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

Summary

・Our purpose is to construct a dataset to support the writing of the Related Work section.
・The problem with our existing dataset is that it only cover Tex sources.
  ・Published papers are often only in PDF format and do not include Tex.
・To augment the dataset from PDF papers, we need to analyze the visual information in PDF and it is not easy, so we solve the problem to analyze them in this study.

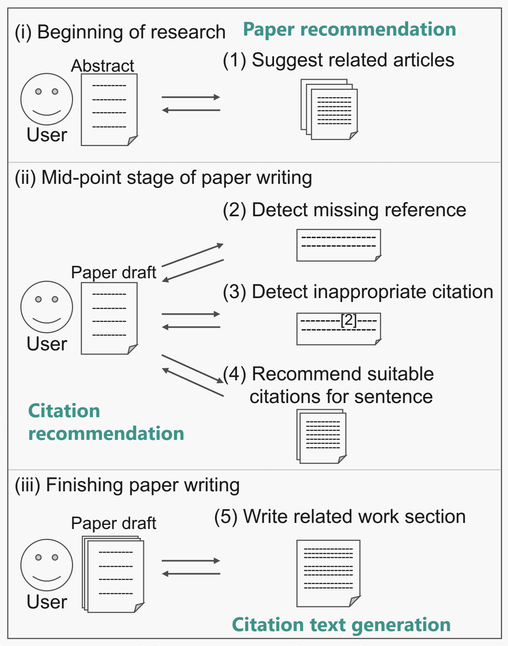
Motivation and Purpose

Background

・Rapid increase of academic papers
  ・Makes it difficult for researchers to survey relevant prior works and appropriately cite them in their papers
・Proposed tasks to reduce the difficulties for researchers such as paper recommendation, citation recommendation and citation text generation

Our Previous Study for integrated Evaluation and its Challenges

・Those existing studies of paper writing support have focused only on specific tasks and evaluated them using private datasets.
・To solve this problem, we integrated tasks (e.g. paper recommendation) to support writing according to the research stage and constructed evaluation dataset from Tex Sources [Narimatsu et al., 2021]
・Problem of this dataset: Papers with available Tex source are limited both quantity and variety.



Writing support tasks we integrated previously

Purpose of This Study

・Automatically construct dataset from PDF papers to significantly increase data volume
  ・Advantages
    ・Increase data quantity so that models for writing support can be trained and evaluated with sufficient data.
    ・Increase variety of paper fields so that researchers can be supported in multiple fields.

Evaluation

Title and Body Text Extraction

・Baseline
  ・Narimatsu et al. [2021]: Our previous work using Tex as sources
  ・GROBID [GROBID, 2008-2021]: Analysis tool of PDF Papers

Result

	Successes	Failures
Narimtsu et al.	**113**	**0**
GROBID	103	10
Proposed	110	3

Number of successes and failures in extracting titles of Related Work sections for 113 papers

	WER	SER
Narimatsu et al.	0.188	0.744
GROBID	0.167	0.542
Proposed	**0.086**	**0.481**

Evaluation result of text extraction and noise removal by Word Error Rate (WER) and Sentence Error Rate (SER)

・The accuracy of our method in extracting section titles is comparable to Narimatsu et al.
・The WER and SER of our method achieve the best score.

Citation Information Extraction

・Data: 2786 papers with Related Work sections randomly selected
・Evaluation process
  1. Detected the citation anchors in the Related Work section, extracted the cited data, and mapped them to the citation anchors
  2. Searched arXiv API for paper titles mapped to citation anchors
  3. Counted the number of papers found in arXiv

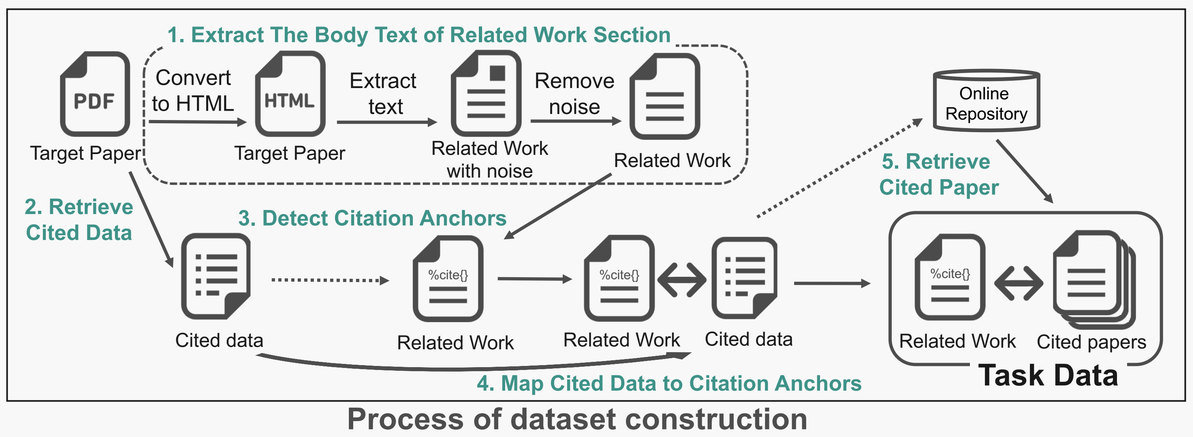
Result

・Our method achieved close accuracy to Narimatsu et al. despite extracting citations from PDF is more difficult than from Tex because PDF has no obvious tags and bib and bbl files.

	Number of matches
Narimatsu et al.	4,874
Proposed	4,225

Evaluation result of text extraction and noise removal by Word Error Rate (WER) and Sentence Error Rate (SER)

Method



Process of dataset construction

1. Extracting The Body Text of Related Work Section

I. Visual Feature Extraction

i.) Convert PDF to HTML format to obtain visual text features

ii.) Presume layout information such as number of columns, start point, font size, font family, and line spacing of body text

Feature : Coordinate, Font type, size

<div x1 y1 ff1 fs1 >text

II. Extract Text of Related Work section

i.) Detect titles of Related Work Section and the next section by classify cases with or without a section number

ii.) Extract all text between the detected two section titles

With section number

<div ...>1. Introduction
<div ... ff1 ...>[Number]. Related Work
<div ... ff1 ...>[Number]+1. Method

Without section number

<div ... [fsL]* ...>Related Work
<div ... fs1 ...>Method

*[fsL]: Font size larger than body text

III. Noise removal from extracted Text

Removed element	Method
Header, Footer Footnote	Remove strings smaller than body text
Caption of figure or table	Remove lines that begin with “Table”, “Figure”, “Fig.” and have wide line spacing
Figure, Formula	Remove strings whose starting x-coordinate is to the right of that of body text
Table	Detect tables using CNN [Casado-García et al., 2020] and remove text in the detected areas

2. Retrieve Cited Data from Reference section

・Basically use GROBID [GROBID, 2008-2021], a tool that can extract cited data with high accuracy, for retrieving cited data by inputting PDF
If citation anchors are numeric: To reduce errors of splitting Reference, split Reference section to pieces of cited data in advance, and input each piece to GROBID in order

3. Detect Citation Anchors

・Integrate two methods of regular expressions [Ahmad et al. 2018], [Gosangi et al. 2021] , which can detect citation anchors with high accuracy and modify them

4. Map Cited Data to Citation Anchors

・Check the number or the name and year in citation anchors against the cited data and map the matches

5. Retrieve Cited Papers

・Search for cited paper’s title on online repository (i.e. arXiv API) and retrieve papers that match input titles with case-insensitive exact match

Conclusion

・We proposed the method of constructing dataset from PDF Papers for scientific-paper writing support tasks.
  ・Dataset from PDF papers in ACL Anthology is available at: https://github.com/citation-minami-lab/acl-citation-dataset
・Evaluation results:
Demonstrated the possibility of augmenting the Tex sourced dataset
  ・Body text: Our method outperformed previous studies.
  ・Citation information: Our method is comparable to Tex sourced method.

Reference

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, 2021. Association for Computational Linguistics.

Grobid. https://github.com/kermitt2/grobid, 2008-2021.

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

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, 2021. Association for Computational Linguistics.

Angela Casado-Garcia, Cesar Dominguez, Jonathan Heras, Eloy Mata, and Vico Pascual. The benefits of close-domain fine-tuning for table detection in document images. In International Workshop on Document Analysis Systems, pp. 199–215. Springer, 2020.