学術論文 PDF からの関連研究章と引用情報の抽出による論文執筆支援のためのデータセット構築

小林恵大 [1]　小山康平 [1]　成松宏美 [2]　南泰浩 [1]
[1] 電気通信大学　[2]NTT コミュニケーション科学基礎研究所
{k1810249,k2131071}@edu.cc.uec.ac.jp
hiromi.narimatsu.eg@hco.ntt.co.jp　minami.yasuhiro@is.uec.ac.jp

概要

我々は論文執筆支援を目的とした研究用データセットの拡充を目指し，PDF 形式の論文から本文および引用に関わる情報を抽出する．従来のデータセットは，本文抽出および引用箇所と引用情報の対応付けが容易な Tex 形式の論文をソースとして構築されていた．しかしながら，一般に公開されている論文の多くは PDF 形式であり，Tex 形式の論文を元に構築されたデータセットは量も分野も限られている．本課題に対し，PDF を対象として同様のデータを構築することで，データセット拡充を行う．構築したデータにおいて本文抽出および引用の対応付けを評価した結果，PDF 形式の論文からも Tex 形式をもとにしたときと近い精度となった．これにより，従来より多くの論文を対象としたデータセットの構築ができることを示した．

1　はじめに

科学技術の発展に伴い，学術論文の数が急速に増加している．そのため，研究者が研究方針検討時に既存研究の調査や，論文執筆時に文章の適切な引用を行うことが困難になっている．

上記のような研究者の負担を削減するために，様々な目的でタスクが定義され研究がなされてきた．論文の閲読時間削減を目的とした論文要約 [1, 2, 3] や，論文の効率的な検索を目的とした参考文献推薦 [4]，論文執筆の効率化を目的とした引用要否判定 [5, 6]，被引用文献割り当て [7]，引用文生成 [8, 9] などである．

一方，執筆支援システムの実用化の観点では，論文執筆を行う研究者が直面する複数の課題に対して統合的に支援できることが期待されている．しかしながら，学術論文執筆支援を目的とした各タスクは独立して研究されており，それぞれのタスクの評価に用いられているデータセットも異なっている．それゆえに，統合的な論文執筆支援システムとして評価することができない．

本課題に対し，Narimatsu ら [10] は，同一のデータソースから構築した複数のタスク評価用データセットを構築し，公開している．彼らは，様々なタスクのデータを自動的に作れるよう，Tex ソースが公開されている論文を対象にデータセットを構築している．Tex のソースを対象とすることで，タスクデータを自動的に構築することが可能であるが，Tex のソースが公開されている論文は数量や分野が限られている．たとえば，Tex のソースを公開している代表的なサイトである arXiv は，物理学，計算機科学，数学が主であり[1)]，他の分野の論文は少ない．

そこで，本研究では，Tex のソースが公開されておらず PDF のみしかない論文も対象に，論文執筆支援に向けたデータを自動的に構築することを目的とする．具体的には，Narimatsu ら [10] と同様に，関連研究の章に着目し，統合的な論文執筆支援システムを評価可能なデータセットを PDF から構築する[2)]．

タスクデータ[3)]の構築には，関連研究の章の本文抽出，本文で引用されている引用文献の情報（タイトルや著者情報）の取得および引用文との対応付け，引用文献情報を用いた論文取得が必要である．しかし，Tex のソースではなく PDF から同様のデータセットを作るためには以下の課題がある．

章タイトルの識別． PDF では章タイトルを表す記号は特にないため，フォントの種別や大きさなどの

1）arXiv submission rate statistic https://arxiv.org/help/stats/2020_by_area/index
2）CC BY 4.0 の下でライセンスされている論文 PDF から構築したデータセット，および任意の論文 PDF を入力としてデータセットを構築するためのソースコードを以下に公開予定である．https://github.com/citation-minami-lab
3）タスクデータの詳細は 付録 A を参照のこと．

視覚情報で識別する必要がある．

抽出した本文のクリーニング． PDF では，章の本文として抽出したテキストに，ヘッダ・フッタ・脚注・数式，図表中の文字列が不自然に挿入され，自然言語処理においてノイズとなる．

引用記号と被引用文献の対応付け． Tex では，引用記号（以下「引用アンカ」とする）と被引用文献は，\cite のような引用を表すタグに囲まれた文字列と bib ファイルを照合することで容易に取得できる．しかしながら，PDF の場合には，明確なタグはなく簡単に対応付けができない．

本稿では，上記の課題を解決するために，既存の PDF からの本文抽出と，引用アンカ検出，参考文献抽出において性能が高い複数のツールを組み合わせ，さらにそれらの精度を高める手法を導入する．

2 関連研究

論文の PDF から情報抽出を行う既存手法について，本研究で関連する，本文抽出と参考文献情報抽出および引用アンカ対応付けに絞って説明する．

本文抽出． GROBID [11] は，CRF モデルを用いて，章タイトル，本文の抽出や数式，図表の認識を行うツールとして提案されているがその精度は高くない[4]．PDFBoT [12] は pdf2htmlEX[5] を使用して PDF 論文を HTML に変換することで得られるテキストの特徴量を用いて，ルールベースにより章タイトル，脚注，図表，タイトル，著者，所属，ディスプレイモードの数式を除去した本文のみを抽出する．このツールは章タイトルや参考文献の章を除去するため本研究の目的に直接使用することはできないが，部分的な利用は可能である．

参考文献情報の抽出・引用アンカ対応付け． 参考文献情報の抽出ができるツールとして，ParsCit [13]，CERMINE [14]，GROBID [11] が提案されている．Tkaczyk ら [15] らは，その中でも GROBID が参考文献情報を高い精度で抽出可能なことを示した．これらのツールは引用アンカの検出，引用アンカと引用対象の文献の対応付けも行うが，GROBID ではその精度は高くなく [5]，CERMINE，ParsCit では精度の評価自体がされていない．Ahmad ら [16] は，本文中の引用アンカを，Gosangi ら [6] は，ACL-Anthology Reference Corpus [17] 中の引用アンカを高精度で検出する手法を提案している．

論文執筆支援を目的とした研究においても，PDF から評価用データを構築しているものがある [18, 19]．これらの研究は，CiteSeerX[6] を用いて本文および引用アンカの対応付けを行なっているが，CiteSeerX では本文として引用アンカ周辺のテキストしか取得できない．

上記の通り，本文抽出および参考文献情報抽出の両方で高い精度が得られている手法はない．そのため，本研究では本文抽出に PDFBoT [12]，参考文献情報抽出に GROBID [11]，引用アンカーの検出に Ahmad ら [20] および Gosangi ら [6] の手法と，それぞれ高い精度が得られている手法を用いる．

3 データセット作成の手法

関連研究の章の本文抽出と，本文で引用された文献情報の抽出および対応付けについて詳説する．

3.1 関連研究の章の本文抽出

関連研究の章の本文を抽出するためには，章タイトルに基づき，本文区間を認識する必要がある．そのため，pdf2htmlEX[7] を用いて，PDF を HTML 形式に変換し，HTML のタグとして得られたテキストの座標情報，フォントの種類・サイズ情報を用いて章タイトルを抽出する．その準備として，初めに，本文の抽出に必要のない<a>，<img>，のタグを除去し，本文のフォントサイズや，行の先頭の x 座標，行間隔などの統計情報を取得する．

続いて，章タイトル “Related Work(Study)” とその次の章タイトルを検出し，その二つの章タイトルの間の関連研究の章の本文範囲のテキストを抽出する．章番号 (Section Number) は範囲検出の大きな手がかりとなるため，その有無により以下の通り場合分けをして検出する．

章番号が付与されている場合: "Introduction"は多くの論文が 1 番目の章として設置しているため，その行の先頭の文字列のスタイルを論文の章番号のスタイルとして特定する．そのスタイルを持った文字列と “Related Work(Study)” から始まる行を関連研究の章タイトルとして検出する．そして，関連研究の章タイトルと章番号が連続し，フォントの種類が一致する行を次の章タイトルとする．

章番号が付与されていない場合: 本文のフォントサイズよりも大きく，“Related Work(Study)” から始

4) GROBID Documentation Benchmark https://grobid.readthedocs.io/en/latest/Benchmarking-pmc
5) https://github.com/pdf2htmlEX/pdf2htmlEX
6) https://citeseerx.ist.psu.edu/index
7) https://github.com/pdf2htmlEX/pdf2htmlEX

まる行を関連研究の章タイトルとして取得する．そして，関連研究の章タイトルとフォントのサイズが等しい行を次の章タイトルとして検出する．
上記手法により，“~shown in Related Work section.” のように参照目的で本文中に含まれる章タイトルを誤って検出するケースを減らす[8]．

以上で抽出した本文には，言語処理上問題となるノイズが含まれるため，以下でその除去を行う．

ヘッダ・フッタ・脚注の除去 一般にヘッダ・フッタ・脚注は本文のフォントサイズより小さいため，PDFBoT [12] と同様に，本文のフォントサイズより 1px 以上小さい行を除去する．

キャプションの除去 本文の途中に挟まれた図表のキャプションは，誤って本文のテキストとして抽出されることが多い．そこで，図表のキャプションである，「Table」「Figure」「Fig.」から始まる行を除去する．ただし，本文中の上記文字列を誤って除去しないよう，これらの文字列を含む行とその上の行の行間が本文の行間よりも広いものだけ除去する．また，Ahmed ら [21] と同様に，複数行のキャプションの場合，1，2 行目の行間よりも大きな間隔をキャプションの切れ目と考え，そこまでを除去する．

数式・図の除去 ここでは，文中に現れる λ などの数式ではなく，ディスプレイモードで記述された数式のみを除去する．PDFBoT と同様に，数式や図に含まれる行は本文よりも始点の x 座標が右に寄っていると仮定し，それらを除去する．

表の除去 論文から画像認識モデルにより表の箇所を抽出する手法 [20] を用いて表の除去を行う．具体的には，論文 PDF を pdf2image[9] を使用して画像形式に変換し，モデルへ入力する．そして，モデルが表として検出した範囲にマージン 10px 加えた範囲に存在する行を除去する．

3.2 引用アンカの検出

引用アンカの形式は論文によって様々である．そのため，過去に提案された次の 2 つの正規表現による検出手法 [16, 6] を組み合わせて引用アンカを検出する．前者 [16] は，引用アンカを高い精度で検出可能である．また，前者では検出対象でないものが後者 [6] では対象とされている．さらに，検出対象の拡大のため，後者 [6] の正規表現に一部改良を加えた[10]．

8） 上記手法で使用する正規表現は付録 B を参照のこと
9） https://pypi.org/project/pdf2image/
10） これらの正規表現の詳細は付録 C を参照のこと．

3.3 参考文献の章の解析および引用論文情報の取得

参考文献の章から著者やタイトルの情報を抽出し，それに対応づく引用論文の情報を取得する．章の解析には GROBID [11] を使用する．GROBID は，入力された論文 PDF に対して，参考文献の章を文献毎に分割し，著者，タイトル，年などのタグ付けを行う．ただし，文献が「[1]」のような数字形式の引用アンカを使用している場合，GROBID では文献毎の分割に失敗すると，失敗した文献以降の文献の順番と実際の順番が誤った対応付けになることがある．これを回避するため，引用アンカが数字形式の論文の場合は，HTML から参考文献の章を抽出し，各文献毎に分割した文字列を GROBID に入力することで解析を行った．その後，GROBID により抽出された参考文献のタイトルを外部検索 API へ入力し，一致する文献がヒットした場合に，アブストラクトを取得し引用論文の情報とする．

3.4 引用アンカと引用論文の対応付け

数字形式の引用アンカにはその数字と一致する順番の参考文献の項目を対応づける．数字形式以外の引用アンカでは，引用アンカから年を抽出した後「et al」，「and」，「&」から後ろの文字列を除去して第一著者名を取得する．その情報と解析した参考文献のタグとを照合し，一致した文献を対応づける．

4 評価

提案手法により構築したデータセットの有効性を，関連研究の章テキストの抽出精度および参考文献の解析精度により評価する．

4.1 関連研究の章の抽出の性能

先行研究の Narimatsu ら [10]，及び GROBID [11] と比較し，関連研究の章タイトル検出，本文の抽出・ノイズ除去の精度を評価した．評価指標には，本文抽出，ノイズ除去を同時に評価できる Word Error Rate (WER) と Sentence Error Rate (SER) を用いた．Axcell [22] の論文リストから無作為に選択した論文の中から，関連研究の章を含む 113 件を評価データとした．それらの論文から手作業でノイズを除去した関連研究の章の文章を正解データとした．

4.1.1 関連研究の章タイトルの検出数の評価

初めに，各手法の関連研究の章タイトルの検出の成功・失敗件数を調べた (表 1).

表 1 各手法での 113 件の論文に対する関連研究の章タイトルの検出成功・失敗件数

	成功	失敗
GROBID [11]	103	10
Narimatsu ら [10]	**113**	**0**
提案手法	110	3

Tex のソースでは章タイトルが陽にタグで示されているため，Narimatsu らは全ての関連研究の章タイトルの検出に成功している．また，提案手法の失敗件数は 3 件であり，十分な検出精度を達成できた．

4.1.2 本文抽出・ノイズ除去の精度評価

本文抽出・ノイズ除去精度の評価では，GROBID が抽出した本文に含まれる，独自のタグや文字列を全て削除した．また，Narimatsu らの手法が抽出する本文の引用アンカは Tex のタグ形式であるため，PDF から作成した正解データと差異がある．そこで，各手法において抽出した本文から全ての引用アンカを消去し，条件を揃えた．正解文中の引用アンカはそのままであるため，3 手法とも WER，SER は実際の値より増えるが，同条件での比較ができる．

以上の条件で提案手法・GROBID・Narimatsu らの手法全てで関連研究の章タイトルの検出に成功した評価データのみを使用して評価した (表 2).

表 2 引用アンカ除去を行なった条件での本文抽出，ノイズ除去精度の WER，SER による評価

	WER	SER
GROBID [11]	0.167	0.542
Narimatsu ら [10]	0.188	0.744
提案手法	**0.086**	**0.481**

結果より，提案手法の WER，SER が最も低かった．Narimatsu ら手法の WER，SER が最も高い原因は，Tex 形式を扱うことによるものである．Tex の記号で図表や注釈などは除去していたが，著者らが独自に設定したコマンドでコメントアウトされた文字列が完全には除去しきれていなかった．GROBID の主なエラーは，本文に含まれるべき単語・文の欠落や，図表の検出の失敗であった．これらに対し，提案手法の主なエラーは，HTML 形式に変換したことによるもので，表記を整えるために単語の途中に挿入されたタグが適切に処理できず，単語途中へスペースが挿入されることがあった．

次に，提案手法と GROBID で，引用アンカを消去せずに実際の WER，SER の値を算出した (表 3).

表 3 提案手法と GROBID の本文抽出，ノイズ除去精度の WER，SER による評価

	WER	SER
GROBID [11]	0.087	0.175
提案手法	**0.010**	**0.054**

結果より，提案手法の WER が 0.01 程度，SER も 0.05 程度と共に GROBID よりも低いことから，正確な関連研究章の本文抽出，ノイズ除去が達成できていると言える．

4.2 参考文献のタイトル抽出件数の評価

引用文献情報の抽出および引用アンカとの対応付けについて，Narimatsu らと比較する．評価には，Narimatsu らがデータセットの構築に使用した Axcell の論文リストから，無作為に選択した 2,786 件の論文を使用した．各手法により，これらの論文の本文の引用アンカ検出，引用文献情報抽出，引用アンカ対応付けを行う．そして引用アンカに対応付けられた文献タイトルを，arXiv API にて入力して検索し，一致する文献がヒットした件数で評価した (表 4).

表 4 抽出した参考文献タイトルを，arXiv API に入力して検索し，検索結果の文献が入力と一致した件数

	一致件数
Narimatsu ら [10]	4874
提案手法	4225

結果より，提案手法の一致件数は Narimatsu らの 87%ほどであり，彼らと近い精度を示した．PDF からの引用文献情報抽出および引用アンカ対応付けでは，Tex ソースとは異なり明示的なタグが無いため，提案手法は十分な性能を達成できたと言える．

5 おわりに

本稿では，論文執筆支援の複数のタスクを評価可能なデータセットを論文 PDF から作成する手法を提案した．そして，提案手法の各プロセスの評価では，本文抽出精度では先行研究に対して優れた結果を示した. 引用文献抽出および引用アンカとの対応付けの評価では，Tex ソースを対象とした先行研究に近い精度を示した．今後は，本手法による新たなデータセットの構築，およびそれを用いて複数の論文執筆支援タスクに取り組みたい．

謝辞

研究の遂行にあたり，ご助言・ご協力をいただきました，NTT コミュニケーション科学基礎研究所杉山弘晃氏，東中竜一郎氏，秋田県立大学堂坂浩二教授，大阪工業大学平博順教授，工学院大学大和淳司教授，農研機構菊井玄一郎チーム長に感謝いたします．

参考文献

[1] Simone Teufel and Marc Moens. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, Vol. 28, No. 4, pp. 409–445, 2002.

[2] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In Proceedings of AAAI 2019, 2019.

[3] Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Enhancing scientific papers summarization with citation graph. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 14, pp. 12498–12506, May 2021.

[4] Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. Scientific paper recommendation: A survey. IEEE Access, Vol. 7, pp. 9324–9339, 2019.

[5] Michael Färber, Alexander Thiemann, and Adam Jatowt. To cite, or not to cite? detecting citation contexts in text. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, Advances in Information Retrieval, pp. 598–603, Cham, 2018. Springer International Publishing.

[6] Rakesh Gosangi, Ravneet Arora, Mohsen Gheisarieha, Debanjan Mahata, and Haimin Zhang. On the use of context for predicting citation worthiness of sentences in scholarly articles. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4539–4545, Online, June 2021. Association for Computational Linguistics.

[7] Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. International Journal on Digital Libraries, Vol. 21, , 12 2020.

[8] Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic generation of citation texts in scholarly papers: A pilot study. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6181–6190, Online, July 2020. Association for Computational Linguistics.

[9] Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. Autocite: Multi-modal representation fusion for contextual citation generation. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21, p. 788–796, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. Task definition and integration for scientific-document writing support. In Proceedings of the Second Workshop on Scholarly Document Processing, pp. 18–26, Online, June 2021. Association for Computational Linguistics.

[11] Grobid. https://github.com/kermitt2/grobid, 2008–2021.

[12] Changfeng Yu, Cheng Zhang, and Jie Wang. Extracting body text from academic pdf documents for text mining. In KDIR, 2020.

[13] Isaac Councill, C. Lee Giles, and Min-Yen Kan. ParsCit: an open-source CRF reference string parsing package. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[14] Dominika Tkaczyk, Paweł Szostek, Piotr Dendek, Mateusz Fedoryszak, and Łukasz Bolikowski. Cermine – automatic extraction of metadata and references from scientific literature. 04 2014.

[15] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL '18, p. 99–108, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Riaz Ahmad and Muhammad Tanvir Afzal. Cad: an algorithm for citation-anchors detection in research papers. Scientometrics, Vol. 117, pp. 1405–1423, 2018.

[17] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. May 2008.

[18] Jian Wu, Athar Sefid, Allen Ge, and C. Giles. A supervised learning approach to entity matching between scholarly big datasets, 12 2017.

[19] Dwaipayan Roy, Kunal Ray, and Mandar Mitra. From a scholarly big dataset to a test collection for bibliographic citation recommendation, 2016.

[20] Ángela Casado-García, César Domínguez, Jónathan Heras, Eloy J. Mata, and Vico Pascual. The benefits of close-domain fine-tuning for table detection in document images. In DAS, 2020.

[21] Muhammad Waqas Ahmed and Muhammad Tanvir Afzal. Flag-pdfe: Features oriented metadata extraction framework for scientific publications. IEEE Access, Vol. 8, pp. 99458–99469, 2020.

[22] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. AxCell: Automatic extraction of results from machine learning papers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8580–8594, Online, November 2020. Association for Computational Linguistics.

表 5　タスクデータの具体例

```
“Title”: “Dataset Construction for Writing Support”
“Sentences” : [ Text1, Text2, Text3, Text4, Text5 ],
“AnswersCitationWorthiness” : [ 0, 1, 0, 1, 0 ],
“CitedNumberList” : [ 0, 2, 0, 1, 0 ],
“CollectedCitedNumberList” : [ 0, 1, 0, 1, 0 ],
“CitationAnchorList” : [ [], [ “(Zhang et al.,2020)”, “(Edo,2019)” ],
            [], [ “(Kar et al.,2021)” ], [] ],
“CitedPaperIndexList” : [ [], [ “1” , “2” ], [], [ “3” ], [], [] ],
“CitedPaperTitle” : { “1” : Title A , “2” : Title B , “3” : Title C },
“CitedPaperArXivId” : { “2” : “2019.3000v1” , “3” : “2021.2000v2” },
“CitedPaperText” : { “2” : Abstract B... , “3” : Abstract C... }
```

付録 A: タスクデータの形式

本研究で作成する，タスクデータの具体例を表 5 に示す．Narimatsu ら [10] と同様の形式であり，各要素は以下を表す．

- Title... 対象となる論文のタイトル
- Sentences... 関連研究の章本文を文単位に分割したリスト
- AnswersCitationWorthiness... 関連研究の章の各文に引用がつけられている場合は"1"，そうでなければ"0"としたリスト
- CitedNumberList... 各文の引用件数のリスト
- CitationAnchorList... 各文に含まれる引用アンカのリスト
- CitedPaperIndexList... 各文の引用の番号のリストであり，この番号が CitedPaperTitle，CitedPaperArXivID，CitedPaperText のキーと対応する
- CitedPaperTitle... 被引用文献のタイトルの辞書
- CitedPaperArXivId...arXiv API から一致する文献が取得できた被引用文献の arXiv 独自の ID の辞書 (外部検索 API に arXiv API を使用した時のみ)
- CitedPaperText... 外部検索 API から取得できた被引用文献のアブストラクトの辞書

付録 B: 章タイトル検出のための正規表現

3.1 章で使用した章タイトルの検出のための正規表現を表 6 に示す．手法 1 において，*Title_Style* はその論文の章タイトルのスタイルの文字，*RW_Title_Font_Fam* は関連研究の章タイトルのフォントの種類，*Next_Style* は関連研究の章の次の数字または英字である．手法 2 において，*Font_Larger_Than_BodyText* は本文より大きいフォントサイズ，*RW_Title_Font_Size* は関連研究の章タイトルのフォントサイズである．

付録 C: 引用アンカ検出のための正規表現

3.2 章で使用した引用アンカ検出のための正規表現について説明する．最初に Ahmed ら [16] の手法を用いて作成した正規表現を表 7 に示す．ここで，

表 6　章タイトル検出のための正規表現

手法 1.

· Related Work(Study) Section (大文字小文字を区別しない)

```
<div class=".{0,60}">Title_Style(related\s*(work|study|studies)(.|\s)*)
```

· Next Section (大文字小文字を区別しない)

```
<div class=".{0,30}RW_Title_FontFam.{0,30}">Next_Style\s*\.?(\s)*
```

手法 2.

· Related Work(Study) Section (大文字小文字を区別する)

```
<div class=".{0,30}Font_Larger_Than_BodyText.{0,30}">
R[eE][lL][aA][tT][eE][dD]\s*([wW][oO][rR][kK]|
[sS][tT][uU][dD][yY]|[sS][tT][uU][dD][iI][eE][sS])(.|\s)*
```

· Next Section (大文字小文字を区別しない)

```
<div class=".{0,30}RW_Title_Font_Size.{0,30}">(.|\s)*
```

表 7　Ahmad ら [16] の手法を元に作成した引用アンカ検出のための正規表現

```
# 数字形式 (大文字小文字を区別しない)
regexNum1 = '\[' + Tag_Value + '\]'
regexNum2 = '\[\s*([1-9][0-9\u2013\-]*\s*[,|;|\u2013](\s|\])*)+' +\
  '[1-9][0-9]*\s*(\-[1-9][09]*)?\]|\[[1-9][0-9]*\]?\s*' +\
  '[\-|\u2013]\s*\[?[1-9][0-9]*\s*\]'
regexNum  = regexNum1 + '|' + regexNum2
# 数字形式以外 (大文字小文字を区別しない)
regexStr1 = '\[[\sA-Za-z0-9\,&\.:;\+\/\(\)\-]*' + First_Author +\
  '[0-9\-&\+\s,:;\.\/\[\(]*' + Year + '[\sA-Za-z0-9\-\.\,:;\+\/\(\) ]*\]'
regexStr2 = '\[[A-Za-z0-9\-&\,:;\s\.\(\)]*' + First_Author +\
  '[A-Za-z\-\,\s]*(\sand\s|&)[A-Za-z\-\,&:\/\s\[\(\)]* \s' + Year +\
  '\s[\)A-Za-z0-9\-&\,:\/\s\(]*\]'
regexStr3 = '\([\sA-Za-z0-9\,&\.:;\+\/\-]*' + First_Author +\
  '[0-9\-&\+\s\,:;\.\/\[]*' + Year + '[\sA-Za-z0-9&\-\.\,:;\+/_]*\)'
regexStr4 = '\([A-Za-z0-9\-&\,:;\s\.]*' + First_Author +\
  '[A-Za-z\-\,\s]*(\sand\s|&)[A-Za-z\-\,&:\/\s\[]*\s' + Year +\
  '\s[A-Za-z0-9\-&\,:\/\s]*\)'
regexStr  = regexStr1 + '|' + regexStr2 +\
      '|' + regexStr3 + '|' + regexStr4
```

表 8　Gosangi ら [6] の手法を元に作成した引用アンカ検出のための正規表現

```
# 大文字小文字を区別しない
author      = '([A-Z][A-Za-z']+)'
etal        = '(et al.?)'
additional  = '(,? ((and |& )?' + author + '|' + etal + '))'
year_num    = '(19|20)[0-9][0-9][a-z]0,1'
page_num    = '(, p.? [0-9]+)?'
yp          = year_num + page_num
year        = '([,]0,1 *' + yp + '(; *' + yp + ')*' +\
'|*[\[\(]([,]0,1 *' + yp + '(; *' + yp + ')*' + ')+[\)\]])+'
regexACL = '([\(\[]?' + author + additional + '*' + year + '[\)\]]?)'
```

Tag_Value は参考文献の章から抽出した各文献の数字形式の引用アンカ，*First_Author* は参考文献の章から抽出した各文献の著者のラストネーム，*Year* は各文献の出版年である．これらの正規表現を参考文献の件数分生成し，マッチングを行う．次に，Gosangi ら [6] の手法を用いて作成した正規表現を表 8 に示す．この正規表現には，[Kobayashi, 2021],Kobayashi [2021] のような形式の引用アンカが検出できるような変更を新たに加えている．以上の正規表現は Python で動作する．