A Context-based Framework for Resource Citation Classification in Scientific Literatures

He Zhao*
Beijing Institute of Technology, School of Computer Science
Beijing, China
zhaohe1995@outlook.com

Zhunchen Luo*
Information Research Center of Military Science, PLA Academy of Military Science
Beijing, China
zhunchenluo@gmail.com

Chong Feng†
Beijing Institute of Technology, School of Computer Science
Beijing, China
fengchong@bit.edu.cn

Yuming Ye
Information Research Center of Military Science, PLA Academy of Military Science
Beijing, China
yuming-ye@163.com

ABSTRACT

In this paper, we introduce the task of *resource citation classification* for scientific literature using a context-based framework. This task is to analyze the purpose of citing an on-line resource in scientific text by modeling the role and function of each resource citation. It can be incorporated into resource indexing and recommendation systems to help better understand and classify on-line resources in scientific literature. We propose a new annotation scheme for this task and develop a dataset of 3,088 manually annotated resource citations. We adopt a neural-based model to build the classifiers and apply them on the large ARC dataset to examine the revolution of scientific resources from trends in their function over time.

CCS CONCEPTS

• **Information systems** → **Entity relationship models**; **Content analysis and feature selection**; *Semi-structured data*; *Extraction, transformation and loading*; • **Software and its engineering** → *Entity relationship modeling*;

KEYWORDS

Scientific literature mining, scientific resource classification

ACM Reference Format:
He Zhao, Zhunchen Luo, Chong Feng, and Yuming Ye. 2019. A Context-based Framework for Resource Citation Classification in Scientific Literatures. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331348

*indicates equal contribution.
†Corresponding author.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331348

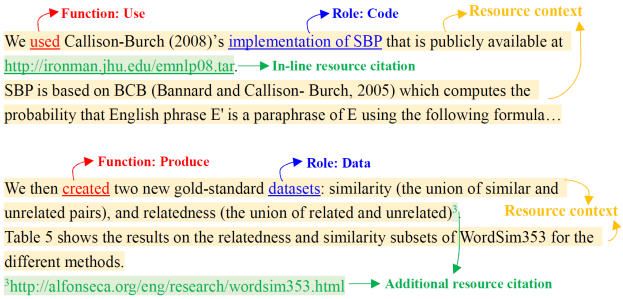

Figure 1: Examples for the two types of resource citations in scientific literature.

1 INTRODUCTION

With the number of scientific publications growing dramatically, numerous on-line resources are mentioned, released and used within the scientific literature. Tracing and modeling these resources such as software, tools and datasets can greatly help researchers by developing scientific resource indexing and recommendation systems or constructing scientific resource knowledge graphs. In this paper, we address a new task of *resource citation classification* for scientific literature. We develop a framework to model the scientific resources by classifying their roles and functions based on the resource citation context. Google has launched a new search engine[1] in 2018 to help scientists find the datasets they need, whereas the datasets can only be matched with their exact names. To help develop more powerful scientific resource searching systems, identifying the resource role will enrich the repository while classifying the resource function is crucial for more flexible searching queries and more exhaustive results. Different from previous works on paper citations [2, 3, 5, 10], our work specially focuses on the on-line resources in scientific text, which is not as well studied as paper citations. To the best of our knowledge, there exists no prior work which tries to classify the role and function on such a fine-grained level in scientific full-text yet.

We first give some definitions for this task. A *resource citation* is defined as a hyperlink the author mentions in the text, which

[1]https://toolbox.google.com/datasetsearch

Table 1: Definitions and examples for the *resource role* categories and *resource function* categories.

Role	*Definition*	*Example*
Tool	The tool consists of toolkits, software, systems or projects	We use a CRF++ based POS tagger for Hi, which is freely available from <CITE>. For En, we use the Twitter POS tagger.
Code	The code consists of codebases, libraries or implementations.	The SVM computations are performed using the freely available Spider Matlab machine learning package available at <CITE>.
Data	The data consists of datasets, databases or corpus.	The selection of C and the RKHS has been done as indicated in <CITE> for Adult and Web data sets and in <CITE> for Banana, Diabetes and Splice data sets.
Website	The website consists of homepages, services, on-line platforms or interfaces.	Answers and Live Search QnA <CITE>, have been rapidly gaining popularity among Web users interested in sharing information online.
Algorithm	The algorithm consists of methods, models or solutions.	The model uses a maximum entropy learner <CITE>, training one binary classifier per sense.
Document	The document consists of supplements, tutorials, specifications or guidelines.	Unfortunately there exists substantial disagreement regarding the interpretation of existing approaches see <CITE>.
Media	The media consists of games, musics or videos.	Formulating MI using Renyi entropy, and Gaussian Example video clips can be viewed at <CITE>.
License	The license provides access to and details of the used licenses.	The Creative Commons Public Domain Dedication waiver (<CITE>) applies to the data made available in this article, unless otherwise stated.
Paper	The paper is a short/long conference paper taken from sites.	In this section, we propose a use case of application of Linked Aoki-Kinoshita et al. Journal of Biomedical Semantics 2014, 6:3 Page 4 of 13 <CITE>.
Function	***Definition***	***Example***
Use	The resource is used in this paper's work.	We use a local search engine, <CITE>, which accepts the SearchTerm and LocationTerm as two query fields and returns the search results from a business listings database.
Produce	The resource is first produced or released by this paper's work.	Finally, to aid other AAC researchers, we have publicly released our crowdsourced AAC collection, word lists and best-performing language models <CITE>.
Introduce	The background, characteristic or applications of the resource is introduced in the context sentences.	The central and most widely-used resource in the field is CyanoBase (<CITE> Nakao et al., 2010).
Compare	The resource is compared with other resources.	Our logistic regression linear parser and re-implementation of Chen and Manning (2014) give comparable accuracies to the perceptron ZPar <CITE> and Stanford NN Parser, respectively.
Extent	The resource is the foundation of this paper's work or some improvements are made based on the resource.	We integrate a phonetic-based encoding scheme, UrduPhone, a feature-based similarity function, and a clustering algorithm, ... and Double Metaphone <CITE>.
Other	The function do not belong to the above 5 categories will be classified into Other.	We would also like to acknowledge the Cognitive Rhythms Collaborative (<CITE>), ...

definitely links to a specific online resource. And a *resource context* is a sequence of words that surround the resource citation, specifically the sentence where a hyperlink appears, as well as the two sentences that appear before and after. Through observing the publications, we find that most resource citations can be divided into two types according to the locations of their hyperlinks: the *in-line resource citations* in bodytext and the *additional resource citations* in footnotes. We further give definitions for the *resource role* and the *resource function*: the *resource role* is the class of a resource and indicates what role the resource plays in its context (e.g. **Tool**, **Data** and **Code**). And the *resource function* is the specific purpose the resource perform with respect to the current paper's work and answers why the author cite that resource here (e.g. **Use**, **Produce** and **Compare**). As Figure 1 shows, there are examples for the two resource citation types, where we note the arguments (which are mostly the target nominals ahead the citation) for identifying the role and the arguments (which are mostly the key verbs before the citation) for identifying the function.

In this paper, we propose a *scientific resource annotation scheme* to frame the role and function of resource citations. We collect a new *scientific resource dataset* from scientific publications of multiple sources. Based on the collection and our annotation scheme, we develop a dataset[2] of 3,088 manually annotated resource context. Then we adopt a novel neural-based method using Bi-LSTM with attention mechanism and position indicator to respectively classify the role and function for each resource citation by analyzing the citation context. Our model shows comparable performance with other reliable baselines in the classification task. We apply our model on the large ARC dataset and examine the changes in resources' function to show the evolution and maturity of the scientific resources.

2 SCIENTIFIC RESOURCE DATASET

Based on the previous argumentative annotation schemes [5, 8**?**] and our hand-analyzed data, we propose a novel resource annotation scheme. By merging some subclasses which have the similar implication into a root class, we finally conduct 9 *resource role* categories and 6 *resource function* categories as shown in Table 1. To the best of our knowledge, due to the difficulties in collecting

[2]https://github.com/zhaohe1995/SciRes

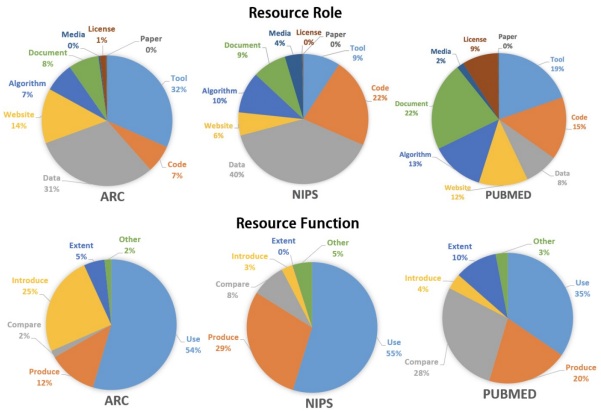

Figure 2: The distribution of resource role and function between different data sources.

large scale of scientific full-text from publications in PDF format, there is no ready-to-use dataset for our task. So we construct a *scientific resource dataset* which will facilitate the future research for context-based resource analyzing in scientific literature. We use the scientific literature from three different sources: the ACL Anthology Reference Corpus (ARC)[3], the NIPS Proceedings (NIPS)[4] and the PubMed[5]. We collect all the 21,411 papers of ARC, all the 7,147 papers of NIPS from 1988 to 2017, and randomly download 11,043 publications from the PubMed. For data processing, we use Omnipage[6] to perform OCR in translating the PDF files to textual format. Then we apply a conditional random field-based parsing tool, ParsCit[7], to extract the metadata and structural information. We extract all the hyperlinks in a scientific literature from both the bodytext and the footnotes. A five-sentence resource context is extracted along with the hyperlink. Finally, we construct a collection of 52,705 data samples.

Annotations were performed by a group of 3 PhD students, of which one majors in NLP, one majors in deep learning and the other majors in biological information. We randomly select 1,100 data samples from each scientific literature source. Since too short texts may not cover sufficient information for identifying the target resource role and fucntion categories, we filter out the samples of which the resource context is less than 10 words. Each resource citation and its context is assigned at least one label in the *resource role* and a unique label in the *resource funciton*. Fleiss's Kappa (κ) is 0.47, indicating a relatively high agreement between annotators considering the number of categories and the difficulties of the task. Finally we get a manually annotated resource dataset of 3,088 data samples. All the categories along with their numbers are shown in Table 2.

There are some findings when comparing the distribution of resource roles and functions between different data sources, as shown in Figure 2. For resource roles, the ARC dataset has relatively more Tools which indicates more research for NLP technology applications in ARC, while the more theoretical NIPS has least Tools but

[3]http://acl-arc.comp.nus.edu.sg/
[4]http://papers.nips.cc/
[5]https://www.ncbi.nlm.nih.gov/pubmed
[6]https://www.nuance.com/
[7]https://github.com/knmnyn/ParsCit

Table 2: The 9 *resource roles* and 6 *resource functions* along with their numbers.

Role	#ARC	#NIPS	#PUBMED	Total	%
Data	333	452	219	1,004	31.0
Tool	340	106	171	617	19.0
Code	74	256	91	421	13.0
Document	81	97	132	310	9.6
Website	146	66	144	256	7.9
Paper	3	3	242	248	7.6
Algorithm	79	118	19	216	6.7
License	16	1	97	114	3.5
Media	4	49	3	56	1.7
Function	#ARC	#NIPS	#PUBMED	Total	%
Use	560	578	347	1,485	48.1
Produce	125	308	201	634	20.5
Introduce	254	89	281	624	20.2
Extent	53	30	39	122	4.0
Other	17	1	106	124	4.0
Compare	18	50	30	98	3.2

most Algorithm and Data citations. Instead of the encapsulated software, the scientific literature from NIPS prefer the implementations in codebases. Moreover, due to the difference in article formats and writing styles in different domains, literature in the field of bioinformatics from PubMed tend to link to the papers by in-line hyperlinks in bodytext while literature from ARC and NIPS tend to cite the papers in the reference lists. For resource functions, we can see that the literature in NIPS tend to **produce** or **release** more new resources while the literature in ARC and PubMed tend to **introduce** more resources as background to support their work. We infer the difference in function is because that the research in NIPS are more theory-based and often put forward new methodologies, whereas the research from ARC and PubMed are more comprehensive and contain more application-oriented works which tend to review a lot of related resources.

3 CLASSIFICATION MODEL

Based on our scientific resource dataset, many challenges make the resource classification task not easy. First, it is very important to well parse, encode and model the information in resource citation context, which is relatively short text having no more than 5 sentences. Second, as the examples shown in Figure 1, by observing the dataset we find in most cases there are key nominals or verbs implying the role and function of the resource located nearby the citation (e.g. the nearest verb before the resource citation such as "use", "apply" and "adopt" often indicates the function of **Use**). For this reason, the citation position in the word sequence is very significant information to be considered in our task. Furthermore, one salient problem is the out-of-vocabulary equations and the spelling errors introduced by OCR process, which are frequent in our dataset. Therefore, effective methods for solving these particular challenges need to be developed for this new scientific resource classification task.

Following previous works [1, 9], our approach takes advantage of attention-based LSTM with character-based embeddings to integrate the features both in char-level and word-level. We develop a

Table 3: Comparison results on *Role* and *Function*.

Method	*Role*	*Function*
AvrgEmbed+LR	0.453	0.475
AvrgEmbed+SVM	0.429	0.466
CNN	0.450	0.498
LSTM	0.413	0.399
RCNN	0.430	0.517
FastText	0.489	0.471
Our modeI	**0.532**	**0.539**

4-layer hierarchical neural model: **1) Word representation layer** concatenates three components for each word: a character-based embedding, a word embedding, and an embedding for capitalized feature and POS feature. **2) Word LSTM layer** applies a Bi-LSTM at the word level taking the concatenated character-word-feature embedding as input. The word representation is obtained by stacking the forward and backward LSTM hidden states. **3) Attention layer** associates each hidden word representation with different weights according to its contribution to the categorized label. **4) Output layer** uses the softmax function to predict the final role label or function label.

4 EXPERIMENTS

For evaluation, we report the F1-score. All hyperparameters are determined on the validation set. The word embedding dimension is 200. The LSTM hidden size is 50. The maximum word sequence length is 200 and word length is 20. The char embedding dimension is 200. The feature embedding dimension is 50, which are initialized at random. For training, we use the Adam optimizer with a mini-batch size of 128 and the learning rate of 0.01.

We compare our method with widely used sentence classification approaches and the state-of-the-art models: Average Embedding + LR/SVM; TextCNN [6]; LSTM: a 3-layer structure with input word embeddings, a bidirectional LSTM layer and a softmax output layer; RCNN [7]; FastText [4]. Our model gets comparable results with the state of the art baselines for both classification tasks, as shown in Table 4. From the table we can see that due to the great difficulties in context-based analyzing for scientific resources, the best F1 results are slightly above 0.53 for both classification task, which provides considerable potential for advancement for future research.

To study what resource functions can tell us about scientific resource development, we apply our resource function classifier trained on our annotated dataset to the ARC dataset. For experiment, we select the top 100 most frequent resources by their hyperlinks from the ARC. And then we filter out the ones which have been existing less than 10 years. Finally, we obtain a set of 33 resources and use our trained classifier to identify the function of each resource citation in its corresponding context. A statistical result is shown in Figure 3. The horizontal axis represents the number of years after the resource first appeared in the scientific corpus. From the figure we can see that a resource will drive to maturity stage and be widely used in 4-5 years after it first exists. And in 8-9 years it will perhaps be out of date and gradually replaced by other new technologies. Moreover, citing the resource as the background and drawing extension based on the resource is progressively increasing along with time, which is consistent with the general expect.

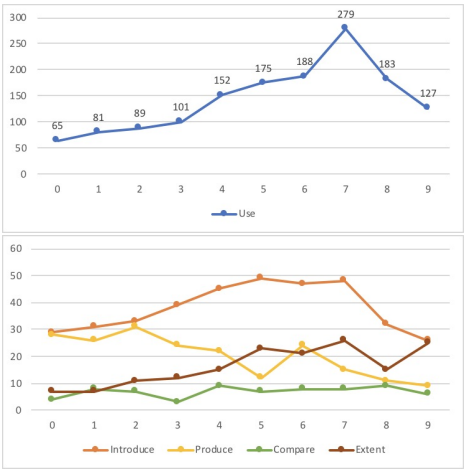

Figure 3: The revolution of resource function in ten years starting from the first appearance.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce a new task of *scientific resource citation classification* for analyzing the role and function of the on-line resources in scientific literature. We first propose a *scientific resource annotation scheme* and develop a dataset of 3,088 manually annotated resource context. Furthermore, we adopt a neural-based method to classify each resource citation and applied it on the large ARC dataset to trace the evolution of resources. For future work, we will explore more tasks, such as scientific resource evaluation, prediction and portrait construction.

6 ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China (No. 2017YFB1002101), National High-tech Research and Development Program (863 Program) of China (No. 2014AA015105), and National Natural Science Foundation of China (No. 1636203, No. 61602490).

REFERENCES

[1] Shi Feng, Yang Wang, Liran Liu, Daling Wang, and Ge Yu. 2018. Attention based hierarchical LSTM network for context-aware microblog sentiment classification. *World Wide Web* (2018), 1–23.
[2] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *WSDM*.
[3] Wenyi Huang, Zhaohui Wu, Liang Chen, Prasenjit Mitra, and C. Lee Giles. 2015. A Neural Probabilistic Model for Context Based Citation Recommendation. In *AAAI*.
[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*.
[5] David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Daniel Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *TACL* 6 (2018), 391–406.
[6] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.
[7] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification. In *AAAI*.
[8] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. In *Bioinformatics*.
[9] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific Information Extraction with Semi-supervised Neural Tagging. In *EMNLP*.
[10] Jie Tang and Jing Zhang. 2009. A Discriminative Approach to Topic-Based Citation Recommendation. In *PAKDD*.