テキスト生成による議論マイニング

川原田 将之 † 平尾 努 § 内田 渉 † 永田 昌明 §
† 株式会社 NTT ドコモ §NTT コミュニケーション科学基礎研究所
{masayuki.kawarada.vw, uchidaw}@nttdocomo.com
{tsutomu.hirao, masaaki.nagata}@ntt.com

概要

本研究では，テキスト生成により議論マイニングを行う手法を提案する．この手法では，議論構造を表すアノテーション付きテキストを生成し，生成したテキストから議論構造を抽出することで議論マイニングを行う．エンコーダ・デコーダモデルを採用し，議論マイニングのモデル構造をシンプルにすることで，従来手法の課題であったハイパーパラメータの調整や複雑な後処理を大幅に簡易化することができる．実験の結果，ベンチマークデータセットに対して，世界最高性能を達成したことを報告する．

1 はじめに

議論マイニングとは，文書内に存在する議論構造を明らかにするタスクである．学生のエッセイ [1, 2] や医療分野 [3] など，多岐にわたる領域で研究が進められており，文書要約 [4] やエッセイの自動採点 [5] への応用が期待されている。

図 1 に示すように，議論構造は木構造や有向グラフによって表現される．この構造を得るためには，(1) テキストから議論に関連するテキストスパンを特定し，(2) それらの種類を判別し，(3) 抽出したスパン同士の関係性を明らかにする，という 3 つのサブタスクが必要である．

従来のアプローチでは, これらのサブタスクごとに異なるモデルを学習し, 逐次的に処理を行うことが一般的であったが，近年ではニューラルネットワークを活用した End-to-End のアプローチが多く見られるようになっている [6, 7]．一方で，End-to-End モデルでは，複数のサブタスクを 1 つのモデルに統合するため，モデル構造が複雑となる傾向がある．これにより，ハイパーパラメータの調整が難しくなり，複雑な後処理が必要となるという課題が存在する．

これらの課題に対処するため，本研究では，Translation between Augmented Natural Languages (TANL) [8] を議論マイニングに適用することを提案する．TANL は，関係抽出や固有表現抽出などの様々な情報抽出タスクを文生成を通じて行う手法であり，入力文にアノテーションを付与した文を生成し，生成文から後処理によって目的の関係や情報を抽出する．TANL は，文を入力とするタスクで高い性能を示しているが，文書全体を入力とし，より広範な関係性の抽出が求められる議論マイニングのようなタスクにおいても同様の効果が得られるかは，まだ明らかではない．

提案手法の有効性を検証するため, 議論マイニングのベンチマークデータセットである Argument-annotated Essays Corpus (AAEC) [2]，AbstRCT [3]，Cornell eRulemaking Corpus (CDCP) [9] を用いた実験を行った．複数のパラメータサイズの T5 [10]，FLAN T5 [11] で実験を行い，FLAN T5-XXL を用いることで，これら全てのデータセットに対して世界最高性能を達成したことを報告する．さらに，生成モデルを使用する際の推論時間の削減にも取り組んだ．議論構造に関係のないテキストスパンの出力を省略することで，AbstRCT において性能を損なうことなく推論時間を約 30 %削減できることを確認した．

2 関連研究

議論マイニングには，議論に関連するテキストスパンを特定する **Span Identification**，抽出したスパンに要素ラベルの付与を行う **Component Classification**，関係ラベルの付与を行う **Relation Classification** という 3 つのサブタスクが含まれる[1)]．初期の議論マイニング研究において，これらのサブタスクは，別々のモデルで解かれることが多かったが [12, 13]，近年では End-to-End のアプローチで行われることが主流となっている [7, 14, 6, 13].

1) 議論マイニングにおいて，主張や理由などの議論に関連するテキストスパンのことを要素 (Component) という．

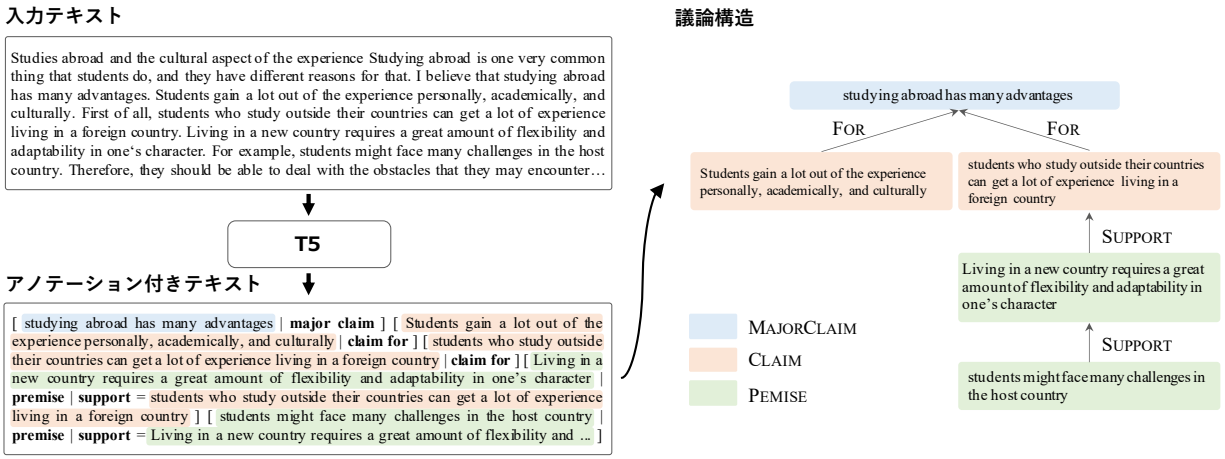

図 1　テキスト生成による議論マイニング手法の全体図.

Morio ら [7] は，Longformer [15] に Biaffine Parser を組み合わせたモデルを用いることで，複数のデータセットに対して世界最高性能を達成している．しかし，このモデルは構造が複雑であるため，ハイパーパラメータの設定が難しく，各サブタスクの出力から最小有向全域木を求めるための後処理が必要になる．これに対し，Bao ら [14] は，BART [16] と Constraint Pointer-mechanism (CPM) を組み合わせたモデルを採用している．Bao らの研究は，エンコーダ・デコーダを用いることで，議論マイニングのプロセス自体をシンプルにしたという点で我々と近いが，我々の手法はテキスト生成に基づいている．テキスト生成によるアプローチを採用することで，事前学習済みモデルをそのまま用いることができ，事前学習で得られた知識を最大限に活用できる．

3　提案手法

図 1 に提案手法の全体図を示す．議論構造を表すアノテーション付きテキストを定義し，テキストを入力，アノテーション付きテキストを出力として，教師あり学習を行う．推論時には，モデルから出力されたアノテーション付きテキストと入力テキストのアライメントを取ることで，議論構造の抽出を行う．

3.1　タスク定義

n 個の単語から成る入力テキスト x は，$x = [x_1, \ldots, x_n]$ と表される．Span Identification の目的は，議論に関連するテキストスパン $s = [x_{\text{start}}, \ldots, x_{\text{end}}]$ を抽出することである．ここで，start，end は，それぞれスパンの開始位置と終了位置を表し，抽出したスパンは $(\text{start}, \text{end})$ のように表される．Component Classification は，得られたスパンに対して，要素ラベル $c \in C$ の付与を行う．ここで，C はデータセットに含まれる全ての種類の要素ラベルを表している．要素ラベルを付与したテキストスパンは，$(\text{start}, \text{end}, c)$ と表される．最後に，Relation Classification により，抽出したスパン同士の関係性の抽出を行う．$(\text{start}^{\text{src}}, \text{end}^{\text{src}})$ が $(\text{start}^{\text{tgt}}, \text{end}^{\text{tgt}})$ に $r \in R$ という関係で依存関係にあるとき，$(\text{start}^{\text{src}}, \text{end}^{\text{src}}, \text{start}^{\text{tgt}}, \text{end}^{\text{tgt}}, \text{r})$ と表される．ここで，R はデータセットに含まれる全ての種類の関係ラベルを表す．通常，議論マイニングモデルの評価は，各サブタスクにおける予測データと正解データの一致率を計測することによって行われる．

3.2　アノテーション付きテキスト

TANL [8] と同様に，入力テキストに対して議論構造を含んだアノテーション付きテキストを定義する．要素ラベルが c であるスパン s^{src} が別のスパン s^{tgt} に関係ラベル r で依存しているとき，アノテーション付きテキストは，“[s^{src} | c | r = s^{tgt}]” と表現される．また，s^{src} が他のスパンと依存関係を持たない場合，関係ラベルと依存先のスパンを省略して，“[s^{src} | c]” と表される．

3.3　アノテーション付きテキストの省略

TANL では，入力テキストをすべて含んだ上でアノテーション付きテキストを出力していた．議論マイニングは，文書を入力としたタスクであるため，出力長が長くなり，それに伴い推論時間の長さが課題となる．そこで，議論構造に関係しないテキストの出力を省略することで効率的な推論を行う．表 1 に TANL のアノテーション付きテキスト（省略な

入力テキスト	Advantages and disadvantages of the prevalent of English With the development of globalization , English became the dominated language in national trade , conference and many important events . This phenomenon has aroused a heated discussion in public . Some people claim that the prevalent of English brings a great number of benefits for people .
省略なし	Advantages and disadvantages of the prevalent of English With the development of globalization , English became the dominated language in national trade , conference and many important events . This phenomenon has aroused a heated discussion in public . Some people claim that [the prevalent of English brings a great number of benefits for people | claim for] .
省略あり	[the prevalent of English brings a great number of benefits for people | claim for]

表 1 TANL のアノテーション付きテキスト (省略なし) と議論構造に無関係なテキストの出力を省略したアノテーション付きテキスト（省略あり）の比較.

		Essay		Paragraph	
	Params	**C**	**R**	**C**	**R**
ILP [2]	-	-	-	62.61	34.74
BLCC [13]	-	63.23	34.82	66.69	39.83
LSTM-ER [13]	-	66.21	29.56	70.83	45.52
BiPAM-syn [6]	110M	-	-	73.5	46.4
BART-CPM [14]	139M	-	-	75.94	50.08
ST Model [7]	149M	76.55	54.66	76.48	59.55
T5-Base	220M	73.75	49.69	74.85	57.16
T5-Large	770M	75.65	51.17	75.55	57.47
T5-3B	3B	77.95	55.95	77.43	59.53
T5-11B	11B	79.48	57.06	77.17	59.02
FLAN T5-Base	220M	75.17	51.99	75.55	58.51
FLAN T5-Large	770M	77.75	56.06	76.93	58.57
FLAN T5-XL	3B	78.51	56.80	77.89	60.94
FLAN T5-XXL	11B	**80.15**	**61.19**	**78.40**	**61.87**

表 2 AAEC の実験結果．C, R はそれぞれ, Component-F1, Relation-F1 を表す.

	Component-F1	**Relation-F1**
ST Model [7]	64.16	38.38
FLAN T5-Base	68.76	38.31
FLAN T5-Large	71.11	44.47
FLAN T5-XL	71.27	45.80
FLAN T5-XXL	**72.86**	**47.66**

表 3 AbstRCT の実験結果.

	Component-F1	**Relation-F1**
BART-CPM [14]	57.72	16.57
ST Model [7]	68.90	31.94
FLAN T5-Base	66.80	23.19
FLAN T5-Large	68.94	28.42
FLAN T5-XL	72.12	31.01
FLAN T5-XXL	**72.68**	**33.96**

表 4 CDCP の実験結果.

し）と議論構造に無関係なテキストを省略したアノテーション付きテキスト（省略あり）の比較を示す.

4 実験

4.1 データセット

議論マイニングのベンチマークデータセットである Argument-annotated Essay Corpus (AAEC) [2] , AbstRCT [3], Cornell eRulemaking Corpus (CDCP) [9] の 3 種類のデータセットを用いた．AAEC は，学生が書いたエッセイに対してアノテーションがされたデータであり，エッセイレベルとパラグラフレベルの 2 種類のデータが存在する．要素ラベルと関係ラベルは，それぞれ，C = {MajorClaim, Claim, Premise}，R = {For, Against, Support, Attack} が存在する．AbstRCT は，医療ドメインのデータセットであり，要素ラベルと関係ラベルは，それぞれ，C = {MajorClaim, Claim, Evidence}，R = {Support, Attack, Partial-Attack} が含まれる．AbstRCT は，他の 2 つのデータセットと比較して，議論に無関係なテキストが最も多く含まれており，その単語が全体に占める割合は 49.30%である．CDCP は，市民からのコメントに対してアノテーションされたコーパスであり，要素ラベルと関係ラベルは，それぞれ，C = {Fact, Testimony, Value, Policy, Reference}，R = {Reasons, Evidence} が含まれる．

4.2 QLoRA による Fine-tuning

TANL では，T5-Base に含まれる全てのパラメータに対して，Fine-tuning を行っていた．本研究では，モデルのパラメータ数を増加させた際の性能についても調査を行うため，QLoRA [17] による Fine-tuning を行う．QLoRA は，モデルに含まれるパラメータとは別に低ランク行列を用意し，量子化を行うことにより学習時の GPU メモリを削減する方法である.

4.3 実験設定

エンコーダ・デコーダとして，T5 [10] と FLAN T5 [11] を用いた上で，Base (220M), Large (770M), XL (3B) and XXL (11B) の 4 種類のパラメータサイズで学習を行った．それぞれのモデルに対して，異なる seed で 3 回の学習を行い，それらの平均値を報告する．詳細な実験設定については，付録 A に示す.

4.4 比較手法

ILP [2], BLCC [13], LSTM-ER [13], BiPAM-syn [18], BART-CPM [14], Single Task (ST) model [7] の 6 種類

出力形式	モデル	AAEC (Essay)		AAEC (Paragraph)		AbstRCT	
		Component-F1	Relation-F1	Component-F1	Relation-F1	Component-F1	Relation-F1
省略なし	FLAN-T5 Base	74.99	50.87	74.97	57.54	65.30	34.55
	FLAN T5-Large	77.76	55.62	76.53	59.09	69.47	39.66
	FLAN T5-XL	78.73	57.21	77.17	61.03	73.13	42.39
	FLAN T5-XXL	**80.59**	60.37	**79.06**	**62.38**	72.78	47.11
省略あり	FLAN-T5 Base	75.17	51.99	75.55	58.51	68.76	38.31
	FLAN T5-Large	77.75	56.06	76.93	58.57	71.11	41.49
	FLAN T5-XL	78.51	56.80	77.89	60.94	71.27	45.80
	FLAN T5-XXL	80.15	**61.19**	78.40	61.87	**72.86**	**47.66**

表 5　議論構造に無関係なテキストを出力する場合（省略なし）と出力しない場合（省略あり）の Component-F1 と Relation-F1 の比較.

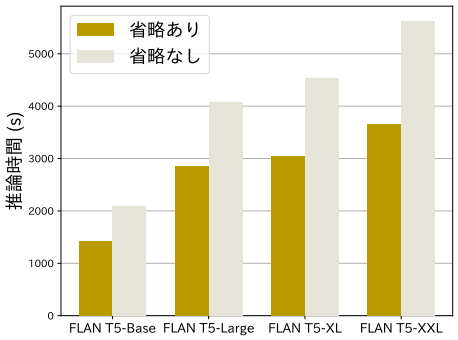

図 2　AbstRCT のテストデータ全てを推論した際の推論時間の比較．推論時の Batch Size は 2 とした.

の手法に対して，比較実験を行った．この中で，BiPAM-syn と Single Task (ST) model は事前学習済みのエンコーダを採用した手法であり，BART-CPM は事前学習済みのエンコーダ・デコーダを採用した手法である.

4.5　評価方法

先行研究 [2, 13, 6, 7, 14] と同様に，議論マイニングにおける一般的な評価方法である Component Classification の F1 スコア（Component-F1）と Relation Classification の F1 スコア (Relation-F1) を用いて評価を行う.

4.6　実験結果

表 2 に AAEC の実験結果を示す．実験結果から，パラメータ数を大きくすることでエッセイレベル (Essay) とパラグラフレベル (Paragraph) の両方で性能向上することがわかる．これらの結果から，大規模モデルへの QLoRA を用いた Fine-tuning は有効であることがわかる．FLAN T5-XXL の場合，エッセイレベルでは Component-F1/Relation-F1 がそれぞれ 80.15/61.19，パラグラフレベルでは 78.40/61.87 という結果が得られており，これは両方のタスクにおいて世界最高性能を達成している．また，すべてのパラメータサイズにおいて，FLAN T5 が T5 を上回っていることから，FLAN [19] による Instruction-tuning は，議論マイニングに効果的であると言える.

次に，表 3 と表 4 に AbstRCT と CDCP における実験結果をそれぞれ示す．AAEC と同様に，パラメータサイズが大きくなると性能向上が見られ，FLAN T5-XXL では，両方のデータセットに対して世界最高性能が得られた．これらの結果から，提案手法は特定のデータセットに限らずに効果的であることがわかる.

最後に，議論構造に無関係なテキストの出力を省略した場合の結果について考察する．表 5 に出力した場合と省略した場合の結果を示す．AAEC のエッセイレベル，AAEC のパラグラフレベル，AbstRCT のすべてのデータセットに対して，省略したとしても性能悪化が起こらないことが分かる．また，AbstRCT において，省略した場合としない場合の推論時間の比較を図 2 に示す．全てのモデルサイズで推論時間を 30%程度削減できており，議論構造に無関係なテキストの出力を省略することは効果的であることがわかる.

5　おわりに

本研究では，シンプルかつ高性能なテキスト生成に基づく議論マイニング手法を提案した．TANL を議論マイニングに適用する際、議論構造に無関係なテキストの出力を削減するための出力形式についても調査を行った．議論マイニングのベンチマークデータセットである AAEC，AbstRCT，CDCP を用いた実験の結果，提案手法は既存手法を上回り，全てのデータセットにおいて世界最高性能を達成した．さらに，不要なテキストの出力を省く出力形式は，モデル性能を損なうことなく，推論時間を約 30%削減できることを明らかにした．今後は，大規模言語モデルにおける提案手法の有効性を確かめたい.

参考文献

[1] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**, pp. 1501–1510, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[2] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. **Computational Linguistics**, Vol. 43, No. 3, pp. 619–659, September 2017.

[3] Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based Argument Mining for Healthcare Applications. In **ECAI 2020 - 24th European Conference on Artificial Intelligence**, Santiago de Compostela / Online, Spain, August 2020.

[4] Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 6187–6194, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[5] Huy V. Nguyen and Diane J. Litman. Argument mining for improving the automated scoring of persuasive essays. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[6] Yuxiao Ye and Simone Teufel. End-to-end argument mining as biaffine dependency parsing. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 669–678, Online, April 2021. Association for Computational Linguistics.

[7] Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. End-to-end argument mining with cross-corpora multi-task learning. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 639–658, 2022.

[8] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In **9th International Conference on Learning Representations, ICLR 2021**, 2021.

[9] Joonsuk Park and Claire Cardie. A corpus of eRulemaking user comments for measuring evaluability of arguments. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[12] Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1384–1394, San Diego, California, June 2016. Association for Computational Linguistics.

[13] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11–22, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[14] Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10437–10449, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. **arXiv:2004.05150**, 2020.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. **arXiv preprint arXiv:2305.14314**, 2023.

[18] Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. Document-level entity-based extraction as template generation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5257–5269, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[19] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **International Conference on Learning Representations**, 2022.

パラメータ	AAEC (Paragraph)	AAEC (Essay)
Batch size	32	8
Max length	512	1024
Step	10,000	
Dropout	0.1	
Adam beta1	0.9	
Adam beta2	0.998	

表 6 AAEC データセットで Fine-tuning を行う際のハイパーパラメータ.

パラメータ	AbstRCT	CDCP
Batch size	8	8
Max length	1024	1024
Step	10,000	
Dropout	0.1	
Adam beta1	0.9	
Adam beta2	0.998	

表 7 AbstRCT と CDCP データセットで Fine-tuning を行う際のハイパーパラメータ.

A 実験設定

A.1 データセット

AAEC では，Egar ら [13] に従い，学習データ/検証データ/テストデータに分割を行った．それぞれのデータセット数は，エッセイレベルのタスクでは 286/36/80 データであり，パラグラフレベルのタスクでは， 1587/199/449 データである．要素ラベルと関係ラベルは，それぞれ，C = {MajorClaim, Claim, Premise}，R = {For, Against, Support, Attack} が存在しているが，AAEC のアノテーションルールで，Claim は常に MajorClaim に対して依存関係にある．そのため，本研究では，要素ラベルと関係ラベルを，C = {MajorClaim, ClaimFor, ClaimAgainst, Premise}，R = {Support, Attack} と設定し学習を行った．Component-F1 の評価では，ClaimFor, ClaimAgainst を Claim として算出を行い，Relation-F1 の評価では，ClaimFor, ClaimAgainst を For, Against として評価を行う．

AbstRCT では，Park ら [9] と同様の分割方法で学習データと検証データの分割を行った．テストデータとして，neoplasm データを用いたため，学習データ/検証データ/テストデータの分割は，を 300/50/100 となった．

CDCP においても，先行研究に従い，全 731 データから抽出した 150 データをテストデータとした．また，全学習データの 15%を検証データとして実験を行った．

パラメータ	
r	16
lora alpha	32
lora dropout	0.05
bias	none
task type	SEQ_2_SEQ_LM
target modules	q, v, k, o, wo, wi_0, wi_1
load_in_4bit	True
bnb_4bit_quant_type	nf4
bnb_4bit_use_double_quant	True
bnb_4bit_compute_dtype	torch.bfloat16

表 8 QLoRA による Fine-tuning のハイパーパラメータ.

A.2 学習パラメータ

表 6 と表 7 に AAEC，AbstRCT，CDCP の学習の際に用いたハイパーパラメータを示す．全てのデータセットで，10,000 STEP の学習を行い，200 STEP 毎にチェックポイントを作成した．その後，Bao ら [14] に従い，検証データを用いて，Component-F1 と Relation-F1 の平均値が最大になるチェックポイントを選び，評価を行った．全てのモデルの学習において，A100 (80GB) を 1 枚用いて行った．

最後に，QLoRA のハイパーパラメータを表 8 に示す．QLoRA による Fine-tuning では，T5 および FLAN T5 に含まれる全ての線形層を対象にして学習を行った．