA-DTB) の統計的な情報を既存の談話依存構造ツリーバンクである SciDTB [18] と比較する．本プロジェクトでは現在拡張フェイズが進行中のため，完了した修練フェイズまでのデータを対象に統計情報を求める．

	GENIA-DTB*	SciDTB
Avg number of EDUs / doc	17.6	14.0
Avg dependency distance	2.6	2.5
Avg Root position	4.6	4.0

表 2　本プロジェクトで構築中の GENIA-DTB と SciDTB の統計情報．アスタリスクは構築途上であることを示す．

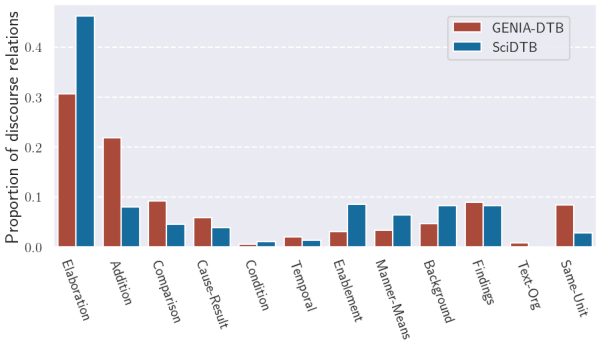

図 6　談話関係の分布の比較．

表 2 から，GENIA の論文要旨のほうが SciDTB の論文要旨よりも長い傾向があることがわかる．しかし，談話依存関係の長さ (親と子の間の距離) の平均については両コーパスでほぼ等しく，談話依存関係の長さは文書長に強く依存しないことがわかる．また，Root EDU の子，すなわち論文要旨中で最も中心的な EDU の位置は GENIA-DTB ではわずかに後方に位置する傾向があることがわかる．談話関係の分布については，図 6 から，GENIA-DTB では Addition と Same-Unit の頻度が SciDTB よりも多いことがわかる．これらの結果から，医学生物学論文要旨は，自然言語処理論文要旨よりも長く，節の埋め込みによる複雑な文の構造を持ち，より等位的な情報の追加が多いことがわかる．このことは，医学生物学分野を対象にした談話依存構造ツリーバンクを構築することの意義を示唆している．

5　今後の計画

今後は，拡張フェイズを引き続き行い，来年度上半期中での合計 1,999 件のアノテーション収集を目指す．また，COVID19-DTB を今回のアノテーションフレームワークにあわせて更新し，GENIA-DTB と将来的に統合することを計画している．

謝辞

本研究は JSPS 科研費 21K17815 の助成を受けたものです。本研究の一部は、JST、AIP 日独仏 AI 研

究、JPMJCR20G9 の支援を受けたものである。

参考文献

[1] Nicholas Asher and Alex Lascarides. **Logics of Conversation**. Cambridge University Press, 2003.

[2] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In **Proceedings of the 2013 Conference of Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1515–1520, 2013.

[3] Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. Text-level discourse dependency parsing. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 25–35, 2014.

[4] 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.

[5] Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. GSN: A graph-structured network for multi-party dialogues. In **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)**, pp. 5010–5016, 2019.

[6] Zhouxing Shi and Minlie Huang. A deep sequential model for discourse parsing on multi-party dialogues. In **Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)**, pp. 7007–7014, 2019.

[7] Annie Louis, Aravind Joshi, and Ani Nenkova. Discourse indicators for content selection in summarization. In **Proceedings of the SIGDIAL 2010 Conference**, pp. 147–156, 2010.

[8] 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**, pp. 1834–1839, 2014.

[9] 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**, pp. 2212–2218, 2015.

[10] Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. Learning-based single-document summarization with compression and anaphoricity constraints. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016)**, pp. 1998–2008, 2016.

[11] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Discourse-aware neural extractive text summarization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 5021–5031, 2020.

[12] Yangfeng Ji and Noah A. Smith. Neural discourse structure for text categorization. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 996–1005, 2017.

[13] Elisa Ferracane, Su Wang, and Raymond J. Mooney. Leveraging discourse information effectively for authorship attribution. In **Proceedings of the The 8th International Joint Conference on Natural Language Processing (IJCNLP)**, pp. 584–593, 2017.

[14] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. Evaluating discourse-based answer extraction for why-question answering. In **Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)**, pp. 735–736, 2007.

[15] 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 (ACL)**, pp. 977–986, 2014.

[16] Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, pp. 1171–1182, 2017.

[17] 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.

[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 (ACL)**, pp. 444–449, 2018.

[19] Noriki Nishida and Yuji Matsumoto. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. **Transactions of the Association for Computational Linguistics**. to appear.

[20] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus -A semantically annotated corpus for bio-textmining. **Bioinformatics**, Vol. 19, No. suppl1, pp. 180–182, 2003.

[21] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation**, 2008.

[22] Harry Bunt and Rashmi Prasad. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)**, 2016.