CiteQA@CLSciSumm 2020

Anjana Umapathy* **Karthik Radhakrishnan*** **Kinjal Jain*** **Rahul Singh***
Language Technologies Institute
Carnegie Mellon University
{aumapath, kradhak2, kinjalj, rahuls2}@cs.cmu.edu

Abstract

In academic publications, citations are used to build context for a concept by highlighting relevant aspects from reference papers. Automatically identifying referenced snippets can help researchers swiftly isolate principal contributions of scientific works. In this paper, we exploit the underlying structure of scientific articles to predict reference paper spans and facets corresponding to a citation. We propose two methods to detect citation spans - keyphrase overlap, BERT along with structural priors. We fine-tune FastText embeddings and leverage textual, positional features to predict citation facets.

1 Introduction

With the ballooning growth in the number of research papers published every year (Larsen and Von Ins, 2010), being able to automatically identify the main contributions of a paper would be a useful tool to aid the ingestion of academic works. Given that citations are expected to focus on the important or note-worthy aspects of a paper (Nakov et al., 2004), tagging a citation with its corresponding reference paper snippet could help identify the main contributions of the reference paper.

While methodologies to find influential papers exist, there is currently no established method to identify the aspects of a reference paper that make it influential. Uncovering important aspects has a range of downstream applications such as discovering interesting insights about salient, frequently cited concepts or empowering researchers to quickly read and understand a large volume of papers. Ability to quickly sift through a large volume of papers is especially useful during times of crisis to get equipped with the most relevant research in a new area (such as COVID-19). External to research papers, this tool could also be leveraged by journalists and paralegals as news articles and legal documents follow a similar citation structure.

In this work, we tackle two tasks - citation span detection and facet identification. We utilize the structural characteristics of scientific articles such as section and sentence importance, location of sentences in conjunction with textual features. We present a keyphrase extraction and a BERT model for citation identification and augment sentence embeddings with hand-crafted features to classify the citation onto a set of pre-defined facets[1].

2 Task and Dataset Description

We break down our objective into two sub-tasks: Citation span detection and Facet identification.

A. Span Detection - Given a reference paper (RP), a citing paper (CP), and the citation sentence (citance), identify the spans in the reference paper which most accurately capture the citation.

B. Facet Identification - Given reference and citing papers, perform a multi-label citation classification onto 5 pre-defined facets {METHOD, RESULT, AIM, HYPOTHESIS, IMPLICATION} indicating the type of citation.

For both Task A and Task B, we make use of the dataset released as part of the shared task at CL-SciSumm (Chandrasekaran et al., 3002forthcoming) for EMNLP 2020.

The CL-SciSumm dataset contains a set of 40 reference papers (in the domain of computational linguistics), each paired with up to 35 citing papers, totalling 753 unique citations. The dataset introduces two tasks: 1) Span detection(A) and

*Equal contribution

[1]Our code and models can be found at https://github.com/karthikradhakrishnan96/CiteQA

Proceedings of the First Workshop on Scholarly Document Processing, pages 297–302
Online, November 19, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

facet identification(B) , and 2) Summarization. In this work, we focus on task 1. Each citation is tagged with the gold reference spans and one or more facets. There is, however, no inter-annotator agreement or human performance reported on this dataset. An example from our dataset is shown below.

Citing Sentence - Given that close to 95% of the word occurrences in human labeled data are tagged with their most frequent part of speech (Lee et al. , 2010)

Reference Sentence - Simply assigning to each word its most frequent associated tag in a corpus achieves 94.6% accuracy on the WSJ portion of the Penn Treebank

Facet - RESULT

We also make use of SciSummNet (Yasunaga et al., 2019) and a cleaned version (Lahiri, 2014) of the ACL-ARC corpus (Bird et al., 2008) for pre-training our models. SciSummNet contains over 1000 reference papers auto-annotated with citation spans and ACL-ARC corpus contains over 10K articles from ACL anthology.

We utilize SciSummNet to fine-tune our BERT (Devlin et al., 2019) model to adapt to 'scholarly document' style of text and use ACL-ARC corpus to generate domain-specific word embeddings using FastText (Joulin et al., 2016).

3 Related Work

Prior work on task A can be broadly classified into two categories - text similarity and deep learning based methods. Text similarity methods typically compute similarity scores between each reference sentence and the citing sentence and rank them to predict the reference spans. Similarity can be computed in different ways - Baruah (2018) use a word-embedding cosine similarity while Syed (2019) and Abura'ed (2018) use multiple similarity metrics like Jaccard, BM25, TF-IDF as features to train a classifier to predict the best reference sentence. PolyU (Cao et al., 2016) groups sentences into chunks and performs predictions by using a RankSVM over these chunks.

Deep learning models such as CiteListNet (Kim, 2019) and NacTem-UoM (Zerva et al., 2019) learn textual feature representations for classification. CiteListNet employs a text-similarity phase followed by a CNN reranker while NacTem-UoM trains BERT to score a reference, citing sentence pair and predicts top-3 reference sentences.

For the facet identification task, Wang (2018) use bag of words in conjunction with some sentence features in multiple facet-specific classifiers, Baruah (2018) use average word2vec embeddings for prediction, and Zerva et al. (2019) use bag of words with random forests.

A common theme across previous works was that they predominantly only utilized the paper text but citations often depend on external factors like section of sentence, position in section etc. We augment our models with these biases to perform a structure aware citation span and facet detection.

4 Proposed Approach

In this section, we briefly describe our data preprocessing and the models used for span and facet identification.

4.1 Data Preprocessing

As noted by previous works (Zerva et al.,2019; Wang,2018) the CL-SciSumm dataset has numerous formatting issues stemming from the Optical Character Recognition (OCR) module used to transcribe the PDF documents. We removed sentences with either over 50% of single-character alphabets or sentences with under 70% of valid words in English dictionary (these sentences usually correspond to tables and figures and annotators do not tag them as gold spans). We replaced common hexadecimal unicode characters with their ascii equivalents and fixed word fragmentation issues occurring in hyphenated words.

We also stripped all citation markers from the reference sentences and removed sentences with over two citations (a reference sentence citing multiple other works is unlikely to be substantial enough to be cited by a different paper). We filtered sentences which were either shorter than 5 or longer than 30 words as these sentences were almost never cited (as measured empirically on the training dataset). Additionally, the marker citing the reference paper was replaced with with a special '##CITATION##' token and all other citations were stripped from the citing sentences.

4.2 Task A

For task A, we incorporate two different methods for span detection - Keyphrase similarity and BERT. We also apply some inductive biases accounting for the underlying writing structure of scientific papers.

4.2.1 Keyphrase similarity

We observed (on over 200 sampled citances) that citing sentences are usually paraphrases of references and tend to reuse the same words from the reference sentence. Though prior works (Wang, 2018) have incorporated word-based similarity methods, they evaluate overlap on complete sentences. We observed (through manual evaluation) that humans focus on important keyphrase similarity as opposed checking similarity over entire sentences. Hence we extract keyphrases from reference and citing sentences through Rapid Automated Keyword Extraction algorithm (Rose et al., 2010) and measure similarity using keyphrase overlap.

4.2.2 BERT for citation identification

Domain-specific BERT models have shown superior language understanding and success on downstream tasks. Though scientific versions of BERT exist (Beltagy et al., 2019), they are trained on the broader domain of scientific text as opposed to just computational linguistics. We make use of the SciSummNet and CL-SciSumm dataset to fine-tune BERT on in-domain computational linguistics papers. We then frame the citation identification problem as a sentence-pair classification task. Positive samples are gold sentences from our reference paper and 5 negative samples were chosen per positive sample from a combination of 3 random and 2 high-word overlap sentences. The word overlap negative samples were found through Jaccard and included to discriminate similar but wrong spans from the correct ones. We use weighted-cross entropy to account for the imbalance between positive and negative pairs. During inference, we picked the top 3 sentences ranked by the probabilities produced by BERT. Our model architecture is shown in figure 1.

4.2.3 Section Importance Bias

From our preliminary analysis, we observed that the introduction and conclusions are cited more than the others. This is especially apparent when the citing text cites multiple papers along with the reference paper (indicating that the citation is relatively generic and is talking about a high-level detail about the reference). To incorporate this, we add an explicit bias to the reference sentences from these sections when the citing sentence has multiple citations, causing them to be weighted higher. More specifically, for sentences from introduction / conclusion sections we add a constant hyperparameter value for every citation present in the citing sentence.

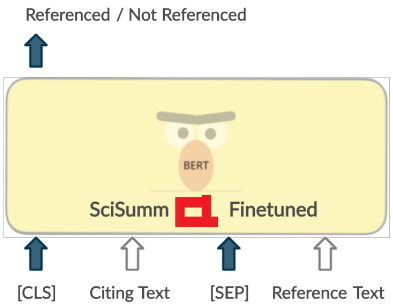

Figure 1: Architecture of our model for Task A. Citing Text is paired with every Reference sentence to predict citation probability

4.2.4 Sentence Importance Bias

Research papers usually have few key contributions and sentences which capture these aspects tend to get cited repeatedly. Beyond relevance to the citing sentence (captured by our model), we propose favoring sentences unequally based on their importance in the reference paper. We incorporate this bias using the TextRank (Mihalcea and Tarau, 2004) score of the reference sentences.

TextRank constructs a weighted graph of sentences based on the keywords present followed by a ranking phase where it assigns a score to each sentence in the graph. This score enables us to incorporate the 'citability' of a reference sentence into our predictions. The TextRank score is linearly interpolated (with the co-efficient as a hyperparameter) with our model score to generate the final score for a reference sentence.

4.3 Task B

For identifying citation facets, we make use of the reference sentences (predicted from our task A model) in conjunction with features capturing facet priors and positional information.

We train FastText on the ACL-ARC corpus to generate domain-specific embeddings for the reference sentences.

While previous works have incorporated rule-based classifiers to include facet priors, we instead

compute prior facet probabilities for each word (in the predicted reference sentences) and each section (in the reference paper) from our training data and use them as additional features. We also add a positional feature - reference SID ratio and textual features - presence of floating points or percentages (as they are highly indicative of RESULT facet).

We then train a Multi-Label Logistic Regression classifier over our input vectors (FastText embeddings + facet features) to predict the citation facet.

5 Experimental Setup

We fine-tune $\text{BERT}_{\text{BASE}}$ on Masked Language Modeling and Sequence Classification tasks using the Transformers library[2] for 2-6 epochs (with early stopping) on an NVIDIA T4 Tensor Core GPU and optimize our models using Adam (Kingma and Ba, 2019). We use 32 out of 40 papers present in the 2018 training data to fine-tune our model on sentence pair classification, reserving the remaining 8 papers for validation. We chose not to use the automatically annotated 2019 training data as we noticed a drop in performance on the validation set. The section and sentence importance bias coefficients were hyperparameters and were varied across runs. Our code repository contains more details on the exact hyperparameter values.

6 Results

Table 1 shows the performance of our models against various baselines. Our BERT model with biases achieves an F_1 of 0.128, outperforming last year's best model (NacTem-UoM). Our model performs competitively with other baseline models on the facet identification task as shown in Table 2.

7 Error Analysis

7.1 Facet Disparity

Our model achieved better performance on identifying citation spans when the underlying facet was METHOD or AIM while it was unable to identify RESULT effectively. This is because the citing and reference sentences are not similar when results are cited. A potential solution would be to identify the facet first and apply facet-specific models for span detection but we would have to ensure that such hard decisions do not cascade errors.

[2]https://huggingface.co/transformers/

Dataset	Model	Micro F_1	Macro F_1
2016-Test	PolyU	0.10	-
	Keyphrase (KP)	0.122	-
	KP + Biases	0.137	0.147
	BERT + Biases	**0.139**	0.117
Test	CiteListNet	0.124	-
	NacTem-UoM	0.126	-
	BUPT	0.087	-
	KP + Biases	0.123	0.127
	BERT + Biases	**0.128**	0.128

'-' → scores not available

Table 1: Performance of different models on Task A. 2016-Test refers to our held-out set and Test refers to the official CL-SciSumm results

Dataset	Model	Micro F_1	Macro F_1
2016-Test	PolyU	0.214	-
	KP + FT	0.285	0.295
	BERT + FT	**0.34**	0.265
Test	NacTem-UoM	0.312	-
	BUPT	**0.389**	-
	KP + FT	0.310	0.315
	BERT + FT	0.299	0.302

'-' → scores not available, FT → FastText

Table 2: Performance of different models on Task B

7.2 Unsolvable examples

Upon an initial manual annotation on a small subset of papers, we noticed that though our predictions seemed perfectly reasonable, gold annotations were often completely different sentences, highlighting the inherent ambiguity of this task and potentially explaining the low performance of all models. An example is shown below.

Citing Sentence - In the large-scale HPSG-based spoken Japanese analysis system developed at ATR, sometimes 98 percent of the elapsed time is devoted to graph unification (Kogure, 1990)

Gold Reference - Furthermore, structure sharing increases the portion of token identical substructures of FSs which makes it efficient to keep unification results of substructures of FSs and reuse them.

Our prediction - Japanese analysis system based on llPSG[Kogure 891 uses 90% - 98% of the elapsed time in FS unification.

Furthermore, we noticed some noisy annotations (P08-1102_sweta.csv contains different reference paper IDs), some papers with 0 sentences (probably owing to OCR errors), and XML formatting issues.

8 Conclusion

In this work, we show that application of biases exploiting the underlying structure of scientific texts is useful on the tasks of citation span and facet identification. In the future, we hope to incorporate these biases into the training process instead of interpolating them during evaluation.

9 Acknowledgement

We would like to thank Dr. Eric Nyberg and Dr. Teruko Mitamura for their valuable suggestions and guidance through the course of this work. We would also like to thank the task organizers and our anonymous peer reviewers for their valuable feedback on our work.

References

Bravo A. Chiruzzo L. Saggion1 H. Abura'ed, A. 2018. LaSTUS/TALN+INCO @ CL-SciSumm 2018 - Using Regression and Convolutions for Cross-document Semantic Linking and Summarization of Scholarly Literature. In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2018)*.

Kolla M. Baruah, G. 2018. Klick Labs at CL-SciSumm 2018. In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2018)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, pages 1755–1759.

Ziqiang Cao, Wenjie Li, and Dapeng Wu. 2016. Polyu at cl-scisumm 2016. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 132–138.

M. K. Chandrasekaran, G. Feigenblat, Hovy. E., A. Ravichander, M. Shmueli-Scheuer, and A De Waard. 3002forthcoming. Overview and Insights from Scientific Document Summarization Shared Tasks 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Ou S. Kim, H. 2019. Ranking-based Identification of Cited Text with Deep Learning. In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2019)*.

Diederik P Kingma and J Adam Ba. 2019. A method for stochastic optimization. arXiv 2014. *arXiv preprint arXiv:1412.6980*, 434.

Shibamouli Lahiri. 2014. ACL ARC Style Browser. http://ec2-54-186-204-149.us-west-2.compute.amazonaws.com/acl_arc_style_browser/.

Peder Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88. Citeseer.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20.

Indurthi V. Srinivasan B.V. Varma V. Syed, B. 2019. Transfer learning for effective scientific research comprehension. In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2019)*.

Li S. Wang T. Zhou H. Tang J. Wang, P. 2018. NUDT @ CLSciSumm-18. In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2018)*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 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.

Chrysoula Zerva, Minh-Quoc Nghiem, Nhung T. H. Nguyen, and Sophia Ananiadou. 2019. NaCTeM-UoM @ CL-SciSumm 2019. In *BIRNDL@SIGIR*.