Classification of URL Citations in Scholarly Papers for Promoting Utilization of Research Artifacts

Masaya Tsunokake*
Research and Development Group,
Hitachi, Ltd.
Tokyo, Japan
masaya.tsunokake@gmail.com

Shigeki Matsubara
Information and Communications,
Nagoya University
Nagoya, Japan
matubara@nagoya-u.jp

Abstract

Utilizing citations for research artifacts (e.g., dataset, software) in scholarly papers contributes to efficient expansion of research artifact repositories and various applications e.g., the search, recommendation, and evaluation of such artifacts. This study focuses on citations using URLs (URL citations) and aims to identify and analyze research artifact citations automatically. This paper addresses the classification task for each URL citation to identify (1) the role that the referenced resources play in research activities, (2) the type of referenced resources, and (3) the reason why the author cited the resources. This paper proposes the classification method using section titles and footnote texts as new input features. We extracted URL citations from international conference papers as experimental data. We performed 5-fold cross-validation using the data and computed the classification performance of our method. The results demonstrate that our method is effective in all tasks. An additional experiment demonstrates that using cited URLs as input features is also effective.

1 Introduction

Open science is an activity for promoting sharing and utilizing research artifacts[1]. One strategy to promote these activities is to provide repositories for research artifacts, and such repositories have been developed recently, e.g., Zenodo[2] and Mendeley Data[3]. In addition, national infrastructures for sharing research artifacts have been developed[4].

To develop a research artifact repository, it is required to register research artifacts and create their metadata[5]. Automating these processes can improve the efficiency of developing repositories and increase the number of research artifacts registered in the repositories. To this end, research artifact citations in scholarly papers can be utilized because scholarly papers citing research artifacts generally describe the name or usage of the artifacts. In addition, information about research artifacts not in existing metadata may be described in the scholarly papers (Kozawa et al., 2010; Singhal et al., 2014). Unlike citations for literature (**paper citations**), there are various ways to cite research artifacts. Therefore, automating the identification of the research artifact citations is not trivial task.

This study focuses on citations using URLs in scholarly papers (**URL citations**) and aims to identify and analyze research artifact citations. Figure 1 shows examples of URL citations. URL citations can refer to not only scholarly papers but also various resources, e.g., datasets, software, homepages, and articles. Therefore, an analysis of URL citations leads to the identification of research artifact citations. In addition, it can clarify the reality of URL citations performed informally.

This paper proposes a method to classify URL citations in scholarly papers according to the following viewpoints:

- The role of resources referenced by the URL in research activities
- The type of resources referenced by the URL

[*] This work was conducted while the first author was a master's student at the Nagoya University in Japan.

[1] This paper denotes research artifacts as digital objects collected, created, generated, or used in the course of research activities such as tools (e.g., software, program) and data (e.g., measurement data, test data). This definition is similar to that provided by the Association for Computing Machinery (Association for Computing Machinery, 2020).

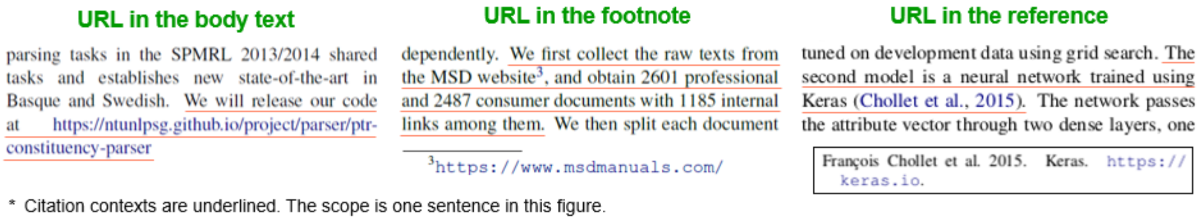
[2] https://zenodo.org/

[3] https://data.mendeley.com/

[4] e.g., Australian Research Data Commons (https://ardc.edu.au/), European Open Science Cloud (https://ec.europa.eu/), National Data Service (Towns et al., 2016) (http://www.nationaldataservice.org), NII Research Data Cloud (https://rcos.nii.ac.jp/en/service/)

[5] Information about research artifacts (e.g., name, creator, type, and usage)

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 8–19
Nov 20, 2022. ©2022 Association for Computational Linguistics



URL in the body text

parsing tasks in the SPMRL 2013/2014 shared tasks and establishes new state-of-the-art in Basque and Swedish. We will release our code at https://ntunlpsg.github.io/project/parser/ptr-constituency-parser

URL in the footnote

dependently. We first collect the raw texts from the MSD website[3], and obtain 2601 professional and 2487 consumer documents with 1185 internal links among them. We then split each document

[3] https://www.msdmanuals.com/

URL in the reference

tuned on development data using grid search. The second model is a neural network trained using Keras (Chollet et al., 2015). The network passes the attribute vector through two dense layers, one

François Chollet et al. 2015. Keras. https://keras.io.

* Citation contexts are underlined. The scope is one sentence in this figure.

Figure 1: Examples of URL citations

- The reason why the authors cited the resources

Zhao et al. (2019) proposed a classification method using multi-task learning for a similar task. That method inputs a word sequence surrounding the citation (**citation context**) into BERT (Devlin et al., 2019), and the representations obtained from the BERT are fed to a classification layer for each task. This paper proposes utilizing the section title and the footnote text used by the URL citation as new input features. Unlike the study by Zhao et al. (2019), this study newly addresses URL citations using reference sections.

2 Related Work

2.1 Citation Classification

Citations in scholarly papers have long been analyzed (Garfield, 1964; Moravcsik and Murugesan, 1975; Spiegel-Rösing, 1977; Cullars, 1990). Garfield (1964) discussed the reasons for citations and listed 15 motivations such as "Paying homage to pioneers" and "Providing background reading". Moravcsik and Murugesan (1975) investigated paper citations in the physics field to consider the appropriateness of using citations as measures of scientific accomplishments. The discussions in these studies were based on manual classification or the authors' insights. With the development of the computer science, some automatic classification methods have been proposed (Teufel et al., 2006; Abu-Jbara et al., 2013; Jurgens et al., 2018; Cohan et al., 2019). Teufel et al. (2006) proposed a method to classify paper citations based on the authors' reason for the citing (**citation function**) such as statement of weakness and comparison with other work. Jurgens et al. (2018) proposed a method to classify paper citations into six categories, e.g., "BACKGROUND," which means a cited paper provides relevant information, "USES," which means a citing paper uses data or methods in the cited paper.

Ding et al. (2014) summarized such approaches for analyzing citations based on their content as Content-based Citation Analysis (CCA). The CCA has been applied to various tasks, e.g., summarizing papers, recommending citations, and improving metrics for papers (Ding et al., 2014). In addition, some studies have demonstrated that considering the citation functions contributes to the analysis of academic trends (Abu-Jbara et al., 2013; Jurgens et al., 2018), automatic generation of citation sentence (Ge et al., 2021), and prediction of the number of citations (Jurgens et al., 2018).

2.2 Research Artifact Citations

Recently, research artifacts, e.g., datasets and software, have been cited increasingly in scholarly papers. Then, there is a growing movement to establish formal rules for data and software citations, as FORCE11 has declared "Data Citation Principles" (Data Citation Synthesis Group, 2014) and "Software Citation Principles" (Smith et al., 2016). However, widespread adoption of this practice among researcher is a long way off. Howison and Bullard (2016) have demonstrated that there were many informal citations in biology papers. One strategy for automatic identification of the informal citations is to identify research artifact mentions in the body text (Krüger and Schindler, 2020). Some studies address the identification of dataset names (Singhal and Srivastava, 2013; Prasad et al., 2019; Ikeda et al., 2020) or software names (Li and Yan, 2018; Schindler et al., 2020; Du et al., 2021). Another approach finds research artifact citations from explicit citations. Ikoma and Matsubara (2020) attempted to identify bibliographic information referring to linguistic resources (e.g., corpus, lexicon) from reference sections. Since some research artifact citations uses URL, identification of URLs referring to research artifacts in scholarly papers has also been studied (Tsunokake and Matsubara, 2021).

Table 1: List of resource roles and resource types

Resource role	Resource type	description
Material	Dataset	corpus, image sets, etc.
	Knowledge	lexicon, knowledge graph, etc.
	DataSource	source data for the Dataset/Knowledge
Method	Tool	toolkit, software, system, etc.
	Code	codebase, library, API, etc.
Supplement	Document	documents on the Web (e.g., specifications, guidelines)
	Paper	scholarly papers
	Media	games, music, videos, etc.
	Website	other resources on the Web (e.g., services, homepages)
Mixed	Mixed	citations referring to multiple resources

2.3 Classification of URL Citations

With the increase in URL citations in scholarly papers, some studies have attempted to utilize resources referenced by URLs. For example, Yamamoto and Takagi (2007) extracted URLs from papers in the life science domain to develop a system for searching online resources. Parmar et al. (2020) extracted URLs from papers and constructed a portal of academic information (e.g., metadata about papers and authors) in the natural language processing field. Nanba (2018) proposed a method to extract a URL in scholarly papers and the tag representing the URL based on their distributed representations obtained from scholarly papers. There is a study addressing the classification of URL citations. Zhao et al. (2019) applied the CCA (Section 2.1) to URL citations in order to construct search/recommendation systems and knowledge graphs for scientific resources. They proposed a classification method to determine the roles of resources referenced by URLs in scholarly papers and the authors' purposes of URL citations based on the citation contexts.

In this study, our goal is to generate metadata for research data automatically. The resource roles defined by Zhao et al. (2019) contain the "Material" and "Method" roles, and we consider that citations corresponding to these labels are equivalent to research artifact citations. Thus, research artifact citations can be identified by solving this classification task. In addition, information on how referenced resources can be used in research activities can be obtained. URL citations are less identifiable and more ambiguous than paper citations whose bibliographic information are regularly listed in the reference sections. Thus, it would be meaningful for the academic community to realize automatic analysis for URL citations.

3 Task Definition

This study addresses three classification tasks determining the followings for each URL citation.

1. The role that resources play in the context of research activities (**resource role**)
2. The type of resources (**resource type**)
3. The reason why resources were cited (**citation function**)

Zhao et al. (2019) defined two levels of resource roles consisting of general resource roles and fine-grained resource roles. The fine-grained resource roles can be regarded as the type of referenced resources; thus, this study redefines them as resource types. Even if the same URL is cited in distinct papers, the resources that one author refers to may differ from those referenced by other authors. Therefore, in any of the classification tasks, it is necessary to infer from the citation contexts. Our target URL citations are as follows:

1. The URL is described in the body text
2. The URL is described in a footnote
3. The URL is described in the bibliographic references, and the corresponding citation anchor is described in the body text

Figure 1 shows an example of each case. If the URL is described in the footnote (case 2) or the reference (case 3), the corresponding surrounding sentences in the body text are the citation contexts. Note that Zhao et al. (2019) only targeted the case 1 and 2. However, when citing online resources, the resources can be cited as a reference, and the corresponding URL is described in

Table 2: List of citation functions

Citation function	Description
Use	Used in the citing paper's research.
Produce	First produced or released by the citing paper's research.
Compare	Compared with other resources.
Extend	Used in the citing paper's research but are improved, upgraded, or changed to work for other problems in the course of the research.
Introduce	The resources or the related information (e.g., background, applications) are introduced.
Other	The URL citation does not belong to the above 5 categories.

the bibliographic information. It is sometimes recommended that online resources are cited as references; thus, classifying URL citations via reference sections is required.

Table 1 presents the labels for the resource role/type. Since each resource type determines the role it can play, there is a correspondence between the resource roles and resource types. While the labels are based on the setting of Zhao et al. (2019), this study applies some alterations with a view to generating metadata for research artifacts. If the extracted information is used for metadata, the resource types are required to be somewhat fine-grained. However, the only resource type corresponding to "Material" (one of the resource roles) is "Data" in the study by Zhao et al. (2019). Therefore, this paper defines "Dataset," "Knowledge," and "DataSource" as more detailed types. In addition, since this paper considers resource types as types of cited digital objects, labels referring to something conceptual rather than actual digital objects (e.g., "algorithm") were dropped from the resource types. In some URL citations, multiple resources may be referenced simultaneously. Since the URL citation cannot be classified into a specific label in this case, "Mixed" is defined as one of the resource roles/types. The "Mixed" label was defined in some studies addressing citation classifications to consider cases where multiple labels are mixed (Cullars, 1990; Ge et al., 2021). Table 2 presents the labels for citation functions. This is the same setting as in Zhao et al. (2019).

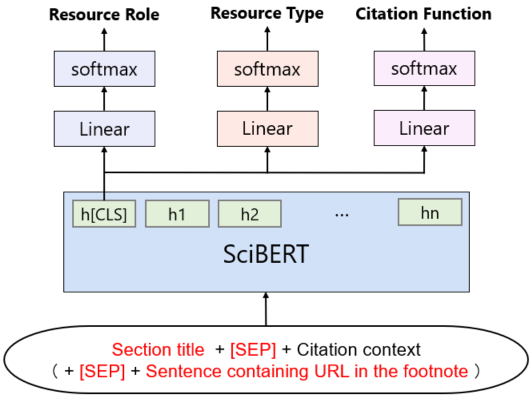


Figure 2: Architecture of our method

4 Method

Zhao et al. (2019) proposed a framework called SciResCLF for a similar classification task. Since there is a certain correlation between labels for each task, they employed multi-task learning in SciResCLF. The SciResCLF employs BERT (Devlin et al., 2019) as the encoder for citation contexts. In the SciResCLF, the citation contexts are taken into the BERT, and the obtained embeddings for the "[CLS]" token are taken into a classification layer for each task. In fine-tuning, the model parameters are optimized based on the weighted sum of the cross-entropy of each task. The SciResCLF only uses citation contexts as the input features. Based on SciResCLF, this paper proposes a classification method using section titles as global context information, and footnote texts used for URL citations.

Jurgens et al. (2018) demonstrated that there was a certain relationship between where paper citations appear in the narrative structure of a citing paper and their citation functions. In our task as well, information about the narrative structure may be useful. For example, scholarly papers may tend to cite used software, code, or datasets by providing the corresponding URL in the sections describing experiments. On the other hand, URLs described in introductory sections may tend to refer to supplements related to the background (e.g., news, web-service). Thus, our method uses the section titles where URL citations appear as input features. In addition, some URL citations do not explain the referenced resources in the body text but explain the resources in the footnotes which the corresponding URL are described in. Therefore, the footnote texts used for URL citations are

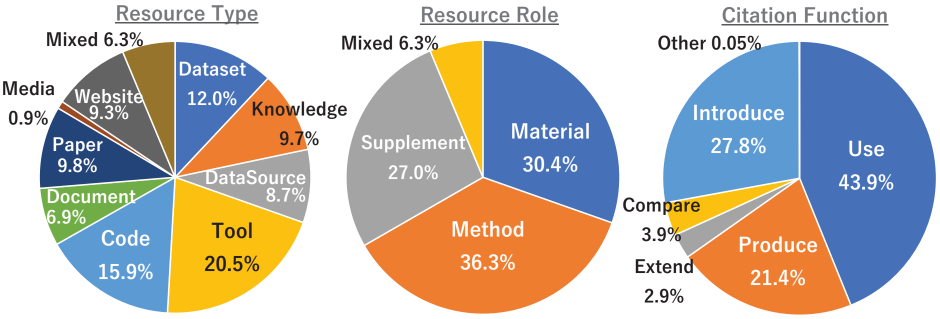
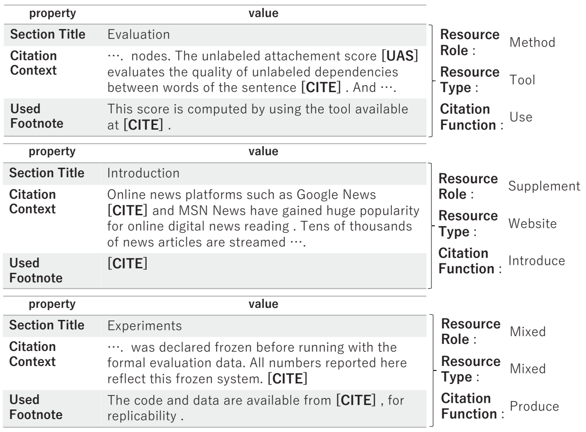


Figure 3: Ratio of each label in the created dataset



property	value
Section Title	Evaluation
Citation Context	···. nodes. The unlabeled attachement score **[UAS]** evaluates the quality of unlabeled dependencies between words of the sentence **[CITE]** . And ···.
Used Footnote	This score is computed by using the tool available at **[CITE]** .

Resource Role : Method
Resource Type : Tool
Citation Function : Use

property	value
Section Title	Introduction
Citation Context	Online news platforms such as Google News **[CITE]** and MSN News have gained huge popularity for online digital news reading . Tens of thousands of news articles are streamed ···.
Used Footnote	**[CITE]**

Resource Role : Supplement
Resource Type : Website
Citation Function : Introduce

property	value
Section Title	Experiments
Citation Context	···. was declared frozen before running with the formal evaluation data. All numbers reported here reflect this frozen system. **[CITE]**
Used Footnote	The code and data are available from **[CITE]** , for replicability .

Resource Role : Mixed
Resource Type : Mixed
Citation Function : Produce

Figure 4: Examples of the created datset

expected to be an effective feature in this classification task.

Figure 2 shows the architecture of our method. In our method, the input for each URL citation is created by concatenating the section title where the citation appears, the citation context, and the footnote sentence containing the cited URL with "[SEP]" [6]. This model is trained in a multi-task learning framework. Thus, the model is optimized based on cross-entropy losses about predicting the resource roles, resource types, and citation functions.

5 Experiment

5.1 Dataset

There was no dataset for the classification of URL citations with the corresponding section titles, footnotes, and this paper's classification labels. Therefore, we created an experimental dataset. We collected the scholarly papers as the source of URL citations from the ACL Anthology[7]. The papers were collected from the proceedings of ACL/EMNLP/NAACL 2000–2021. We collected a total of 15,761 papers. The PDF of each paper was converted to text by PDFNLT-1.0[8](Abekawa and Aizawa, 2016). The URLs[9], footnote numbers in the body text that refer to the footnotes, and the citation anchors referring to bibliographic information in the reference section were detected for each paper. The citation anchors were detected by regular expressions[10] based on those described by Gosangi et al. (2021). They are compatible with both the Harvard and Vancouver referencing style. Using the detected results, paragraphs where the URL citations appeared were extracted as the citation contexts of the citations. We evaluated the performance of identifying the location of URL citations using 65 randomly selected papers. As a result, precision was 0.995 (199/200), and recall was 0.948 (199/210).

The extracted URL citations were annotated by an expert in the natural language processing field. Before the annotation, a part of URL removed mechanically, such as URL citations whose citation context had only a few words and the URLs attached as an auxiliary to the bibliographic information in the paper citation. The annotator was instructed to refer to the label definitions and examples of annotated URL citations before the work and could refer to them anytime during the work.

The created dataset contained 2,037 URL citations from 652 papers. Figure 3 shows the distribution of labels. Although the distribution of labels is skewed, there is a certain balance in the ratio

[6] If a URL citation does not use a footnote, or a footnote used by the URL citation only contains the URL, the second "[SEP]" and the footnote sentence are not concatenated.

[7] https://aclanthology.org/

[8] https://github.com/KMCS-NII/PDFNLT-1.0

[9] Strings beginning with either "http://," "https://," or "ftp://" were identified as URLs.

[10] Details are described in the appendix.

Table 3: Evaluation results for each task

Method	Resource role				Resource type				Citation function			
	ACC.	P.	R.	F1	ACC.	P.	R.	F1	ACC.	P.	R.	F1
Baseline	0.653	0.682	0.598	0.621	0.430	0.450	0.348	0.357	0.663	0.563	0.429	0.437
Our method	†0.694	0.711	0.653	†0.670	0.459	0.452	†0.385	0.391	†0.703	0.571	0.438	0.448

Table 4: Cases where baseline failed to predict but our method correctly predicts

Inputs of our method	True	Prediction	
		Baseline	Our method
Introduction [SEP] Recently, a new benchmark MRC dataset called Natural Questions [CITE] (NQ) has presented a substantially greater challenge for the existing MRC models. [SEP] **NQ provides some visual examples of the data [CITE] .**	Supplement	Material	Supplement
Data [SEP] WikiSum consist of Wikipedia articles each of which are associated with a set of reference documents. [CITE] [SEP] **We take the processed Wikipedia articles from [CITE] released on April 25th 2018.**	DataSource	Knowledge	DataSource
Conclusion [SEP] We have described a dependency-based system [CITE] for semantic role labeling of English in the PropBank framework. [SEP] **Our system is freely available for download at [CITE] .**	Produce	Use	Produce

of corresponding resource types for each resource role. For example, "Dataset," "Knowledge," and "DataSource" defined by this paper correspond to "Material," and there is not much difference in their ratios. In the dataset, the rate of URL citations using footnotes is 0.725, the rate of URL citations using the reference sections is 0.170, and the rate of URL citations in the body texts is 0.105. Figure 4 shows the examples of dataset[11]. Another researcher in the natural language processing field annotated 100 citations in the dataset as with the original annotator. As a result, the Cohen's kappa of the resource roles, resource types and citation functions were 0.644, 0.456, and 0.615, respectively.

5.2 Experimental Setup

A 5-fold cross-validation was performed using the created dataset. Randomly 20% of the dataset was used as the development set, and the rest was used as the training or test set by dividing it into 5 parts. Thus, the training set contained 1,304 samples, the development set contained 407 samples, and the test set contained 326 samples for each split.

The SciResCLF proposed by Zhao et al. (2019) was employed as the baseline, and both the baseline and our method were evaluated by the 5-fold cross-validation. Both methods used SciBERT (Beltagy et al., 2019) as the encoder for the input features. In our method, the section title used as the input was the top-level heading, and the footnote text was the 1 sentence containing the URL in the footnote used by the URL citation. The loss function was the sum of the cross-entropy losses for each task. The optimization function was Adam (Kingma and Ba, 2014)[12].

To assess the classification performance, both methods were evaluated by accuracy and the macro-averaged F1. Accuracy tends to be more dominated by the results of frequent classes than the F1 averaging the result of each class.

5.3 Experimental Results

Table 3 presents the average of evaluation result for each split[13]. Our method outperformed the baseline in all metrics on all tasks. Table 4 presents cases where the baseline failed to predict but our method predicted correctly. Note that the section titles before the first [SEP]s and the footnotes after the second [SEP]s were not taken into the baseline. In the first row, the footnote indicates that the referenced resource is not a dataset but an example visualization as a supplemental resource. In the second row, the footnote indicates that the referenced resource is not a created dataset but the source used for creation of the dataset. These footnotes contributed to the prediction of our method. In the third row, using section titles and footnotes, our model may catch a tendency that authors are

[11] In the same way as Zhao et al. (2019), we replaced the citation locations and cited URLs with "[CITE]."

[12] Details are described in the appendix.

[13] Daggers (†) mean that there was a significant difference between the baseline and our method by the paired t-test. The significance level was 0.05. ACC., P., and R. are accuracy, macro-averaged precision, and macro-averaged recall, respectively.

Table 5: Results of ablation study and additional experiment

Method	**Resource role**		**Resource type**		**Citation function**	
	ACC.	F1	ACC.	F1	ACC.	F1
Baseline	0.653	0.621	0.430	0.357	0.663	0.437
Our method	0.694	0.670	0.459	0.391	0.703	0.448
- w/o section title	(−) 0.674	(−) 0.653	(+) 0.481	(+) 0.409	(−) 0.701	(+) 0.451
- w/o footnote	† (−) 0.663	(−) 0.626	(−) 0.423	† (−) 0.348	(−) 0.688	(+) 0.457
- w/ URL	(−) 0.679	(−) 0.631	(+) 0.501	† (+) 0.437	(+) 0.715	(+) 0.454

Table 6: F1-score for each label

Resource role	**F1-score**		**Resource type**	**F1-score**		**Citation function**	**F1-score**	
	Baseline	Our method		Baseline	Our method		Baseline	Our method
Material	0.659	0.680	Dataset	0.466	0.448	Use	0.715	0.751
Method	0.688	0.728	Knowledge	0.217	0.243	Produce	0.615	0.729
Supplement	0.605	0.686	DataSource	0.513	0.514	Compare	0.230	0.172
Mixed	0.532	0.585	Tool	0.494	0.533	Extend	0.029	0.000
			Code	0.410	0.476	Introduce	0.671	0.667
			Document	0.179	0.204	Other	0.000	0.000
			Paper	0.493	0.606			
			Media	0.000	0.000			
			Website	0.343	0.348			
			Mixed	0.459	0.536			

likely to add a URL referring to their own created resources at the end of the paper.

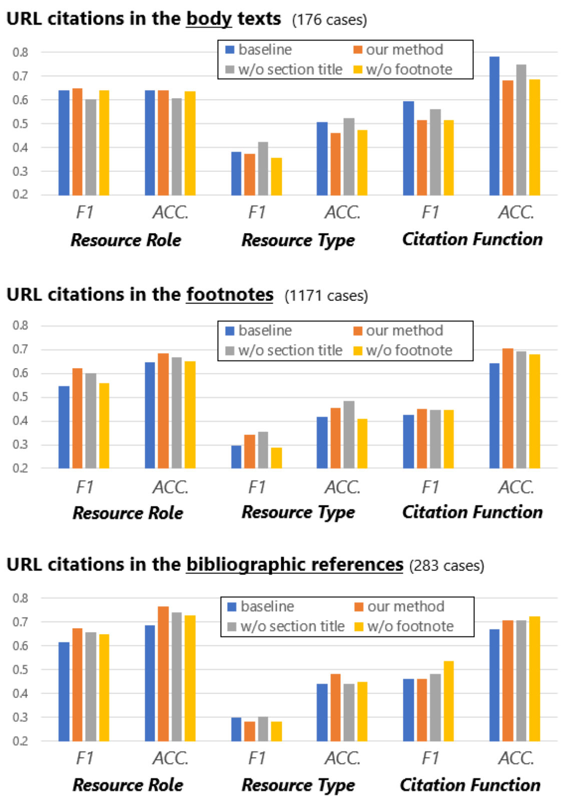
5.4 Disscussion

Table 5 presents the result of an ablation study when one of the proposed input features was excluded[14]. As to resource roles, excluding section titles or footnotes from our method degraded the classification performance, which indicates that both features are effective. Similarly, excluding footnotes from our method degraded the performance in resource types, and thus using footnotes is effective. In contrast, excluding section titles from our method improved the performance in the classification of resource types. In addition, “w/o footnote”, which added only the section title to the input features of baseline, was inferior to the baseline. These results demonstrate that using the section titles has a negative effect on the classification of resource types. As to citation functions, excluding one feature from our method improved the performance of the F1. However, the both F1s of “w/o section title” and “ w/o footnote” were higher than the baseline. Independently, each of these features was effective in the classification of citation functions; however, combining them or the ways by which they were combined resulted in a negative effect.

Table 6 presents the F1 for each label[15] for the baseline and our method. In the classification of resource roles and types, our method outperformed the baseline for all labels except for “Dataset.” There were some cases where our method misclassified citations whose resource type was “Dataset” as “DataSource” because the footnote included the text “from [CITE]” (e.g., “The corpus can be downloaded from [CITE]”). While the ratio of “DataSource” in all predicted labels by our method was 0.089, that for cases where the input text included “from [CITE]” is 0.276. However, it was effective in the second row of Table 4. In this case, the resource type is “DataSource” because the URL refers to not the WikiSum dataset but its source articles. Ideally, the ability to identify the target of the citation and infer the relationship between it and the surrounding words indicating research artifacts is required.

[14] The results of the ablation study that were lower and higher than that of our method are marked with “(−)” and “(+),” respectively. Daggers (†) mean that there was a significant difference compared to our method by the paired t-test. The significance level was 0.05.

[15] It is the average for splits in the cross-validation.



URL citations in the body texts (176 cases)

URL citations in the footnotes (1171 cases)

URL citations in the bibliographic references (283 cases)

Figure 5: Evaluation results for each URL citation's type based on how the URL are described

As described in Section 3, the URL citations are divided into three types based on how the URLs are described. Figure 5 shows the evaluation results for each type of URL citation. The block of each bin shows the results of baseline, our method, our method without section titles, and our method without footnotes, from left to right. An overview of Figure 5 shows that the valid features depend on the combination of task and how to cite (i.e., the type of the URL citation). Our method basically outperformed the baseline when classifying URL citations in the footnotes and the bibliographic references; however, it tended to exhibit inferior performance compared to the baseline when classifying URL citations in the body. As to the classification of citation functions for URL citations in the body texts, our method was inferior to "w/o section title." In addition, "w/o footnote," which adds section titles to the baseline, was also inferior to the baseline. These results indicate that section titles have a negative effect on the classification of citation functions for URL citations in the body texts. However, there is a different trend for URL citations in footnotes and bibliographic references, which indicates that section titles are effective in the classification of citation functions for both types of URL citations. Different approaches depending on tasks or types of URL citations are required.

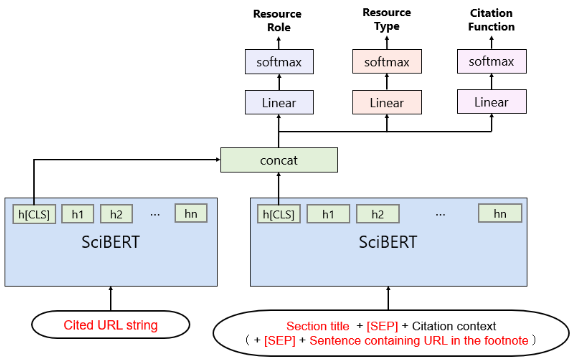


Figure 6: Architecture of classification model using the cited URL strings

URL citations whose URLs are described in the references do not use footnotes. However, in the classification of such URL citations, "w/o section title" feeding cited footnotes into inputs tends to outperform the baseline using only citation contexts. Interestingly, training footnote texts is also useful for some citations that do not use footnotes.

5.5 Improving Classification Performance for Resource Types

While our method was effective for the classification of resource types, the F1-score was lower than that of other tasks. Thus, we extended our method by utilizing the substrings of URLs as input feature for classification. For some cases, the type of resources can be inferred from the domain or directory name constituting the URL. For example, it can be inferred from "data" and "tweets," that the URL "http://trec.nist.gov/data/tweets/" points to data related to tweets. Each substring constituting the URLs can contain information about resources on the website. In this approach, for each URL citation, the string of the cited URL is tokenized and encoded by SciBERT. The hidden layer corresponding to the "[CLS]" token is employed as the embedding for the entire substring sequence. Figure 6 shows the classification model used in this approach. The model concatenates the embedding of the cited URL string and the embedding of context information in the citing paper, and the obtained vector is used as the input feature for the linear layer of each task.

Table 5 presents the experimental results[16] The method utilizing the cited URLs (the row of "w/ URL") improved the classification performance of resource types. In addition, the classification of citation functions was also improved.

6 Limitation

In this paper, experimental data was constructed from one domain. Since paper styles, including structure of sections, how to use footnotes, and which type of URL citation the authors prefer may differ according to the domain, constructing experimental data from other domains and verifying our method on the data remain as the future works.

This paper defined the "Mixed" label for multiple resource citations. If citations are classified to the "Mixed" label, the resource roles and types can not be identified. Therefore, in practice, additional classification is required. Otherwise, it can be considered to employ the multi-label classification as with Zhang et al. (2022)'s study which applied the multi-label formulation to the classification of citation function. In that case, it is necessary to discuss how citations using ambiguous terms as referenced resources should be regarded (e.g., "All code and resources are available at [CITE].").

This paper addressed the automatic classification of URL citation to generate metadata of research artifacts. It contributes to the efficient expansion of research artifact repository, enrichment of the existing repositories, and automatic analysis of research artifact citations. However, resources cited by URLs tend to become unreachable within some years (Zeng et al., 2019). To promote utilization of research artifacts cited by URLs, establishing systems and platforms to preserve the artifacts and maintaining them are also required. As for the maintaining, the automatic predicting the longevity of research artifacts cited by URLs (Acuna et al., 2022) might be useful.

7 Conclusion

This paper addressed the classification task of identifying the resource role, resource type, and citation function, for each URL citation in scholarly papers. This paper proposed the classification method using not only citation contexts but also section titles and footnote texts as input features. Our method was evaluated experimentally and the results demonstrated the effectiveness of our method on all tasks. However, the effective features differ depending on the task and how the URL is cited. When classifying resource types, an approach that obtains and uses an embedding for the URL string used for the citation was effective.

Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 21H03773) of JSPS.

References

Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly paper browsing system based on PDF restructuring and text annotation. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations (COLING 2016)*, pages 136-140, Osaka, Japan. The COLING 2016 Organizing Committee.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2013)*, pages 596-606, Atlanta, Georgia. Association for Computational Linguistics.

Daniel E. Acuna, Jian Jian, Tong Zeng, Lizhen Liang, and Han Zhuang. 2022. Predicting the longevity of resources shared in scientific publications. *ArXiv preprint*, arXiv:2203.12800.

Association for Computing Machinery. 2020. Artifact review and badging version 1.1. (accessed 26 October 2022)

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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-IJCNLP2019)*, pages 3615-3620, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2019) Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

[16] If there was a significant difference compared to our method by the paired t-test, daggers (†) are assigned. The significance level was 0.05.

John Cullars. 1990. Citation characteristics of Italian and Spanish literary monographs. *The Library Quarterly*, 60(4):337-356.

Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. FORCE11, San Diego, CA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2019) Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the association for information science and technology*, 65(9):1820-1833.

Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72(7):870-884.

Eugene Garfield. 1964. Can citation indexing be automated?. In *Proceedings of Statistical Association Methods for Mechanized Documentation, Symposium Proceedings*, pages 189–192, Washington, USA.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge- and content-based framework for citing sentence generation. In *Proceedings of The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL2021) (Volume 1: Long Papers)*, pages 1466-1478, Online. Association for Computational Linguistics.

Rakesh Gosangi, Ravneet Arora, Mohsen Gheisarieha, Debanjan Mahata, and Haimin Zhang. 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 (NAACL-HLT2021)"*, pages 4539–4545, Online. Association for Computational Linguistics.

James Howison and Julia Bullard. 2016. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9):2137-2155.

Daisuke Ikeda, Kota Nagamizo, and Yuta Taniguchi. 2020. Automatic identification of dataset names in scholarly articles of various disciplines. *International Journal of Institutional Research and Management*, 4(1):17-30.

Tomoki Ikoma and Shigeki Matsubara. 2020. Identification of research data references based on citation contexts In *Proceedings of The 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)*, pages 149-156, Kyoto, Japan. Springer.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ArXiv preprint*, arXiv.1412.6980

Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, and Shigeki Matsubara. 2010. Collection of usage information for language resources from academic articles. In *Proceedings of The 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1227-1232, Valletta, Malta. European Language Resources Association (ELRA).

Frank Krüger and David Schindler. 2020. A literature review on methods for the extraction of usage statements of software and data. *Computing in Science Engineering*, 22(1):26-38.

Kai Li and Erjia Yan. 2018. Co-mention network of R packages: Scientific impact and clustering structure. *Journal of Informetrics*, 12(1):87-100.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86-92.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. Doccano: Text annotation tool for human. Software available from `https://github.com/doccano/doccano`.

Hidetsugu Nanba. 2018. Construction of an academic resource repository. In *Proceedings of the Toward Effective Support for Academic Information Search Workshop*, pages 8-14, Hamilton, New Zealand. Kyushu University.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319-327, Florence, Italy. Association for Computational Linguistics.

Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. NLPExplorer: Exploring the universe of NLP papers. *Advances in Information Retrieval*, 12036:476-480.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems 32 (NIPS19)*, pages 8024-8035, Vancouver, Canada. Curran Associates Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825-2830.

Animesh Prasad, Chenglei Si, and Min-Yen Kan. 2019. Dataset mention extraction and classification. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications (ESSP)*, pages 31-36, Minnesota, USA. Association for Computational Linguistics.

David Schindler, Benjamin Zapilko, and Frank Krüger. 2020. Investigating software usage in the social sciences: A knowledge graph approach In *Proceedings of the 17th European Semantic Web Conference Semantic Web (The Semantic Web)*, pages 271-286, Crete, Greece. Springer.

Ayush Singhal, Ravindra Kasturi, and Jaideep Srivastava. 2014. DataGopher: Context-based search for research datasets. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pages 749–756, California, USA. Institute of Electrical and Electronics Engineers

Ayush Singhal and Jaideep Srivastava. 2013. Data extract: Mining context from the web for dataset extraction. *International Journal of Machine Learning and Computing*, 3(2):219-223.

Arfon M. Smith, Daniel S. Katz, and Kyle E. Niemeyer. 2016. Software citation principles. *PeerJ Computer Science*, 2:e86.

Ina Spiegel-Rösing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7(1):97-113.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing (EMNLP2006)*, pages 103-110, Sydney, Australia. Association for Computational Linguistics.

John Towns, Christine Kirkpatrick, Kenton McHenry, and Kandace Turner. 2016. Towards a U.S. national data service - inaugural report. The National Date Service, Illinois, USA. (accessed 26 October 2022)

Masaya Tsunokake, Shigeki Matsubara. 2021. Classification of URLs citing research artifacts in scholarly documents based on distributed representations. In *Proceedings of the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2021) co-located with JCDL2021*, pages 20–25, Online. CEUR-WS Team.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP2020): System Demonstrations*, pages 38-45, Online. Association for Computational Linguistics.

Yasunori Yamamoto and Toshihisa Takagi. 2007. OReFiL: An online resource finder for life sciences. BMC bioinformatics, 8(1):1-8.

Tong Zeng, Alain Shema, and Daniel E. Acuna. 2019. Dead science: Most resources linked in biomedical articles disappear in eight years. In *Proceedings of the 14th International Conference on Information - iConference 2019*, pages 170-176, Washington, USA. Springer.

Yang Zhang, Yufei Wang, Quan Z. Sheng, Adnan Mahmood, Wei Emma Zhang, and Rongying Zhao. 2022. TDM-CFC: Towards document-level multi-label citation function classification. In *Proceedings of the 22nd International Conference on Web Information Systems Engineering - WISE 2021*, pages 363-376, VIC, Australia. Springer.

He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*, pages 5206-5215, Hong Kong, China. Association for Computational Linguistics.

A Supplement for Creating Dataset

Citation anchors in scholarly papers were detected by regular expressions based on those used by Gosangi et al. (2021). The following code shows the regular expressions for the Harvard referencing style, which was implemented by Python.

```
AUTHOR_NAME = r"([A-Z][\w\-’]*?)"
ETAL = "(et ?als?\.?)"
AUTHOR_SECTION = AUTHOR_NAME +
                 r"(?: (?:(?:and|&) (?:de )?" +
                 AUTHOR_NAME + ’|’ + ETAL + ’))?’
YEAR = r"((?:18|19|20)[0-9]{2}[a-z]?)"
PAGE = r"(?:, (?:pages|pp?\.?) \d+(?:-\d+)?)"
YP = f"{YEAR}{PAGE}?"

LEFT_BRACKET = r"[\(\[]"
RIGHT_BRACKET = r"[\)\]]"

CITET = f"{AUTHOR_SECTION} {LEFT_BRACKET}{YP}
          {RIGHT_BRACKET}"
CITEP_SINGLE = f"{LEFT_BRACKET}{AUTHOR_SECTION}
                 , {YP}{RIGHT_BRACKET}"
CITEP_MULTI_BEGIN = f"{LEFT_BRACKET}{AUTHOR_SECTION}
                      , {YP};"
CITEP_MULTI_INSIDE = f"(?<=; ){AUTHOR_SECTION}
                      , {YP};"
CITEP_MULTI_END = f"(?<=; ){AUTHOR_SECTION}
                   , {YP}{RIGHT_BRACKET}"

CITATION_ANCHOR = f"(?:{CITET}|{CITEP_SINGLE}|
                     {CITEP_MULTI_BEGIN}|
                     {CITEP_MULTI_INSIDE}|
                     {CITEP_MULTI_END})"
```

In addition, the following code is for the Vancouver referencing style.

```
NUMBER = r"(?:([1-9]\d*)(?:(-[1-9]\d*))?)"
CITATION = f"\\[{NUMBER}(?:, ?{NUMBER})*?\\]"
```

The annotation environment was implemented by Doccano (Nakayama et al., 2018).

B Experimental Setup

In the experiment, the following procedure was performed in each split of the 5-fold cross-validation. For each candidate of hyperparameters, the classification model was trained for up to 50 epochs. Note that training was terminated if the minimum loss for the development set could not be updated within 10 epochs. Then, for each classification task, the trained model with the best classification performance[17] for the development set was applied to the test set to evaluate the method. In this evaluation, accuracy (i.e, micro-averaged F1) and macro-averaged F1 were computed from the classification results obtained on the test set in each split.

The following hyperparameters were verified in the experiment.

- Batch size: 16, 32, 64
- Learning rate: 1.0e-4, 5.0e-5, 1.0e-5, 5.0e-6
- Scope of citation contexts[18]: 1 sentence, 3 sentences (citing sentence and 1 sentence before and after the citing sentence), 5 sentences (citing sentence and 2 sentences before and after the citing sentence)
- Dropout rates: 0.0, 0.3, 0.6
- Maximum sequence length of inputs: 256

The weight of each task in the loss was set equally at 1.0.

In addition, scikit-learn[19] (Pedregosa et al., 2011), PyTorch[20] (Paszke et al., 2019), and Hugging Face’s transformers library[21] (Wolf et al., 2020) were used to implement the experiment. Sentence segmentation was performed by ScispaCy[22] (Neumann et al., 2019).

[17] macro-averaged F1-score

[18] A paragraph is one of the semantic units. Therefore, in this study, the scope of the citation context was limited to the paragraph containing the URL citation even when the employed scope included sentences before and after the citing sentence.

[19] https://scikit-learn.org/stable/

[20] https://pytorch.org/docs/1.8.1/

[21] https://github.com/huggingface/transformer

[22] https://github.com/allenai/scispacy