文の類似度と Extractive QA による被引用文特定の一手法

西海　真祥† 　金澤　輝一†† 　上野　史††† 　太田　学†††

† 岡山大学大学院自然科学研究科　〒 700–8530 岡山市北区津島中 3–1–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 岡山大学学術研究院自然科学学域　〒 700–8530 岡山市北区津島中 3–1–1
E-mail: †p3ng4cj3@s.okayama-u.ac.jp, ††tkana@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　学術論文の引用箇所から被引用論文中の適切な被引用箇所を特定することは引用箇所の理解に貢献するが，被引用個所の特定は難しい．EMNLP 2020 の ASD ワークショップの CL-SciSumm Shared Task には引用箇所から被引用文を特定するサブタスクが存在する．本稿では Semantic Textual Similarity と Extractive Question Answering を用いて被引用文を特定する手法を提案する．さらに被引用文を特定した結果について考察する．

キーワード　学術論文，引用箇所，被引用文

1 はじめに

学術論文では論文の立ち位置や根拠を示すために，多数の参考文献が引用される．したがって学術論文を理解する上で，引用箇所の引用内容を把握することは重要である．しかし，日々学術論文が増加していく中で，多くの参考文献を参照し，その内容を理解することは大きな労力を要する．EMNLP 2020 の ASD ワークショップの CL-SciSumm Shared Task [1] では，その負担を軽減するために引用箇所に関連づけて被引用論文を要約するタスクがあり，その過程において引用箇所から被引用箇所を特定するサブタスクが存在する．我々は引用箇所に注目した論文閲覧支援を研究しており，被引用文の特定はその研究においても重要な要素である [2]．本稿では，文の類似度を測る Semantic Textual Similarity[1] (STS) と Extractive Question Answering[2] (EQA) の 2 種類の手法を利用し．引用箇所から被引用箇所を特定する手法を提案する．

本稿の構成を示す．2 節で被引用文の特定に関連する研究を紹介し，3 節で被引用文の特定手法，4 節で被引用文を特定する実験について述べ，5 節でまとめる．

2 関連研究

被引用文の特定を含むタスクである CL-SciSumm Shared Task は，Text Analysis Conference 2014[3] (TAC 2014) において実施された生物医学の被引用論文要約にはじまり，Empirical Methods in Natural Language Processing 2020 (EMNLP 2020) のワークショップで採用された [3]．Chai らは，学術論文により事前学習した SciBERT [4] と，BERT の出力に畳み込み層を経て得られた文の分散表現と文の意味役割ラベルから得られた埋め込み表現を組み合わせる Semantics-Aware BERT (SemBERT) [5] による被引用文の特定手法を提案した [6]．この手法は，被引用文を特定する CL-SciSumm Shared Task のサブタスクで最高スコアを記録した．

Elkiss らは PubMed[4]に登録されている 2,497 論文を用い，引用箇所を含む論文と被引用論文中のアブストラクトと引用文のベクトルから self-cohesion と cross-cohesion を計算することで，文集合における文のばらつき具合を求め，その結果を分析した [7]．引用文は被引用論文のアブストラクトの内容の一部を共有しているが，全てではないことを実験的に示すことで，学術論文の要約においてその論文を引用している引用文が役立つ可能性を示した．

Cohan らは情報検索のベクトル空間モデルを基にした被引用文の特定手法を提案した [8]．引用文からストップワード，数字，引用マーカを除くことに加え，引用文を品詞分解して最大 3 語で構成された名詞句に制限したり，学術ドメインに関する知識を利用することにより被引用文を特定する性能が向上することを示した．生物医学に関わるドメイン知識を利用するため，National Library of Medicine (NLM)[5]が提供する Unified Medical Language System (ULMS) から生物医学用語の同義語などを得た．

3 被引用文の特定手法

被引用文の特定は，引用箇所で引用した内容と最も近い文を被引用論文の中から特定するタスクである．ファインチューニングした STS と EQA のモデルにより被引用文を特定する手法を提案する．被引用文の特定において，STS と EQA の計算では共に，文の分散表現に特化した Sentence-BERT [9] を採用し，Sentence-Transformers [9] を利用する．モデルのファインチューニングには CL-SciSumm Shared Task で提供された表 1 のデータセットを使用する．SciSummNet 2019 はプログラムで自動生成された引用文と被引用文のペアであり，14,987

1：https://huggingface.co/tasks/sentence-similarity
2：https://huggingface.co/tasks/question-answering
3：https://www.nist.gov/tac/2014
4：https://www.ncbi.nlm.nih.gov/pmc/
5：https://www.nlm.nih.gov/

表 1　CL-SciSumm Shared Task データセット

	SciSummNet 2019	CL-SciSumm 2018
作成法	自動作成	手動作成
引用文と被引用文のペア数	14987	753

件ある．CL-SciSumm 2018 は人手で作成された引用文と被引用文のペアであり，753 件ある．

3.1　STS による被引用文の特定

Sentence-BERT の事前学習済みモデルである all-MiniLM-L6-v2 [9] を表 1 の CL-SciSumm 2018 でファインチューニングして引用文と被引用文中の文の類似度を求める．入力文長は最大 256 トークンである．入力するファインチューニングデータは，引用文と被引用文とその 2 文間の類似度 $[-1, 1]$ の組み合わせである．なお，ファインチューニングデータの正例は，引用文とそれにふさわしい被引用文の 2 文とし，負例は，引用文と被引用論文中の正例に含まれない文の 2 文とした．正例の類似度は 1.0，負例の類似度は all-MiniLM-L6-v2 で求めた 2 文間の類似度とした．

ファインチューニングしたモデルに引用文と被引用論文中の文を入力すると，2 文間の類似度が得られる．STS による被引用文の特定では，この類似度が高い文を被引用文の候補とする．

3.2　EQA による被引用文の特定

Sentence-BERT の事前学習済みモデルである msmarco-distilbert-base-tas-b [10] を表 1 の CL-SciSumm 2018 でファインチューニングして，引用文を質問の内容としたクエリの応答として適切な文の度合いを被引用論文中の文から求める．入力文長は最大 512 トークンである．ファインチューニングデータは引用文であるクエリと被引用文，そしてクエリに対する応答としての文の適切さ $[-1, 1]$ の組み合わせである．ファインチューニングデータは STS の場合と同様に正例と負例を用意する．すなわち，正例は，引用文とそれにふさわしい被引用文とその 2 文間の類似度 1.0 である．負例は，引用文と被引用論文の正例に含まれない文の 2 文とその 2 文間の msmarco-distilbert-base-tas-b で求めた類似度である．

ファインチューニングしたモデルに引用文と被引用論文中の文を入力すると，その引用文の応答としての被引用文の適切さが得られる．EQA による被引用文の特定では，この適切さが大きい文を被引用文の候補とする．

3.3　STS と EQA による被引用文の特定

3.1 節で説明した STS と 3.2 節で説明した EQA を両方利用して被引用文を特定する．

まず図 1 に実験に用いた検証データにおける，STS と EQA の各スコアの上位の被引用文を選択した時の F_1 を示している．F_1 は予測した被引用文と正解の被引用文の一致状況を示す指標の 1 つであり，高いほどモデルの精度が良いことを示す．図 1 が示すように EQA の上位 2 件を被引用文とした場合 F_1 が最高となった．そのため STS と EQA による被引用文の特定では，STS と EQA のスコアを利用して 2 件を選択する．

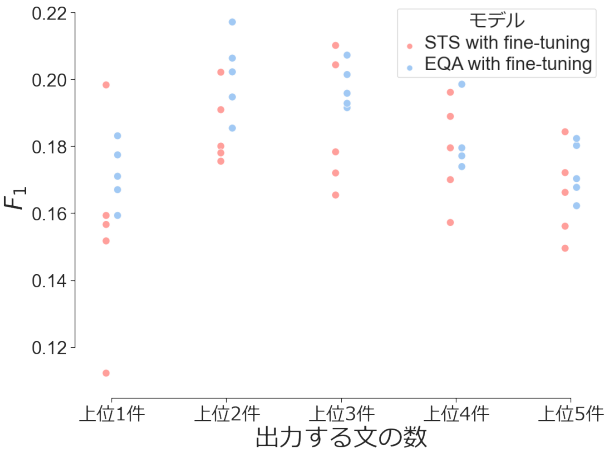

図 1　出力する被引用文の数と F_1

具体的には，STS と EQA の各上位 2 件の文から，出力する 2 文を選択する．候補は最大 4 文あるが，STS と EQA によって選ばれた文は重複する場合がある．STS with fine-tuning の上位 2 件を STS 1st，STS 2nd，EQA with fine-tuning の上位 2 件を EQA 1st，EQA 2nd とした時，下記の通り，文の重複は場合分けできる．

- STS 1st = EQA 1st かつ STS 2nd = EQA 2nd
- STS 1st = EQA 1st かつ STS 2nd ≠ EQA 2nd
- STS 1st ≠ EQA 1st かつ STS 2nd = EQA 2nd
- STS 1st = EQA 2nd かつ STS 2nd = EQA 1st
- STS 1st = EQA 2nd かつ STS 2nd ≠ EQA 1st
- STS 1st ≠ EQA 2nd かつ STS 2nd = EQA 1st
- 重複なし

出力する 2 文の選択は文の重複の有無によって変更する．

- 文の重複がある場合，EQA の上位 2 件を選ぶ．これは図 1 が示すように，単独では EQA が STS より優れているからである．
- STS の上位 2 件と EQA の上位 2 件に文の重複がない場合，STS 1st と EQA 1st を選ぶ．これは 2 つが異なる傾向を示した場合には，それぞれの特徴を取り入れるためである．

4　実　　験

4.1　実 験 概 要

CL-SciSumm Shared Task のテストデータと評価プログラムを利用して提案する被引用文特定手法の性能を測る．このタスクでは 20 の被引用論文がある．各引用箇所の引用内容を示す被引用箇所を含む被引用文を，被引用論文から最大 5 つ選んで提出する．正解の被引用文との一致状況をマイクロ平均 F_1 で評価する．被引用論文と引用文のペア全ての再現率と適合率，そしてそれらの調和平均を求めた F_1 の平均である．

5 分割交差検証により，ファインチューニングのパラメータを決定した．ファインチューニングデータは手動作成した CL-SciSumm 2018 の 753 件のみ使用する．自動作成のデータによるファインチューニングは性能を低下させたため実験で

は使用しない．事前に被引用論文の各文のスコアをファインチューニング前の事前学習済みモデルで算出する．訓練データの 1 件の正例に対し，負例を被引用論文からそのスコアが高い順に 10 件選ぶ．ファインチューニングのパラメータは, バッチサイズを 16, エポック数を 1, warmup step 数を学習 step 数の 10%，学習率を 2e-05, 最適化アルゴリズムを AdamW，weight decay を 0.01 とする．

20 の被引用論文への引用文 339 件と，提案した被引用文特定手法で得られた被引用文のペアで評価する．

4.2 評価指標

正解の被引用文と予測した被引用文から以下の F_1 を算出する．

まず再現率を以下の式で求める．

$$\frac{\text{被引用文と予測した文のうち実際に被引用文であった数}}{\text{正解の被引用文の数}}$$

適合率を以下の式で求める．

$$\frac{\text{被引用文と予測した文のうち実際に被引用文であった数}}{\text{予測した被引用文の数}}$$

F_1 は以下の式で求まる．

$$\frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

4.3 被引用文特定の実験結果

表 2 は被引用文を特定した実験結果の評価指標をまとめたものである．このうち上 3 つは，EMNLP 2020[9]の ASD ワークショップの CL-SciSumm 2020 における CL-SciSumm Shared Task の上位 3 チームの最高スコアとそのモデルを示す．“SciBERT_SemBERT (N=2) [6]” は，意味役割付与情報を利用する Semantics-aware BERT の BERT の部分を，科学論文誌で事前学習済みモデルである SciBERT とする手法であり，F_1 が 0.1716 で最高の値だった．“uniHD intersection 2 field [11]” は，BM25 [12] と top-k re-ranking with BERT [13] の 2 ステップによる文の順位付け手法であり，F_1 が 0.1608 で 2 位のチームだった．“CMU run110 [14]” は，節情報や TextRank [15] を考慮した BERT モデルであり，F_1 が 0.1284 で 3 位のチームだった．一方その下に並ぶ STS without fine-tuning は STS の事前学習済みモデル all-MiniLM-L6-v2，EQA without fine-tuning は EQA の事前学習済みモデル msmarco-distilbert-base-tas-b，STS with fine-tuning と EQA with fine-tuning はそれぞれ all-MiniLM-L6-v2 と msmarco-distilbert-base-tas-b をファインチューニングしたモデルであり，いずれのモデルも被引用文件を 2 件選択する．なお，STS with fine-tuning3. 1 節で，EQA with fine-tuning3. 2 節で説明した提案手法である．2 件を選択する理由は，図 1 に示したように，上位 2 件を選択した EQA with fine-tuning の F_1 が最も高かったからである．表 2 に示した通り，ファインチューニングした EQA はファインチューニングした STS よりも性能が上回っている．

9：https://2020.emnlp.org/

表 2　被引用文特定の実験結果

モデル	適合率	再現率	F_1
SciBERT_SemBERT (N=2) [6]	**0.1318**	0.2459	**0.1716**
uniHD intersection 2 field [11]	0.1164	**0.2597**	0.1608
CMU run110 [14]	0.0869	0.2459	0.1284
STS without fine-tuning	0.0842	0.1588	0.1100
STS with fine-tuning	0.1032	0.1948	0.1349
EQA without fine-tuning	0.0944	0.1782	0.1234
EQA with fine-tuning	0.1223	0.2307	0.1598
Combination	0.1252	0.2362	0.1636

3. 3 節で説明した方法で 2 文を選択する提案手法を表 2 では Combination と記しており，F_1 が 0.1636 となった．これが提案手法の中で最も高い F_1 の値であった．

4.4 考　　察

テストデータにおける STS with fine-tuning と EQS with fine-tuning の選択文の重複状況と正解数を表 3 にまとめる．本研究のように 2 文を選択するならば，STS with fine-tuning と EQA with fine-tuning の出力する組み合わせにおいて，Combination の選択方法が最も F_1 が高くなることが分かる．正解している場合において，STS with fine-tuing と EQA with fine-tuing は共通する部分が多いが，無視できない違いもある．また，STS 1st，STS 2nd，EQA 1st，EQA 2nd から最適な 2 文を選択すると F_1 が 0.1981 まで向上するため，STS with fine-tuing と EQA with fine-tuing の違いを捉えれば，性能向上が図れる．図 2 に，STS with fine-tuning で選択した 2 文と EQA with fine-tuning で選択した 2 文に重複がなく，STS 1st または EQA 2nd が正解の場合の STS 1st と EQA 2nd の STS スコアの差を示す．この結果を仮に利用し，STS よりも EQA の被引用文特定精度が高いことから，EQA 1st を必ず選ぶこととすると，STS 1st と EQA 2nd の STS スコアの差が 0.05 を越えると EQA 1st と STS 1st を選択，下回ると EQA 1st と EQA 2nd を選択することで，F_1 は 0.1694 となる．すなわち，EQA 1st が STS 1st よりも STS のスコアが一定程度高ければ，EQA 2nd ではなく STS 1st を選択することで性能向上を図れる可能性がある．

5 おわりに

本稿では引用文に対する被引用論文中の被引用文の特定手法を提案し，実験では，CL-SciSumm Shared Task のデータを利用して，STS と EQA のタスクに合わせて事前学習した Sentence-BERT をファインチューニングした被引用文判定器を作成した．2 つの被引用文判定器が出力する上位 2 件ずつの文を合わせた最大 4 文から，2 文を選別して被引用文と予測した．この最大 4 文の重複状況によって場合分けし，文の重複があれば EQA の上位 2 件，文の重複がなければ STS と EQA の 1 位を出力とする提案手法では，引用文と被引用文の一致する精度を示すマイクロ平均 F_1 が 0.1636 となった．さらに，STS の上位 2 件と EQA の上位 2 件に重複がない場合に，STS の最

表 3　テストデータにおける STS with fine-tuning と EQA with fine-tuning の選択文の重複と正解数

重複の状況	STS		EQA	
	1st	2nd	1st	2nd
STS 1st = EQA 1st STS 2nd = EQA 2nd	21	4	21	4
STS 1st = EQA 1st STS 2nd ≠ EQA 2nd	28	6	28	20
STS 1st ≠ EQA 1st STS 2nd = EQA 2nd	1	9	6	9
STS 1st = EQA 2nd STS 2nd = EQA 1st	9	10	10	9
STS 1st = EQA 2nd STS 2nd ≠ EQA 1st	7	4	7	7
STS 1st ≠ EQA 2nd STS 2nd = EQA 1st	6	11	11	7
文の重複なし	15	10	17	11
合計	87	54	100	67

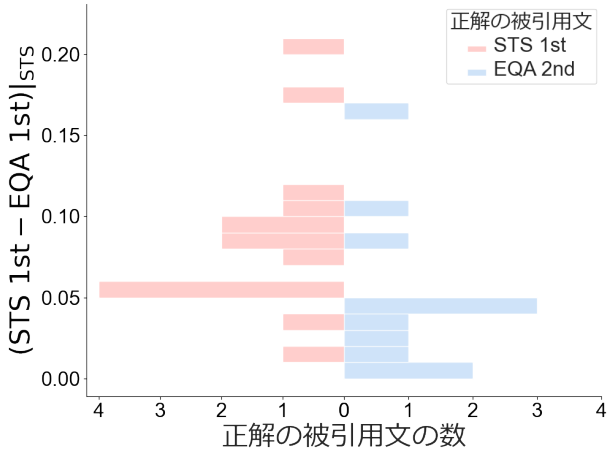

図 2　STS と EQA で出力する 2 文に重複がなく，STS 1st または EQA 2nd が正解の文とその場合の STS 1st と EQA 1st の STS スコアの差

高スコアの文と EQA の最高スコアの文間の，STS スコアの差に閾値を導入して 2 文を選択するようにすれば，マイクロ平均 F_1 は 0.1694 となることも分かった．なお最大 4 文の中から理想的に 2 文を選択できればマイクロ平均 F_1 が 0.1981 まで向上するため，STS と EQA で選択した文の違いについてさらに分析したい．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 (B)(課題番号 22H03904)，同基盤研究 (C)(課題番号 18K11989) および 2022 年度国立情報学研究所共同研究 (22FC01) の援助による．

文　　献

[1] Chandrasekaran, M. K. et al., “Overview and Insights from the Shared Tasks at Scholarly Document Processing 2020: CL-SciSumm, LaySumm and LongSumm,” In Proceedings of the First Workshop on Scholarly Document Processing, Online, Association for Computational Linguistics, pp. 214—224 (online), DOI: 10.18653/v1/2020.sdp-1.24, 2020.

[2] 西海真祥 他, “引用意図を利用した初学者向け学術論文閲覧支援方法の検討,” FIT2022 第 21 回情報科学技術フォーラムの講演論文集第 2 分冊,pp. 39–44, 2022.

[3] Chandrasekaran, M. K. et al., “Overview and Results: CL-SciSumm Shared Task 2019,” In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL 2019), 2019.

[4] Beltagy, I. et al., “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), Hong Kong, China, Association for Computational Linguistics, pp. 3615–3620 (online), DOI: 10.18653/v1/D19-1371, 2019.

[5] Zhang, Z. et al., “Semantics-Aware BERT for Language Understanding,” In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, pp. 9628–9635 (online), DOI: 10.1609/aaai.v34i05.6510, 2020.

[6] Chai, L. et al., “NLP-PINGAN-TECH @CL-SciSumm 2020,” In Proceedings of the First Workshop on Scholarly Document Processing, Online, Association for Computational Linguistics, pp. 235–241 (online), DOI: 10.18653/v1/2020.sdp-1.26, 2020.

[7] Elkiss, A. et al., “Blind men and elephants: What do citation summaries tell us about a research article?,” Journal of the Association for Information Science and Technology, Vol. 59, pp. 51–62, 2008.

[8] Cohan, A., et al., “Matching Citation Text and Cited Spans in Biomedical Literature: a Search-Oriented Approach,” In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Association for Computational Linguistics, pp. 1042–1048 (online), DOI: 10.3115/v1/N15-1110, 2015.

[9] Reimers, N. et al., “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,” Association for Computational Linguistics, (online), available from ⟨https://arxiv.org/abs/1908.10084⟩, 2019.

[10] Hofstätter, S. et al, “Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling,” In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 113–122, DOI: 10.1145/3404835.3462891, 2021.

[11] Aumiller, D. et al., “UniHD@CL-SciSumm 2020: Citation Extraction as Search,” In Proceedings of the First Workshop on Scholarly Document Processing, Online, Association for Computational Linguistics, pp. 261–269 (online), DOI: 10.18653/v1/2020.sdp-1.29, 2020.

[12] Robertson, S. et al., “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, Vol. 3, pp. 333–389, 2009

[13] Qiao, Y. et al., “Understanding the Behaviors of BERT in Ranking,” CoRR, abs/1904.07531, 2019.

[14] Umapathy, A. et al., “CiteQA@CLSciSumm 2020,” Proceedings of the First Workshop on Scholarly Document Processing, Online, Association for Computational Linguistics, pp. 297–302 (online), DOI: 10.18653/v1/2020.sdp-1.34, 2020.

[15] Mihalcea, R. et al., “TextRank: Bringing Order into Text,” In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Association for Computational Linguistics, pp. 404–411, 2014.