Local Citation Recommendation using Retrieval Augmented Classification

Gowri Shankar Raju Kurapati
gowri.kurapati@colorado.edu

Abstract

Keeping up with the latest work related to a topic is difficult due to the exponential increase in scientific publications. This difficulty affects how we cite others' work during article writing. Previous research have attempted to assist by offering citation-oriented support, but often falls short in their accuracy, relevance, and response times. In this work, we propose to use retrieval augmented classification, and a combination of transformer-based embeddings and semantic retrieval, to increase both relevance and accuracy. We create an end-to-end architecture that combines citation worthiness prediction and citation recommendation at a local level. Further, we distill the architecture to achieve high inference rates, making it usable for real-time writing. We deploy our system as a plugin for multiple popular writing services including Google Docs and others.

Keywords: Sentence Level Citation Recommendation, Knowledge Retriever

1 Introduction

Linking citation context to the particular topic of the referenced work is a necessary step in the citation creation process. Scientists take a great deal of effort to find relevant work to cite while writing (Färber et al., 2018). They keep track of the immense body of work, find the relationship that work and their writing, and find the proper place where to put such citations. Citing sources is an intrinsically complicated action that is influenced by a variety of personal and social standards (Zhang et al., 2013).

There has been an astounding growth in published research. According to Jinha (2010), from the eighteenth century, the number of scientific publications has grown at a pace of 3% a year. When Lutz and Rüdiger (2015) evaluated publications published between 1980 and 2012, they discovered that the quantity doubles every 24 years and expands exponentially. In 2018, the annual growth rate of articles grew to 4%, while the annual growth rate of journals increased to above 5% (Johnson et al., 2018). As a result, locating and identifying references among the enormous and growing number of publications is a labor-intensive effort and is becoming more difficult. Even the best manuscript writer leave out important citations unintentionally (Tang and Zhang, 2009).

Numerous research teams have looked into the best ways to provide citations for individual articles and local contexts automatically. Küçüktunç et al. (2012) employ graph-based approaches to estimate the citation linkages between articles at the document level. By relying on collective citation patterns, McNee et al. (2002) and Torres et al. (2004) use collaborative filtering to generate these recommendations. These methods usually rely on being aware of the proper location for citations in order to produce recommendations for article-level citations. He et al. (2010) suggest context aware citation recommendation at the local level using the local contextual data of the locations where citations are made. Recent studies by Huang et al. (2015), Ebesu and Fang (2017), and Bhagavatula et al. (2018) use distributed representations, autoencoders, and two-step process of embedding documents into vector representations & ranking them according to the relevance respectively, to estimate a semantic representation of sentences.

Finding the proper location for a citation has received significantly less attention than other tasks. Though Recurrent Neural Networks (Färber et al., 2018), Convolution Neural Networks (Bonab et al., 2018) and Attention based networks (Zeng and Acuna, 2020) have been used for detecting the citation worthiness of sentence, extensive research hasn't been done on the finding the citation recommendations when the sentence is deemed worthy. SenCite (Wang et al., 2022) finds the citation worthiness using a convolution recurrent neural network and then suggests citations based on the conspicuous similarity between the sentences in the target articles' abstract, full text, and in-link context. In Yang et al. (2019) learns relationships between variable-length texts of the two text objects, citation contexts and scientific articles, and incorporates venue information and author information in the attention mechanism to do local-level citation recommendation. Not much work has been done on integration of the citation worthiness and retrieving the recommendations in a single architecture.

We explored the Open Domain Question Answering which has been widely explored task in the field of Natural Language Processing which tries to answer a question in natural language using large scale documents. The quantity of literature review on OpenQA has recently increased, notably on strategies that combine with Knowledge Retrievers (Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020). We leverage the idea of retrieval-augmented Language Model to predict the citation worthiness of a given sentence or a query. In contrast to storing the knowledge in the parameters of the language model, we ask the model to retrieve the knowledge from a large corpus of documents of <DatasetName> and use during the classification of citation worthiness. The retrieved documents from the corpus which are used to predict the citation worthiness are the recommendations that we use to cite for the query. This way, we are able to combine the classification of worthiness and giving recommendations to cite in a single architecture. In sum, the contributions of our paper are as follows:

- Integrate citation worthiness with local citation
- Create a new dataset for the task based on open access publications
- Use ideas from Open Domain Question Answering to solve the location and citation recommendation task
- Propose a new idea that performs significantly better than traditional models
- (Optional) To decrease inference time and size of the model, we distill our big model into a smaller one, and show that we halve inference time while retaining competitive performance.

References

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*.

Hamed Bonab, Hamed Zamani, Erik Learned-Miller, and James Allan. 2018. Citation worthiness of sentences in scientific reports. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1061–1064.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1093–1096.

Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. To cite, or not to cite? detecting citation contexts in text. pages 598–603.

K Guu, K Lee, Z Tung, P Pasupat, and MW Chang. 2020. Realm: Retrieval-augmented language model pre-training. arxiv 2020. *arXiv preprint arXiv:2002.08909*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Twenty-ninth AAAI conference on artificial intelligence*.

Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned publishing*, 23(3):258–263.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. The stm report: An overview of scientific and scholarly publishing.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. 2012. Direction awareness in citation recommendation.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. 2002. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125.

Jie Tang and Jing Zhang. 2009. A discriminative approach to topic-based citation recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 572–579. Springer.

Roberto Torres, Sean M McNee, Mara Abel, Joseph A Konstan, and John Riedl. 2004. Enhancing digital libraries with techlens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 228–236.

Hei-Chia Wang, Jen-Wei Cheng, and Che-Tsung Yang. 2022. Sentcite: a sentence-level citation recommender based on the salient similarity among multiple segments. *Scientometrics*, 127(5):2521–2546.

Libin Yang, Zeqing Zhang, Xiaoyan Cai, and Tao Dai. 2019. Attention-based personalized encoder-decoder model for local citation recommendation. *Computational Intelligence and Neuroscience*, 2019.

Tong Zeng and Daniel E Acuna. 2020. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. *Scientometrics*, 124(1):399–428.

Guo Zhang, Ying Ding, and Staša Milojević. 2013. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, 64(7):1490–1503.