Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models

Tong Zeng Daniel E Acuna

2020-03-28

Scientist learn early on how to cite scientific sources to support their claims. Sometimes, however, scientists have challenges determining where a citation should be situated—or, even worse, fail to cite a source altogether. Automatically detecting sentences that need a citation (i.e., citation worthiness) could solve both of these issues, leading to more robust and well-constructed scientific arguments. Previous researchers have applied machine learning to this task but have used small datasets and models that do not take advantage of recent algorithmic developments such as attention mechanisms in deep learning. We hypothesize that we can develop significantly accurate deep learning architectures that learn from large supervised datasets constructed from open access publications. In this work, we propose a bidirectional long short-term memory network with attention mechanism and contextual information to detect sentences that need citations. We also produce a new, large dataset (PMOA-CITE) based on PubMed Open Access Subset, which is orders of magnitude larger than previous datasets. Our experiments show that our architecture achieves state of the art performance on the standard ACL-ARC dataset (F1 = 0.507) and exhibits high performance (F1 = 0.856) on the new PMOA-CITE. Moreover, we show that it can transfer learning across these datasets. We further use interpretable models to illuminate how specific language is used to promote and inhibit citations. We discover that sections and surrounding sentences are crucial for our improved predictions. We further examined purported mispredictions of the model, and uncovered systematic human mistakes in citation behavior and source data. This opens the door for our model to check documents during pre-submission and pre-archival procedures. We discuss limitations of our work and make this new dataset, the code, and a web-based tool available to the community.

Introduction

Scientists and journalists have challenges determining proper citations in the ever increasing sea of information. More fundamentally, when and where a citation is needed—sometimes called citation worthiness—is a crucial first step to solve this challenge. In the general media, some problematic stories have shown that claims need citations to make them verifiable—e.g., *the debunked A Rape on Campus* article in the Rolling Stone magazine (Wikipedia contributors 2018). Analyses of Wikipedia have revealed that lack of citations correlates with an article's immaturity (Jack et al. 2014; Chen and Roth 2012). In science, the lack of citations leaves readers wondering how results were built upon previous work (Aksnes and Rip 2009). Also, it precludes researchers from getting appropriate credit, important during hiring and promotion (Gazni and Ghaseminik 2016). The sentences surrounding a citation provide rich information for common semantic analyses, such as information retrieval (Nakov et al. 2004). There should be methods and tools to help scientists cite; in this work, we want to understand where citations should be situated in a paper with the goal of automatically suggesting them.

Relatively much less work has been done on detecting where a citation should be. He et al. (2011) were the first to introduce the task of identifying candidate location where citations are needed in the context of scientific articles. Jack et al. (2014) studied how to detect citation needs in Wikipedia. Peng et al. (2016) used the learning-to-rank framework to solve citation recommendation in news articles. These are very diverse domains, and therefore it is difficult to generalize results. We contend that a large standard dataset of citation location with open code and services would significantly improve the systematic study of the problem. Thus, the task of citation worthiness detection is relatively new and needs further exploration.

The attention mechanism is a relatively recent development in neural networks motivated by human visual attention. Humans get more information from the region they pay attention to, and perceive less from other regions. An attention mechanism in neural networks was first introduced in computer vision (Sun and Fisher 2003), and later applied to NLP for machine translation (Bahdanau et al. 2014). Attention has quickly become adopted in other sub-domains. Luong et al. (2015) examined several attention scoring functions for machine translation. Li et al. (2016) used attention mechanisms to improve results in a question-answering task. Zhou et al. (2016) made use of an attention-based LSTM network to do relational classification. Lin et al. (2017) used attention to improve sentence embedding. Recently, Vaswani et al. (2017) built an architecture called transformer that promises to replace recurrent neural networks (RNNs) altogether by only using attention mechanisms. These results show the advantage of attention for NLP tasks and thus its potential benefit for citation worthiness.

In this study, we formulate the detection of sentences that need citations as a classification task that can be effectively solved with a deep learning architecture that relies on an attention mechanism. Our contributions are the following:

1. A deep learning architecture based on bidirectional LSTM with attention and contextual information for citation worthiness
2. A new large scale dataset for the citation worthiness task that is 300 times bigger that the next current alternative
3. A set of classic interpretable models that provide insights into the language used for making citations
4. An examination of common citation mistakes—from unintentional omissions to potentially problematic mis-citations
5. An evaluation of transfer learning between our proposed dataset and the ACL-ARC dataset
6. The code to produce the dataset and results, a web-based tool for the community to evaluate our predictions, and the pre-processed dataset.

Data sources and data pre-processing

PMOA-CITE

In this paper, We constructed a new dataset called `PMOA-CITE` based on PubMed central open access subset.

PubMed Central Open Access Subset (PMOAS) is a full-text collection of scientific literature in bio-medical and life sciences. PMOAS is created by the US's National Institutes of Health. We obtain a snapshot of PMOAS on August, 2019. The dataset consists of more than 2 million full-text journal articles organized in well-structured XML files by the National Information Standards Organization (`ANSI/NISO 2013`).

We prepare the `PMOA-CITE` in the following steps:

1. Sentence segmentation and outlier removal. Text in a PMOAS XML file is marked by a paragraph tag, but there might be other XML tags inside paragraph tags. Therefore, we needed to get all the text of a paragraph from XML tags recursively and break paragraphs into sentences. We used spaCy Python package to do the sentence splitting (`Honnibal and Montani 2017`). However, there are some outliers in the sentences (e.g., long gene sequences with more then 10 thousand characters that are treated as one sentence). Base on the distribution of sentence length (see Fig. 3), we remove the sentences that are outliers either in character or word length. We winsorize 5% and 95% quantiles. For character-wise length, this amounts to 19 characters for 5% quantile and 275 characters for 95% quantile. For word-wise length, it is 3 words and 42 words, respectively.
2. Hierarchical tree-like structure.
3. Citation hints removal. The citing sentence usually has some explicit hints which discloses a citation. This provides too much information for the model training and it does

Table 1: regular expression to remove the citation hints

Regular expression	Description	Example
(?<!^)([\[\(])[\s]*([\d][\s\,\-\–\;\-]*)*[\d][\s]*[\]\)]	numbers contained in parentheses and square brackets	“[1, 2]”, “[1- 2]”, “(1-3)”, “(1,2,3)”, “[1-3, 5]”, “[8],[9],[12]”, “(1-2; 4-6; 8)”
[\(\[]\s*([^\(\)\[\]]*(((16|17|18|19|20)\d{2}(?!\d))|(et[\. \s\\xa0]*al\.))[^\(\)]*)?[\)\]]	text within parentheses	“(Kim and li, 2008)”, “(Heijman , 2013b)”, “(Tárraga , 2006; Capella-Gutiérrez , 2009)”, “(Kobayashi et al., 2005)”, “(Richart and Barron, 1969; Campion et al, 1986)”, “(Nasiell et al, 1983, 1986)”
et[\. \s\\xa0]+al[\.\s\(\[]*((16|17|18|19|20)\d{2})*[]\] \s]*(?=\D)	remove et al. and the following years	“et al.”, “et al. 2008”, “et al. (2008)”

not faithfully represents a real-world application scenario. Thus, we removed all the citation hints by regular expression (see Table 1).

4. Noise removal

After the processing, we get a dataset (PMOA-CITE) with approximately 309 million sentences. However, due to the computational cost and in order to make all of our analysis manageable, we randomly sample articles whose sentences produce close to one million sentences. We further split the one million sentences, 60% for training, 20% for validating, and 20% for testing. We present some characteristics of the whole dataset and one million sentence sample in Table 2.