Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5673–5682
Marseille, 20-25 June 2022

Dataset Construction for Scientific-Document Writing Support by Extracting Related Work Section and Citations from PDF Papers

Keita Kobayashi[1], **Kohei Koyama**[1], **Hiromi Narimatsu**[2], **Yasuhiro Minami**[1]
[1]The University of Electro-Communication, [2]NTT Communication Science Laboratories
[1]1-5-1 Chofugaoka, Chofu, Tokyo, [2]2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto
{k2231042, k2131071}@edu.cc.uec.ac.jp
hiromi.narimatsu.eg@hco.ntt.co.jp minami.yasuhiro@is.uec.ac.jp

Abstract

To augment datasets used for scientific-document writing support research, we extract texts from "Related Work" sections and citation information in PDF-formatted papers published in English. The previous dataset was constructed entirely with Tex-formatted papers, from which it is easy to extract citation information. However, since many publicly available papers in various fields are provided only in PDF format, a dataset constructed using only Tex papers has limited utility. To resolve this problem, we augment the existing dataset by extracting the titles of sections using the visual features of PDF documents and extracting the Related Work section text using the explicit title information. Since text generated from the figures and footnotes appearing in the extraction target areas is considered noise, we remove instances of such text. Moreover, we map the cited paper's information obtained using existing tools to citation marks detected by regular expression rules, resulting in pairs of cited paper information and text of the Related Work section. By evaluating body text extraction and citation mapping in the constructed dataset, the accuracy of the proposed dataset was found to be close to that of the previous dataset. Accordingly, we demonstrated the possibility of building a significantly augmented dataset.

Keywords: Scientific Document Analysis, PDF Text Analytics, PDF Information Extraction, Corpus, Bibliometrics

1. Introduction

With the development of science and technology, the number of published academic papers is rapidly increasing[1]. Therefore, when reviewing their research methodology, it is difficult for researchers to survey relevant prior works and to cite them appropriately in their own papers. Despite researchers' best efforts in these tasks, they may still unintentionally overlook certain existing research or neglect certain citations where credit is due. Consequently, much work has been put into developing systems that support the writing of scientific papers.

To reduce the burden on researchers, the following tasks have been defined and studied for various purposes: paper summarization aimed at reducing the reading time of already searched papers (Teufel and Moens, 2002; Yasunaga et al., 2019; An et al., 2021), paper recommendation aimed at reducing the search time for unfamiliar papers (Bai et al., 2019), prediction of citation worthiness aimed at improving the efficiency of writing a paper (Färber et al., 2018; Gosangi et al., 2021), cited document allocation (Färber and Jatowt, 2020), and citation text generation (Xing et al., 2020; Wang et al., 2021).

Despite the large amount of work carried out in assisting researchers with scientific paper writing, prior studies have focused only on specific tasks, which were evaluated independently using private datasets. Therefore, it is impossible to verify the usefulness of scientific paper writing support systems in general, as applied to actual research fields, based on the results of these previously developed technologies. To solve this problem, Narimatsu et al. (2021) defined a series of tasks related to scientific-paper writing that can be pipelined as shown in Figure 1 and developed a pub-

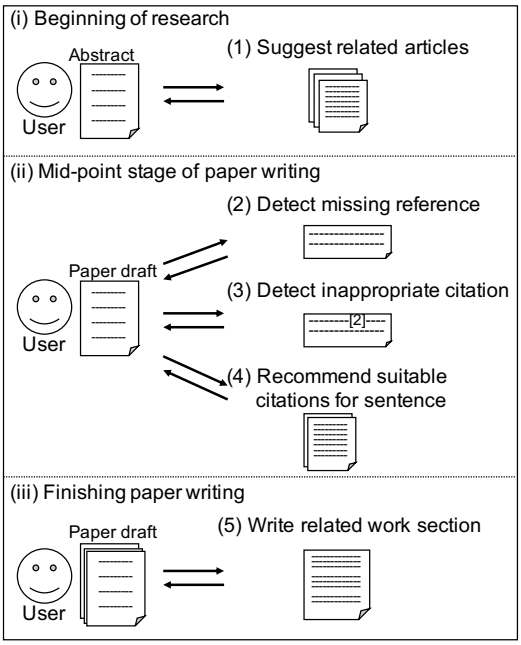

Figure 1: Scientific paper writing support for each phase of research defined in Narimatsu et al. (2021).

[1]arXiv monthly submission https://arxiv.org/stats/monthly_submissions

licly available dataset for evaluating each task. This made it possible to evaluate the system as a single pipelined writing support system, bringing this solution one step closer to a practical environment.

However, there are still challenges to overcome in achieving their targeted dataset construction. They constructed a dataset from papers having available Tex source files so that the data used for various tasks could be created automatically. Although it is possible to automatically construct task data by targeting Tex sources, papers with available Tex sources are limited in both quantity and variety of fields. For example, arXiv,[2] as one of the most popular online repositories of Tex sources, mainly focuses on the fields of physics, computer science, and mathematics, and thus it is not possible to properly evaluate the performance of tasks in paper-writing support in other fields. In addition, some papers are only available in PDF format on arXiv. In fact, out of 2,400 papers submitted to arXiv during January 2021, 584 of them (24%) did not include Tex sources.

Therefore, the purpose of this study is to automatically construct a dataset for paper writing support, even for papers that are only available as PDF-formatted documents. Specifically, as in the previous study (Narimatsu et al., 2021), we focus on the Related Work section and construct a dataset from PDF papers that can be used to evaluate an integrated system of scientific-document writing support[3][4]. In this paper, although the target text is limited to the Related Work section, the data can be expanded by increasing the number of section titles from which the text is extracted since the method itself is not dependent on this. The steps in data construction are (1) to identify the body text of the Related Work section, (2) to extract the cited data for each cited paper in the Reference section, including the title, the authors, the year of publication, etc., (3) to map the cited data to the sentence containing the citation, and (4) to retrieve the cited paper using the cited data. However, the following challenges must be overcome in order to create an equivalent dataset from a PDF on par with that from a Tex source.

- **Identifying section titles** In a Tex source, the \section{} command makes it easy to extract the section title and body text in the section. However, in contrast, there is no specific mark for a title in PDF, so it must be identified by visual information such as the type and size of the font.
- **Cleaning the extracted text** The text extracted from a PDF document often includes headers, footers, footnotes, formulas, and strings in figures and tables that are unrelated to the sentences in that location. Consequently, these become noise in natural language processing.
- **Mapping cited datum to citation anchor** Cited datum refers to a group of the authors, the paper title, the journal title, etc., and citation anchor refers to the in-text citation marks such as [1] and Kobayashi et al. (2021). In Tex source, the citation anchor can be easily retrieved and mapped by using \cite{} commands and bib files. In PDF, however, not only is there no explicit mark, the format of a citation anchor also varies from paper to paper, so it is difficult to map the citation anchor to the cited datum.

To tackle these challenges, we adopted several existing high-performance tools and proposed methods to improve accuracy. Specifically, to identify section titles and clean the extracted text, the method of PDFBoT (Yu et al., 2020) is used in our method. Visual features are obtained by converting PDF format to HTML with reference to PDFBoT, and section titles are identified with our own method using these features. Then, the text in the Related Work section is extracted by removing headers, footers, formulas, and figures with PDFBoT, and tables are removed by our own method. Next, we used and improved upon the methods of Ahmad and Afzal (2018) and Gosangi et al. (2021) to obtain citation anchors and applied GROBID (GROBID, 2008 - 2021) to obtain cited data.

Using above methods, we evaluated if our method improves the accuracy of text extraction in a specific section and maps citation anchors to cited data with an accuracy close to that of using Tex as a source.

2. Related Work

Many studies and tools have aimed to extract information from PDF papers. In this section, we focus on the aspects relevant to our research, i.e., the extraction of body text, the extraction of cited data, and citation anchor mapping.

For body text extraction, CERMINE (Tkaczyk et al., 2014) has been proposed as a tool to extract section titles and body text from PDF papers using a conditional random field (CRF) model. However, this tool does not consider the identification and removal of noise such as symbols or strings that are not part of the body text itself but are unintentionally included in the body text. PDFdigest (Ferrés et al., 2018) extracts section titles and body text, removes footnotes, and identifies figure and table captions by converting PDF papers to HTML, using a rule-based algorithm according to HTML text features. GROBID (GROBID, 2008 - 2021) has been proposed as a tool for extracting section titles

[2] arXiv submission rate statistic https://arxiv.org/help/stats/2020_by_area/index

[3] Dataset constructed from PDF papers licensed under CC BY 4.0 on https://aclanthology.org/ is available at https://github.com/citation-minami-lab/acl-citation-dataset.

[4] Details of the task data are described in Appendix.

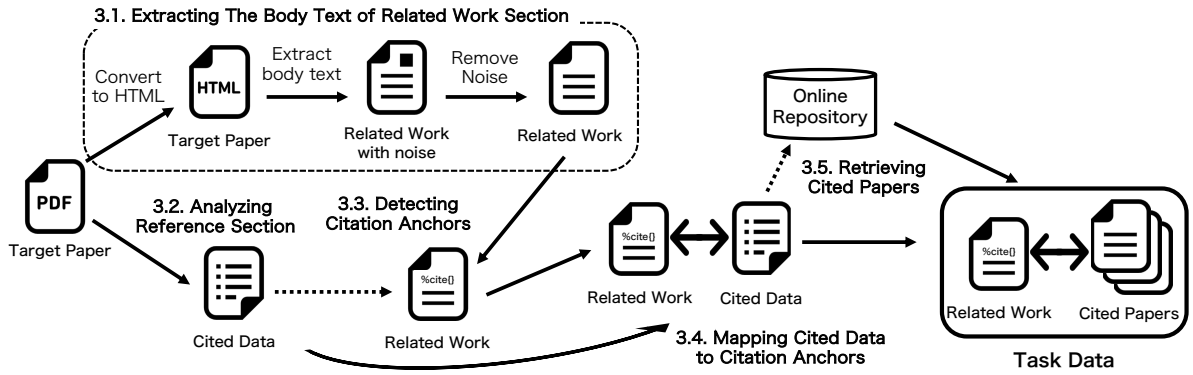

Figure 2: Process of creating task data. The dotted arrow represents an optional passing of information that is used as a cue.

and body text as well as recognizing mathematical expressions, figures, and tables using the CRF model. However, because it does not consistently recognize section titles and figures[5], it cannot be reliably used to identify the Related Work section, and it even sometimes mistakenly includes text from figures and tables in its extracted text. PDFBoT (Yu et al., 2020) extracts only the body text by removing section titles, footnotes, headers, footers, figures, tables, display mode formulas, and the Reference section from a paper in PDF format. Similar to PDFDigest, this tool uses a rule-based algorithm, leveraging text features obtained by converting PDF to HTML using pdf2htmlEX[6], which achieves the removal of unwanted elements with high accuracy. However, because PDFBoT also removes section titles and the Reference section, it cannot be applied directly for the purpose of this research.

The following tools have been proposed for extracting cited data, detecting citation anchors, and mapping cited data to citation anchors: ParsCit (Councill et al., 2008), CERMINE (Tkaczyk et al., 2014), and GROBID (GROBID, 2008 - 2021). Among the three proposed methods, it is clear that GROBID can extract cited data with the highest accuracy (Tkaczyk et al., 2018), and would seem at first to be the most suitable choice. However, it cannot be applied to our methods because it cannot take as input only the Related Work section but requires the input of the entire PDF paper.

Ahmad and Afzal (2018) proposed a method to detect citation anchors with high accuracy using previously extracted cited data. Furthermore, Gosangi et al. (2021) proposed a method to detect citation anchors while building a dataset using the ACL Anthology Reference Corpus (Bird et al., 2008). Although the format of the citation anchors that they managed to detect is limited, they could still detect citation anchors that are not covered by Ahmad and Afzal (2018).

In some studies on paper-writing support, the datasets for evaluation were constructed from PDF papers (Wu et al., 2017; Roy et al., 2016). These studies used CiteSeerX[7] for body text extraction and citation anchor mapping. However, CiteSeerX only extracts the text around the citation anchors, and it cannot retrieve the entire specific section. In addition, CiteSeerX uses GROBID (GROBID, 2008 - 2021) for body text extraction, which does not sufficiently remove noise in the body text.

As shown above, no method has achieved an overall high accuracy in extracting both body text and cited data, so we decided to combine these few methods that have high accuracy in extracting either body text or cited data. Specifically, we used PDFBoT for extracting body text, GROBID for extracting cited data, and the methods of Ahmad and Afzal (2018) and Gosangi et al. (2021) for detecting citation anchors.

3. Dataset Construction Methods

To create the task data for our targeted paper-writing support, it is necessary to extract the body text of the Related Work section, extract the cited data, map the cited data to the citation anchors in *a target paper*, and retrieve cited papers. The process of data creation is shown in Figure 2 and explained in this section. The labeled number at each process in the figure corresponds to the section number below.

3.1. Extracting Body Text of Related Work Section

In our proposed method, we first detect the title of the Related Work section and the title of the next section, and then we extract all text between the two section

[5]GROBID Documentation Benchmark https://grobid.readthedocs.io/en/latest/Benchmarking-pmc

[6]https://github.com/pdf2htmlEX/pdf2htmlEX

[7]https://citeseerx.ist.psu.edu/index

In the case with a section number
· Related Work(Study) Section (case-insensitive) *Title_Style*(related\s*(work|study|studies)(.|\s)*)
· Next Section (case-insensitive) *Next_Style*\s*\.?(\s)*
In the case without a section number
· Related Work(Study) Section (**case-sensitive**) 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 (case-insensitive) (.|\s)*

Table 1: Regular expressions for detecting section titles. *Title_Style* is the characters for the style of the section title of the paper, *RW_Title_Font_Family* is the font type of the title of the Related Work section, and *Next_Style* is the number or letter of the next section after the Related Work section in the case with a section number. *Font_Larger_BodyText* is the font size larger than the body text, and *RW_Title_Font_Size* is the font size of the title of the Related Work section in the case without a section number.

titles as the body text. We considered all sections with titles beginning with "Related Work" or "Related Study (Studies)" as sections for related work.

3.1.1. Converting PDF to HTML (Preprocessing)

By converting the target paper in PDF to a HTML-formatted paper using pdf2htmlEX with reference to PDFBoT (Yu et al., 2020), the visual features obtained from HTML tags such as font size and font type are used to identify section titles and remove noise from the body text. First, using the converted HTML, we remove unnecessary symbols, obtain the coordinate of each line for the precise removal of noise, and collect the statistics of the features in the body text, as described below.

(a) Remove tags, i.e., `<a>`, `<img>`, and `<span>`, that are not needed for extracting section titles and body text.

(b) As in PDFBoT (Yu et al., 2020), the x-coordinate of the beginning of each line is obtained using the HTML structure for use in noise removal.

(c) The font size with the highest frequency of occurrence is regarded as the font size of the body text.

(d) Count the x-coordinates at the beginning of every line and determine whether the layout is single-column or double-column based on the coordinates and their frequency. Specifically, the top two x-coordinates with the highest frequencies are compared. If the coordinate difference is large and their frequencies are comparable, the layout is considered to be double-column. Otherwise, the layout is considered to be single-column. In practice, a layout is considered to be double-column if the difference between the top two x-coordinates is larger than 100 px and the frequency difference is less than two-fifths (0.4) of the frequency of the Top1 coordinate. This threshold was determined empirically.

(e) For all lines, measure the line spacing, and regard the line spacing larger than 7 px and with the highest frequency as the size of the line spacing in the body text.

3.1.2. Extracting Related Work Section

Then we detect the title of the Related Work section and the title of the next section and extract all text between the two section titles as the Related Work section. Since the section number is an important clue in detecting section titles, we distinguish between those with a section number and those without a section number in the following cases.

In the case with a section number: Since many papers have "Introduction" as the first section, we extract that line and identify the style of the paper's section numbers from the first letter (e.g., "1.", "I.", "A."). A line that begins with a character that matches that style and "Related Work" or "Related Study (Studies)" is detected as the title of the Related Work section. Then, a line that matches the font type of the title of the Related Work section and begins with a character that is consecutive to the section number of the Related Work is detected as the next section title.

In the case without a section number: A line that is larger than the font size of the body text and begins with "Related Work" or "Related Study (Studies)" is detected as the title of the Related Work section. Then, the line whose font size is equal to the title of the Related Work section is detected as the next section title.

The regular expressions used to detect section titles are shown in Table 1. Our method avoids the false detection of inline section titles in the body text for reference purposes, such as "∼ shown in Related Work section", by using the section title number and font style.

Finally, all text between the title of the detected Related Work section and the title of the next section is extracted as the body text of the Related Work section.

3.1.3. Removing Noise from the Body Text

The extracted body text often contains headers, footers, footnotes, mathematical formulas, and text in figures and tables that are not related to the sentences in that

```
# Numeric Format (case-insensitive)
regexForNumeric1 = '\[ (1|2|3|4| ... |The_Number_of_Cited_Data) \]'
regexForNumeric2 = '\[\s*([1-9][0-9\u2013\-]*\s*[,|;|\u2013](\s|\])*)+[1-9][0-9]*\s*' +\
                   '(\-[1-9][09]*)?\]|\[[1-9][0-9]*\]?\s*[\-|\u2013]\s*\[?[1-9][0-9]*\s*\]'
regexForNumeric  = regexForNumeric1 + '|' + regexForNumeric2
# Non-numeric Format (case-insensitive)
regexForStr1 = '\[[\sA-Za-z0-9\,&\.:;\+\/\(\)\-]*' + First_Author_Name +\
               '[0-9\-&\+\s,:;\./\[\(]*' + Year + '[\sA-Za-z0-9\-\.\,:;\+/\(\) ]*\]'
regexForStr2 = '\[[A-Za-z0-9\-&\,:\s\.\(\)]*' + First_Author_Name +\
               '[A-Za-z\-\,\s]*(\sand\s|&)[A-Za-z\-\,&:/\s\[\(\)]* \s' + Year +\
               '\s[\)A-Za-z0-9\-&\,:\/\s\(]*\]'
regexForStr3 = '\([\sA-Za-z0-9\,&\.:;\+/\-]*' + First_Author_Name +\
               '[0-9\-&\+\s\,:;\./\[]*' + Year + '[\sA-Za-z0-9&\-\.\,:;\+/_]*\)'
regexForStr4 = '\([A-Za-z0-9\-&\,:\s\.]*' + First_Author_Name +\
               '[A-Za-z\-\,\s]*(\sand\s|&)[A-Za-z\-\,&:/\s\[]*\s' + Year +\
               '\s[A-Za-z0-9\-&\,:\/\s]*\)'
regexForStr  = regexForStr1 + '|' + regexForStr2 + '|' + regexForStr3 + '|' + regexForStr4
```

Table 2: Regular expressions for citation anchor detection based on the method of Ahmad and Afzal (2018). *The_Number_of_Cited_Data* is the number of cited data in the target paper in Numeric Format. *First_Author_Name* is the last name of the first author in the cited datum and *Year* is the year of publication in the cited datum in Non-numeric Format.

```
# case-insensitive
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 + ')*' + '| *[\[\(][,]{0,1} *' + yp + '(; *' + yp + ')*' + ')+[\)\]])+'
regexForACL = '([\(\[]?' + author + additional + '*' + year + '[\)\]]?)'
```

Table 3: Regular expressions for citation anchor detection based on the method of Gosangi et al. (2021). The boldface characters in *regexForACL* are items we have added.

location. Since these elements become noise in natural language processing, we remove them as follows.

Removing headers, footers, and footnotes In general, the font size of non-body text such as text in headers, footers, and footnotes is smaller than that of the body text. Therefore, as in PDFBoT (Yu et al., 2020), strings that are more than 1 px smaller than the font size of the body text are assumed to be non-body text, and thus they are removed.

Removing captions of figures and tables Captions of figures and tables inserted in the middle of the text are often mistakenly extracted as the body text. Therefore, we remove lines starting with "Table", "Figure", or "Fig.", which indicate the captions of the figures and tables. If a caption spans multiple lines, a line spacing wider than the line space between the first and second lines is considered a break in the caption. We remove the lines that are regarded as captions. Note that we remove only lines containing these strings and that have a wider line spacing than that of the body text, to avoid incorrect removal of text for reference purposes, such as "∼ shown in Figure 1".

Removing figures and formulas Here, we remove only mathematical formulas written in display mode, not those such as "λ" that appear inline. The reason for this is that when a mathematical formula appears in a sentence, deleting it often results in the sentence becoming unnatural. As in PDFBoT (Yu et al., 2020), we assume that strings in mathematical formula and figures have a starting x-coordinate of more than 30 px to the right of the body text, and thus we remove them.

Removing tables We use an image recognition model (Casado-García et al., 2020) to detect the coordinates of a table with high accuracy and then remove that table. Specifically, the target PDF paper is converted to image format using pdf2image[8] and input into the model. Then, we remove all strings that appear within the range detected by the model plus a margin of 10 px.

After the above removal procedures, we normalize the text by measures such as removing HTML tags, removing consecutive spaces, and decomposing ligatures.

[8] https://pypi.org/project/pdf2image/

Target anchors (a)	[1], [1,2], [1;2], [1-3], [1]-[3], [1,3-5] smith 2021, (Smith, 2021), (Smith and Jones, 2021), Smith et al. (2021a) (Smith et al., 2020, 2021a), [Smith et al., 2021; Jones et al., 2020]
Target anchors (b)	Smith et al. [2021]
Not target anchors (c)	[sfc+01], [BG*], [DB-Main], [Wol94,GV95], [N], [N,P,X], $\sim^{1-3}$.
Not target anchors (d)	$\sim^{1}$.

Table 4: Examples of citation anchors to be targeted or not targeted for detection

3.2. Analyzing Reference Section

We use GROBID (GROBID, 2008 - 2021) to analyze the reference section and retrieve the cited data. GROBID divides the reference section in the input PDF paper into each cited paper, identifies each cited paper with its title, authors, published year, etc., and outputs the results in TEI XML format. We extract cited data from this output. However, GROBID sometimes fails to properly divide the reference section. Taking into account that numeric citation anchors such as "[1]" are mapped to the citation anchors based solely on the order of the cited data, if the cited data could not be split correctly, there would be a mismatch between the citation anchor and its mapped cited data, since the order of the cited data itself would be out of sequence. To avoid this problem, we do not directly input target PDF papers with numerical citation anchors into GROBID. Instead, strings divided into individual papers are passed into GROBID to obtain the results in TEI XML format tagged with author, title, year, etc.

3.3. Detecting Citation Anchors

The format of the citation anchor varies from paper to paper. To detect a large number of citation anchors, we combine two previously proposed regular expression detection methods (Ahmad and Afzal, 2018; Gosangi et al., 2021). We also provide a way to detect citation anchors that are not covered by those two methods. Their regular expressions are shown in Table 2 and Table 3.

These methods allow detection with high accuracy, but then a problem arises: If there is a word that starts with a capital letter just before the citation anchor, such as "Later, Kobayashi (2021) did $\sim$", that word would also be included in the citation anchor. Since such words are mainly adverbs, we listed commonly used adverbs to be excluded from the citation anchors. On the other hand, citation anchors that are difficult to detect and map with our method are not targeted. Table 4 shows the target and non-target citation anchors. Here, (a) is the citation anchor that is the target of the existing two methods and also the target of this study, and (b) is the citation anchor that we newly added as the target. We excluded (c), which is difficult to map and detect using existing methods, and (d), which is not covered by existing methods. To elaborate further, the first six citation anchors in (c) are difficult to map to the cited data extracted by GROBID, and the final ones in (c) and (d) are difficult to distinguish from references to footnotes. We examined the number of citation anchors that were not targeted using 500 randomly collected articles from arXiv, and we found 40 of these in total. Therefore, even if we excluded the papers with these types of citation anchors, the impact on the total data volume would be negligible.

3.4. Mapping Cited Data to Citation Anchors

For numeric citation anchors, we map them to the cited data in the order that match the citation anchor numbers. For non-numeric citation anchors, we extract the year from the citation anchor and retrieve the first author's name by removing the string after "et al.", "and", and "&". Using this information, we map the cited data to the citation anchors.

3.5. Retrieving Cited Papers

We input the titles of the cited data mapped to citation anchors into the online repository's external search API and retrieve the abstracts as the information of the cited papers if the same paper is found. We consider them to be the same paper only if the title obtained as a search result matches the input title exactly, although case-insensitive matches are allowed.

4. Evaluation

We evaluated our proposed method by comparing the constructed dataset from the two viewpoints of extracting the Related Work section and extracting citation information.

4.1. Performance of Extracting Related Work Section

The accuracy of detecting the title and extracting the body text of the Related Work section is evaluated by comparing these tasks' performances with those of previous works, i.e., Narimatsu et al. (2021) and GROBID (GROBID, 2008 - 2021).

4.1.1. Metrics

For evaluation, we used Word Error Rate (WER) and Sentence Error Rate (SER), which can evaluate the accuracy of both text extraction and noise removal simultaneously. These are expressed in the following formulas (1),(2):

$$WER = \frac{I_w + D_w + S_w}{N_w}, \quad (1)$$

	Successes	Failures
GROBID (2008 - 2021)	103	10
Narimatsu et al. (2021)	**113**	**0**
Proposed	110	3

Table 5: Number of successes and failures in detecting titles of Related Work sections for 113 papers.

	WER	SER
GROBID (2008 - 2021)	0.167	0.542
Narimatsu et al. (2021)	0.188	0.744
Proposed	**0.086**	**0.481**

Table 6: Evaluation of body text extraction and noise removal by WER and SER under the condition of removing all citation anchors.

where I_w is the number of words inserted incorrectly, D_w is the number of words deleted incorrectly, S_w is the number of words substituted incorrectly, and N_w is the number of words in the correct answer data.

$$SER = \frac{I_s + D_s + S_s}{N_s}, \quad (2)$$

where I_s is the number of sentences inserted incorrectly, D_s is the number of sentences deleted incorrectly, S_s is the number of sentences included in the correct data but with some errors, and N_s is the number of sentences in the correct data.

4.1.2. Data

We prepared 120 target papers containing related work sections, randomly selected from the list of papers in Axcell (Kardas et al., 2020) used by Narimatsu et al. (2021) to construct their dataset. Then, we manually created the correct answer by extracting all of the sentences in the related work sections and removing noise such as figures, tables, and headers as defined above.

In addition, since WER and SER directly compare the correct answer text and extracted text, it is difficult to make accurate comparisons between methods if the text contains mathematical formulas or expressions due to the influence of slight differences in character codes. Therefore, we excluded papers including mathematical formulas in the body text of the Related Work section and used the remaining 113 papers for the evaluation of body text extraction.

4.1.3. Evaluation of the Number of Detected Related Work Section Titles

First, we used this evaluation data to examine the number of successes or failures of each method by detecting the titles of the Related Work sections (Table 5). Since the detected section titles are enclosed by <head> tags in GROBID, we checked whether the title of the Related Work section appeared in the area enclosed by <head> tags.

Narimatsu et al. (2021) succeeded in detecting the title of the related work section in all of the evaluation data because they used Tex source and could extract them by identifying \section{} tags. The results of our method are better than those of GROBID, and the difference is significant. Therefore, our method's performance can be considered satisfactory.

4.1.4. Evaluation of Body Text Extraction and Noise Removal

To make a fair evaluation, we devised a method and an evaluation standard. For the method, we removed typical noise using the following process for the body text extracted by GROBID.

(a) Remove <formula> tags and enclosed formulas.

(b) Remove <ref> tags surrounding citation anchors and references to figures, tables, and formulas contained in the body text.

(c) Remove or add the spaces in citation anchors that cause slight differences from the correct answer data (e.g., GROBID extraction string: [1,2], correct answer data: [1, 2]).

(d) Revert the automatically converted expressions such as "(Kobayashi et al., 2011)(Kobayashi et al., 2012)" and "[1][2][3]" to "(Kobayashi et al., 2011,2012)" and "[1]-[3]".

Note that, since GROBID outputs the detected figures and tables separately from the body text, it was not necessary to delete the figures and tables.

Moreover, we modified the extracted results to align the evaluation conditions between methods. Narimatsu et al. (2021) extracted the body text from a Tex source, and since the citation anchor is in the form of a Tex tag, it differs from the correct data created from a PDF source. Therefore, we removed all citation anchors from the extracted body text in each method. Since the citation anchors in the correct answer text are left unchanged, the WER and SER of all three methods are larger than their actual values. However, they still can be compared under the same conditions. Under the above conditions, we evaluated our proposed method, GROBID, and Narimatsu et al. (2021) using only the evaluation data that successfully detected the titles of Related Work sections in all three methods (Table 6).

The results show that our method achieves the best score for both WER and SER. The reason that Narimatsu et al. (2021) has the highest WER and SER is that they did not completely remove the commented-out strings using the authors' original commands, even though they removed figures, tables, footnotes, etc. using Tex symbols. The main errors of GROBID are due to the lack of words and sentences that should be included in the body text as well as failure to detect figures and tables. In contrast, the main error of our method is the insertion of a space in the middle of a word. This is because the tag in HTML for-

	WER	SER
GROBID (2008 - 2021)	0.087	0.175
Proposed	**0.010**	**0.054**

Table 7: Evaluation of body text extraction and noise removal accuracy by WER and SER under the condition of **not** removing citation anchors.

	WER	SER
Target: PDF+Tex	0.010	0.054
Target: PDF	0.007	0.028

Table 8: Evaluation of body text extraction of our proposed method using papers with Tex sources (Target: PDF+Tex) and randomly collected papers (PDF) without regard to the availability of Tex sources.

mat, inserted in the middle of a word for notation purposes, could not be handled properly. The use of language models could possibly correct these errors.

Next, we calculated the actual WER and SER for our method and GROBID without removing citation anchors (Table 7). The results show that the WER and SER of our proposed method are both lower than those of GROBID, and about 0.01 and 0.05, which indicate success in accurately extracting the body text of the Related Work section and in removing noise.

In the above evaluation, the target papers used were limited to those in which Tex source is available. Therefore, to confirm that our proposed method works for the set of papers without Tex sources, we also compared performance using a different dataset selected randomly without knowing whether the paper contains Tex sources. These data include 120 papers: 70 papers randomly selected from arXiv and 50 papers randomly selected from ACL-Anthology. As in the above evaluation, we excluded 21 papers containing mathematical formulas in the body text of the Related Work section and used the remaining 99 papers for this evaluation.

Table 8 shows a comparison between the results of papers having Tex source (Target: PDF+Tex) and those of papers that do not publish their Tex sources (Target: PDF). In this evaluation data, the number of failed detections of titles of the Related Work section by our method was 5. The WER and SER were calculated for 94 cases, excluding those 5 papers.

The results show that both WER and SER of Target:PDF are significantly lower than Target:PDF+Tex. This indicates that the accuracy of body text extraction does not change significantly depending on whether the Tex source is available. Therefore, it was confirmed that our method performs as well for papers in PDF format.

4.2. Evaluation of Cited Data Extraction and Mapping

The detection of citation anchors, the extraction of cited data, and the mapping of them to citation anchors were evaluated in an integrated manner with the previous study Narimatsu et al. (2021).

For the evaluation data, 2,786 papers were randomly selected from the list of Axcell (Kardas et al., 2020). Each method detected the citation anchors in the Related Work section of these papers, extracted the cited data, and mapped them to the citation anchors. Then, we searched for the titles of the cited data mapped to the citation anchors by inputting them into the arXiv API.

	Number of papers
Narimatsu et al. (2021)	4,874
Proposed	4,225

Table 9: Number of papers that match the papers mapped to the citation anchors found in arXiv.

Table 9 shows the number of papers in citation which can be found in arXiv, and our proposed method reaches about 87% of Narimatsu et al. (2021). In the Tex source they targeted, the cited data that should map to the citation anchors could be easily retrieved by searching the bib file using a string surrounded by citation tags. In addition, the title of the cited data is clearly indicated as a string following "title =". On the other hand, there is no such tag to provide hints for extracting cited data and mapping to the citation anchors in a PDF source. Therefore, it is clear that our method achieves satisfactory performance.

5. Conclusion

In this paper, we proposed a method to construct a dataset from papers in PDF format that can be used to evaluate various tasks of scientific-document writing support. In the evaluation of body text extraction and noise removal of the Related Work section, our method outperformed an existing work (Narimatsu et al., 2021) and the PDF paper analysis tool GROBID. Moreover, in the evaluation of the extraction of citation information, our method performed well compared to conventional methods.

In future work, we will expand the dataset using our method and then use it to work on various tasks for the practical application of a scientific-document writing support system.

6. Acknowledgements

We are grateful to Dr. Hiroaki Sugiyama and Dr. Ryuichiro Higashinaka of NTT Communication Science Laboratories, Prof. Kohji Dohsaka of Akita Prefectural University, Prof. Hirotoshi Taira of Osaka Institute of Technology, Prof. Junji Yamato of Kogakuin University, and Dr. Genichiro Kikui of Japan Science and Technology Agency/Kogakuin University for their advice and cooperation in carrying out this research.

“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... }

Table 10: Example of task data

Appendix: Format of Task Data

An example of task data created in this study is shown in Table 10. It has the same format as Narimatsu et al. (2021), with each element indicating the following contents.

- Title: Title of the target paper.
- Sentences: A list of sentences divided from the body text of Related Work section.
- AnswersCitationWorthiness: A list of “0” or “1” that indicates whether a sentence has a citation.
- CitedNumberList: A list of the number of citations in each sentence.
- CollectedCitedNumberList: A list of the number of cited papers in which the information was retrieved from an external API in each sentence.
- CitationAnchorList: A list of citation anchors in each sentence.
- CitedPaperIndexList: A list of citation numbers in each sentence. This number corresponds to the keys for CitedPaperTitle, CitedPaperArXivID, and CitedPaperText.
- CitedPaperTitle: Dictionary of cited paper titles.
- CitedPaperArXivId: A dictionary of unique arXiv IDs of cited papers retrieved from the arXiv API (only if arXiv API is used).
- CitedPaperText: A dictionary of abstracts of cited papers retrieved from an external search API.

7. Bibliographical References

Ahmad, R. and Afzal, M. T. (2018). CAD: an algorithm for citation-anchors detection in research papers. Scientometrics, 117:1405–1423.

An, C., Zhong, M., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2021). Enhancing scientific papers summarization with citation graph. 35(14):12498–12506.

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., and Xia, F. (2019). Scientific paper recommendation: A survey. IEEE Access, 7:9324–9339.

Casado-García, Á., Domínguez, C., Heras, J., Mata, E., and Pascual, V. (2020). The benefits of close-domain fine-tuning for table detection in document images. In International Workshop on Document Analysis Systems, pages 199–215. Springer.

Councill, I., Giles, C. L., and Kan, M.-Y. (2008). ParsCit: an open-source CRF reference string parsing package. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08). European Language Resources Association (ELRA), May.

Färber, M. and Jatowt, A. (2020). Citation recommendation: approaches and datasets. International Journal on Digital Libraries, 21(4):375–405.

Färber, M., Thiemann, A., and Jatowt, A. (2018). To cite, or not to cite? detecting citation contexts in text. In European conference on information retrieval, pages 598–603. Springer.

Ferrés, D., Saggion, H., Ronzano, F., and Bravo, À. (2018). PDFdigest: an adaptable layout-aware PDF-to-XML textual content extractor for scientific articles. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), May.

Gosangi, R., Arora, R., Gheisarieha, M., Mahata, D., and Zhang, H. (2021). 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, pages 4539–4545. Association for Computational Linguistics, June.

GROBID. (2008 - 2021). https://github.com/kermitt2/grobid.

Kardas, M., Czapla, P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., and Stojnic, R. (2020). AxCell: Automatic extraction of results from machine learning papers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8580–8594. Association for Com-

putational Linguistics, November.

Narimatsu, H., Koyama, K., Dohsaka, K., Higashinaka, R., Minami, Y., and Taira, H. (2021). Task definition and integration for scientific-document writing support. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 18–26. Association for Computational Linguistics, June.

Teufel, S. and Moens, M. (2002). Articles summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 28(4):409–445.

Tkaczyk, D., Szostek, P., Dendek, P., Fedoryszak, M., and Bolikowski, c. (2014). Cermine – automatic extraction of metadata and references from scientific literature. 04.

Tkaczyk, D., Collins, A., Sheridan, P., and Beel, J. (2018). 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, page 99–108. Association for Computing Machinery.

Wang, Q., Xiong, Y., Zhang, Y., Zhang, J., and Zhu, Y. (2021). 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, page 788–796. Association for Computing Machinery.

Xing, X., Fan, X., and Wan, X. (2020). Automatic generation of citation texts in scholarly papers: A pilot study. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6181–6190. Association for Computational Linguistics, July.

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., and Radev, D. R. (2019). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 7386–7393.

Yu, C., Zhang, C., and Wang, J. (2020). Extracting body text from academic pdf documents for text mining. In KDIR.

8. Language Resource References

Bird, Steven and Dale, Robert and Dorr, Bonnie and Gibson, Bryan and Joseph, Mark and Kan, Min-Yen and Lee, Dongwon and Powley, Brett and Radev, Dragomir and Tan, Yee Fan. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. European Language Resources Association (ELRA), ISLRN 150-170-243-077-5.

Dwaipayan Roy and Kunal Ray and Mandar Mitra. (2016). From a Scholarly Big Dataset to a Test Collection for Bibliographic Citation Recommendation.

Wu, Jian and Sefid, Athar and Ge, Allen and Giles, C. (2017). A Supervised Learning Approach To Entity Matching Between Scholarly Big Datasets.