サーベイ論文自動生成に向けた大規模ベンチマークデータセットの構築

笠西 哲 [1]　磯沼 大 [1]　森 純一郎 [1,2]　坂田 一郎 [1]
[1] 東京大学大学院　[2] 理研 AIP
kasanishi@ipr-ctr.t.u-tokyo.ac.jp　isonuma@ipr-ctr.t.u-tokyo.ac.jp
mori@mi.u-tokyo.ac.jp　isakata@ipr-ctr.t.u-tokyo.ac.jp

概要

サーベイ論文の自動生成は、自動文章要約において最も挑戦的なタスクの一つである。近年、大規模言語モデルによりサーベイ論文生成が挑まれているものの、大規模データセットの欠如がその進歩の足枷となっている。本研究では、1 万本超のサーベイ論文と 69 万本超の被引用論文で構成されたサーベイ論文生成データセットを公開する。本データセットをもとに、近年の Transformer 要約モデルをサーベイ論文生成用に改良し、サーベイ論文生成の評価実験を行なった。人手評価により、モデルにより生成された要約の一部は人が作成したサーベイ論文と遜色ないことが示された一方で、自動サーベイ論文生成の課題が明らかになった。

1　はじめに

科学論文ドメインの文書要約において、サーベイ論文の自動生成は最も重要な課題の一つである。サーベイ論文は、研究者によって過去の研究成果を調査するために執筆される、複数の論文の要約である [1]。サーベイ論文の自動生成が実現すれば、研究者がこれまでサーベイ論文がなかった分野へ進出する際の大きな一助となる。しかし、これまでにサーベイ論文生成に取り組んだ研究はごく少数である。例えば、Taylor ら [2] は科学論文で学習を行った大規模言語モデルである GALACTICA を用いて、サーベイ論文を自動生成するデモンストレーションを公開したが、内容の正確性について批判が寄せられ、数日中に公開が中止された [3]。サーベイ論文生成タスクに利用可能な大規模データセットの不在が、教師あり要約モデルの同タスクへの適用を困難にし、サーベイ論文自動生成の障壁となっている。

本研究は、10,269 本の論文からなるサーベイ論文自動生成タスクのための大規模データセットを公開し、サーベイ論文生成の研究を開拓する。図 1 にタスクの概要を示す。本研究では、サーベイ論文の章分けと各章で引用される論文が与えられた時に、章ごとにそれらの論文を要約するタスクに取り組む。実際のサーベイ論文執筆では、引用する論文を選定し、章構成を決定するというプロセスが必要だが、本研究はそれらのプロセスはスコープ外とし、章ごとに複数の論文を要約するプロセスに焦点を当ててサーベイ論文生成タスクの実験を行った。さらに、既存の Transformer ベースの要約モデルを拡張することで、タイトルや章タイトルを要約のクエリとした複数文書要約としてサーベイ論文を生成するモデルを提案する。提案手法の生成文章に対して詳細な人手評価を行い、サーベイ論文の完全な自動生成に向けた今後の課題について議論する。

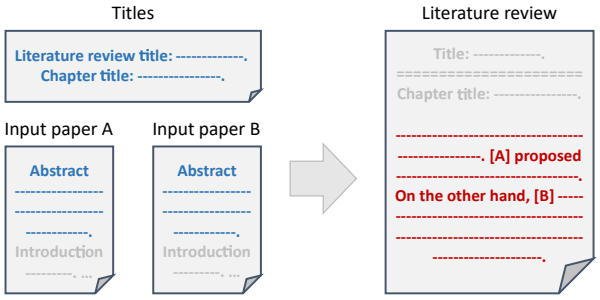

図 1　サーベイ論文自動生成タスクの概要

2　タスクとデータセット

2.1　サーベイ論文生成のタスク定義

出力　本研究ではサーベイ論文を章ごとに生成するタスクを考える。これは、現在の要約モデルの殆どが数百単語以下の短い要約文を生成するタスク [4][7] に適合しており、一般に数千単語以上の長さとなるサーベイ論文全文を現在の要約モデルで生成

表 1 Multi-News[4], Multi-XScience[5], MSˆ2[6] との比較。train/valid/test における参照要約の数、入力文章の合計の長さ（単語数）、参照要約の長さ（単語数）、入力文書の数、novel n-gram の割合を示す。

dataset	train/valid/test	total input length	target length	number of inputs	unigrams	bigrams	trigrams	4-grams
Multi-News	44,972/5,622/5,622	2,103	264	2.79	16.87%	55.57%	74.44%	81.23%
MSˆ2	14,188/2,021/1,667	6930	61	22.80	15.24%	62.35%	87.23%	95.27%
Multi-XScience	30,369/5,066/5,093	778	116	4.42	35.28%	81.57%	94.88%	97.89%
Our dataset	79,103/8,217/7,475	1,789	605	7.01	31.28%	80.11%	95.17%	98.09%

するのは困難なためである。

入力　本タスクでは、**被引用論文のアブストラクト**および**サーベイ論文のタイトル・章タイトル**の情報を入力として用いる。ここで、被引用論文とはサーベイ論文の各章で引用されている論文を指す。被引用論文のアブストラクトは、被引用論文の内容を表す主要な情報源として利用される。本来であれば被引用論文の本文を入力することが望ましいが、構築するデータセットにおいて本文情報が得られない被引用論文が約 30 %存在することを鑑みて、アブストラクトのみを入力とする問題設定とした。また、タイトル情報は、生成される各章でどのような内容が記述されるのかを示唆する、いわば要約の方向性を示すクエリとしての機能を持つ。

2.2　データセットの構築

我々は、科学論文の大規模コーパスである S2ORC[8] を用いて、10,269 本のサーベイ論文からなるサーベイ論文生成タスクのデータセットを構築した。構築方法の詳細は Appendix A に記載した。

2.3　データセットの統計

構築したデータセットと代表的な複数文書要約のデータセットを比較した統計を表 1 に示す。ここで、表中の novel n-gram とは、参照要約中の n-gram のうち入力文書中に含まれないものの割合を示し、novel n-gram の割合が高いデータセットほど要約の抽象度が高い [7]。我々のデータセットは他のデータセットと比較して約 2 倍以上のデータ数を有し、データドリブンなニューラル要約モデルにより適したデータセットとなっている。さらに、我々のデータセットは Multi-News や MSˆ2 と比較して novel n-gram の割合が高く、より抽象的要約手法に適した難易度の高いタスクであることが示唆される。

3　実験

我々が提案したサーベイ論文データセットを用いて、文書要約モデルによるサーベイ論文生成の実験を行う。本章では実験に使用するモデルについて述べ、実装詳細については Appendix B に記載した。

Cited paper 1
Literature review title <s> Chapter title <s> Abstract of paper 1 <s> BIB001
Cited paper 2
Literature review title <s> Chapter title <s> Abstract of paper 2 <s> BIB002

図 2　文書要約モデルへの入力データのフォーマット

3.1　ベースライン手法

ベースライン手法として、教師なし抽出型要約手法である LEAD と LexRank[9]、Transformer ベースの抽象型要約手法である Big Bird[10] と Fusion-in-Decoder[11] を使用する。Big Bird は、通常の Transformer の Self-attention の計算を簡略化し、約 16,000 単語までの長文の入力に対応した Transformer モデルである。今回は、単文書要約タスクにより arXiv データセット [12] で事前学習したモデルを、サーベイ論文データセットで fine-tuning して実験を行う。Fusion-in-Decoder (FiD) は、元々は Open domain question answering 用の複数文書を入力できる Transformer Encoder-Decoder モデルであるが、多文書要約タスクにも適用されている [6]。FiD は複数の文書を個別にエンコードして、その hidden_states を連結してデコーダーに入力し、一括で出力することで文書間の関係性も考慮しながら複数文書を一度に処理できる。今回は、エンコーダとデコーダの初期重みとして CNN/Daily Mail データセットで事前学習を行った BART-Large モデルの重みを使用し、サーベイ論文データセットで fine-tuning して実験を行う。これらの要約モデルに入力するデータのフォーマットを図 2 に示す。今回は、サーベイ論文のタイトルと章タイトル、各被引用論文のアブストラクト、および異なる被引用論文を識別するための識別記号を連結して入力する。Big Bird では全ての被引用論文の情報を連結してモデルに入力する。FiD では各被引用論文の情報を Encoder に別々に入力する。

3.2 提案手法

本節では、前節で紹介した FiD を改良した我々のサーベイ論文生成モデルについて述べる。3.1 節で述べた通り、タイトル・章タイトルはサーベイ論文の各章でどのような内容が記述されるのかを示唆する、いわば要約の方向性を示すクエリとしての機能を持つ。これらのクエリとの類似度によって各被引用論文を重みづけし、よりクエリに沿った要約を出力できるように FiD を改良した。具体的には、クエリと各被引用論文のアブストラクトをそれぞれ Encoder に入力し、得られた hidden_states を average pooling したそれぞれのベクトルの内積をクエリと各被引用論文の類似度とする。その類似度で各被引用論文の hidden_states を重みづけして Decoder に入力することで、モデルにクエリと各被引用論文の類似度の情報を与えている。

n 本の被引用論文の入力トークン列をそれぞれ $R_1, R_2, ..., R_n$ とし、それぞれの長さを $l_1, l_2, ..., l_n$ とする。タイトル・章タイトルを結合したクエリのトークン列を Q とし、その長さを l^q とする。BART Encoder を Enc とすると、m 番目の被引用論文、クエリの hidden_states がそれぞれ以下のように得られる。

$$H_m \in \mathbb{R}^{M \times l_m} = \mathrm{Enc}(R_m) \quad (1)$$

$$H^q \in \mathbb{R}^{M \times l^q} = \mathrm{Enc}(Q) \quad (2)$$

ここで、M は hidden_states の行数であるが、BART の場合は $M = 1024$ である。また、行列の各行の平均をとってベクトルを得る操作を Avgpool とすると、m 番目の被引用論文、クエリの特徴ベクトルをそれぞれ以下のように得られる。

$$\boldsymbol{h}_m \in \mathbb{R}^M = \mathrm{Avgpool}(H_m) \quad (3)$$

$$\boldsymbol{h}^q \in \mathbb{R}^M = \mathrm{Avgpool}(H^q) \quad (4)$$

そして、クエリと m 番目の被引用論文の類似度 w_m をこれらの内積として求め、w_m で重みづけした H_m を結合して BART Decoder に入力することで、出力となるサーベイ論文の章が得られる。

$$w_m = \mathrm{dot}(\boldsymbol{h}_m, \boldsymbol{h}^q) \quad (5)$$

$$\mathrm{output} = \mathrm{Dec}(\mathrm{concat}(w_1 * H_1; ...; w_n * H_n)) \quad (6)$$

ここで、dot は内積、concat は行方向に行列を結合する操作、$*$ は行列の各要素にスカラー値をかける操作、Dec は BART Decoder を表す。

表 2 ベースライン手法と提案手法の ROUGE スコアによる自動評価の結果

Models	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	22.58	4.48	11.54
LexRank	24.14	4.90	12.42
Big Bird	23.79	3.71	15.47
FiD	30.72	7.09	16.05
Ours	**34.93**	**7.86**	**16.76**

4 結果

本章では、ベースライン手法と提案手法を用いたサーベイ論文生成実験の結果について述べる。

4.1 ROUGE スコアによる自動評価

ベースライン手法と提案手法の ROUGE スコアを表 2 に示す。表 2 より、BigBird のスコアが FiD よりも低く、LEAD や LexRank と同程度であることが分かる。長文の入力に対応した単文書要約モデルである BigBird を fine-tuning するのみでは性能が著しく低いことから、サーベイ論文生成が通常の文書要約と比較してより難易度の高いタスクであることが示唆される。また、我々の改良手法とベースラインの FiD を比較すると、改良手法のスコアが高いことがわかる。我々の改良手法がタイトル・章タイトルを要約のクエリとして考慮することで、より適切な文章を生成できていることが確認された。

4.2 人手評価

1 年以上の研究経験を有する画像処理研究従事者 3 人に生成文章の人手評価を依頼し、評価対象として画像処理分野の 5 本のサーベイ論文中の 30 章を選定した [13, 14, 15, 16, 17]。これらの論文は、学習・検証用データセット中のいずれのサーベイ論文とも被引用論文の重複が 20%未満である。評価項目として以下の 5 つを定め、各項目について提案手法によって生成された章と実際のサーベイ論文の章のどちらが優れているか、あるいは同程度であるかの 3 段階で評価するように依頼した。評価者は、提示された 2 つの章のうちどちらが実際のサーベイ論文であるかを伏せられた状態で評価した。

- Relevance: 文章がタイトル・章タイトルと関連しているか
- Coherence: 文章に一貫性があり、情報が構造化されているか

表 3　提案手法の生成文章と実際のサーベイ論文を比較した人手評価の結果

Human Evaluation Results	Relevance	Coherence	Informativeness	Factuality	Overall score
Human > Machine-Generated	25.6%	48.9%	64.4%	40.0%	68.9%
Comparable	56.7%	31.1%	20.0%	48.9%	8.9%
Machine-Generated > Human	17.8%	20.0%	15.6%	11.1%	22.2%

Title: **Deep Learning for Image Super-resolution: A Survey**
Chapter title: **Progressive Upsampling Super-resolution**

The Progressive Upsampling Super-Resolution (PSR) BIB002, BIB004 is proposed to progressively reconstruct the sub-band residuals of high-resolution images at multiple pyramid levels. This method does not require the bicubic interpolation step as the pre-processing step, and thus dramatically reduces the computational complexity. The key idea of PSR is that the network upsamples an image in intermediate steps, while the learning process is organized from easy to hard, as is done in curriculum learning. To obtain more photorealistic results, a generative adversarial network (GAN), named ProGanSR, is proposed, which follows the same progressive multi-scale design principle. This not only allows to scale well to high upsampling factors (e.g., 8×) but also constitutes a principled multiscale approach that increases the reconstruction quality for all up-sampling factors simultaneously BIB005. …

図 3　Overall score で高い評価が得られた生成章の例

- Informativeness: 一般的な情報だけではなく、引用文献に関する具体的な情報を記述しているか
- Factuality: 入力アブストラクトと矛盾する情報や、事実に反する情報がないか
- Overall score: 全体的なサーベイ論文としての完成度

人手評価の結果を表 3 に示す。数値は、各指標について実際のサーベイ論文の章が優れている（Human > Machine-Generated）、同程度である（Comparable）、生成された章が優れている（Machine-Generated > Human）と評価された割合を示している。Relevance に関しては、生成された章が優れている、もしくは同程度であると評価された割合の合計が約 75%に達しており、提案手法が要約のクエリを適切に考慮でき、おおむね実際のサーベイ論文と遜色なくタイトル・章タイトルと関連した文章を生成できていることが明らかとなった。また、Factuality に関しては、生成された章が優れている、もしくは同程度であると評価された割合の合計が 60%に達しており、おおむね半分以上の生成文章で実際のサーベイ論文と同程度以上に誤りのない記述を生成できていることが明らかとなった。一方、Informativeness に関しては生成された章が優れている、もしくは同程度であると評価された割合の合計が約 36%にとどまっており、引用文献に関する具体的な情報を記述できているかという点では、生成文章は実際のサーベイ論文に及んでいないといえる。

そして、Overall score に関しては、生成された章が優れていると評価された割合が 22.2%に達している。これは、生成された文章のほうが本物のサーベイ論文よりも「サーベイ論文らしい」という例が一定程度存在することを示しており、非常に興味深い結果であるといえる。Overall score で 2 人の評価者に優れていると評価された生成章の例を図 3 に示す。図 3 より、生成文章は “Progressive Upsampling Super-resolution” について一貫した流暢な説明ができていることが分かる。

4.3　今後の方向性

第 4.2 節の人手評価の結果、生成文章が優れていると評価された割合が特に少ないのは Informativeness と Factuality の項目であった。これは、被引用論文の情報としてアブストラクトのみを入力しているため、参照要約が入力文書にない情報を含むことが原因であると考えられる [18]。このため、アブストラクトのみならず本文や本データセットで利用可能な引用文などの追加情報を入力することで、Informativeness と Factuality をともに改善できると考えられる。

本研究では、人手評価において生成文章が全体的に優れていると評価された割合が 20%以上に達するなど、サーベイ論文の自動生成に向けた第一歩として有望な結果を示した。一方で、具体的な情報の欠如や誤った情報の出力など、未解決の重要な課題も明らかになった。現状の性能においては、サーベイ論文の執筆時に草稿を自動生成するツールなど、人間による修正を前提とした応用が期待される。

5　おわりに

本研究では、サーベイ論文自動生成タスクとそのための大規模なデータセットを提案し、さらに同タスクに適した文書要約手法を提案した。人手評価を含む詳細な実験により、同タスクの課題と今後の方向性を明らかにした。本研究が、サーベイ論文生成という挑戦的な課題に取り組む今後の研究の基礎となることを願う。

謝辞

本研究は、NEDO JPNP20006、JST ACT-X JPMJAX1904 及び JST CREST JPMJCR21D1 の支援を受けたものである。

参考文献

[1] Kokil Jaidka, Christopher Khoo, and Jin-Cheon Na. Deconstructing human literature reviews – a framework for multi-document summarization. In **Proceedings of the 14th European Workshop on Natural Language Generation**, pp. 125–135, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. **arXiv preprint arXiv:2211.09085**, 2022.

[3] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica demo. https://galactica.org/, 2022. (Accessed on 12/26/2022).

[4] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8068–8074, Online, November 2020. Association for Computational Linguistics.

[6] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. MSˆ2: Multi-document summarization of medical studies. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7494–7513, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[8] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics.

[9] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. **Journal of artificial intelligence research**, Vol. 22, pp. 457–479, 2004.

[10] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 17283–17297, 2020.

[11] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 874–880, Online, April 2021. Association for Computational Linguistics.

[12] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[13] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. **IEEE transactions on pattern analysis and machine intelligence**, Vol. 43, No. 10, pp. 3365–3387, 2020.

[14] Licheng Jiao and Jin Zhao. A survey on the new generation of deep learning in image processing. **IEEE Access**, Vol. 7, pp. 172231–172263, 2019.

[15] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. **ACM Computing Surveys**, Vol. 51, No. 6, pp. 1–36, 2019.

[16] Hamid Laga. A survey on deep learning architectures for image-based depth reconstruction. **arXiv preprint arXiv:1906.06113**, 2019.

[17] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. **Neural Networks**, Vol. 131, pp. 251–275, 2020.

[18] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, 2022. Just Accepted.

[19] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

A　データセット構築手法の詳細

我々のデータセットは、タイトル、章タイトル、アブストラクト、本文情報、さらには引用情報を含んだ科学論文の大規模コーパス S2ORC[8] を使用して構築された。まず、サーベイ論文の候補として、Computer Science 分野の論文の中から “survey”, “overview”, “literature review”, “a review” のいずれかをタイトルに含み、かつ本文情報にアクセスできる論文を抽出した。これによって、13,984 本の論文がサーベイ論文の候補として得られた。

これらの候補の中にはサーベイ論文として適切ではない論文が含まれているため、候補からサーベイ論文として適切な論文を抽出するための SciBERT[19] ベースのサーベイ論文分類器を学習した。まず、コンピュータサイエンスの研究に従事している大学院生に、候補論文のうち 889 本の論文について、サーベイ論文として適切かどうかを以下の基準でアノテーションするように依頼した。

- 複数の科学論文をレビューしていること
  - 一般的なツールや書籍、記事をレビューしていないこと
  - 特定のプロジェクトや Shared Tasks の解説を行っていないこと
- 文献レビューのみを行っており、新規手法の提案・先行研究の再実験・アンケート等をしていないこと（＝引用文献の情報だけでは原理的に生成不可能な情報が含まれていないこと）

アノテーターは、候補論文のタイトルとアブストラクトを対象として上記の基準でアノテーションを行い、3 人のアノテーターのうち 2 人以上が一致した判断を採用することで最終的なアノテーション結果を得た。アノテーションによって、候補論文 889 本のうち 583 本が適切、306 本が不適切として分類され、$Fleiss'kappa = 0.65$ となった。このアノテーション結果を train:valid:test=589:150:150 に分け、SciBERT を fine-tuning してサーベイ論文分類器を学習した。その結果、$accuracy = 89\%, precision = 88\%, recall = 97\%, f1 = 92\%$ の精度を持つサーベイ論文分類器を学習できた。このサーベイ論文分類器を使用して、候補論文からサーベイ論文として適切な論文を抽出したところ、698,049 本の被引用論文と 210,049 章を含む 10,269 本のサーベイ論文が残った。

最後に、これらのサーベイ論文の合計 210,049 章のうち、S2ORC の引用データを利用して被引用論文のアブストラクトを合計 1 本以下しか抽出できなかった章を除外したところ、664,319 本の被引用論文と 94,795 章を含む 9,607 本のサーベイ論文が残った。これらの操作によって、被引用論文のアブストラクトとサーベイ論文のタイトル・章タイトルを入力とし、サーベイ論文の各章を生成するタスクのデータセットを作成した。

B　実装詳細

モデルは PyTorch[1] と HuggingFace Transformers[2] を用いて実装された。BigBird の初期重みは google/bigbird-pegasus-large-arxiv[3] を、FiD の初期重みは facebook/bart-large-cnn[4] を使用した。文書要約モデルはいずれもサーベイ論文データセットで 10epoch 学習を行い、検証時に ROUGE-2 スコアが最も高くなった epoch のモデルを評価に使用した。なお、今回は実験時間の都合上、検証用データセットの 8,217 章からランダムに 1,000 章サンプリングしたデータを使用して評価を行った。学習の optimizer は AdamW を使用し、$\beta_1 = 0.9$, $\beta_2 = 0.999$, $lr = 5e-05$ とした。モデルの出力は $beam size = 4$ のビームサーチによってデコードされ、maximumgenerationlength = 256 とした。なお、FiD では各被引用論文の入力が最大 1024 単語になるように入力を切り詰めている。

1） https://pytorch.org/
2） https://huggingface.co/
3） https://huggingface.co/google/bigbird-pegasus-large-arxiv
4） https://huggingface.co/facebook/bart-large-cnn