ParsCit: A Reference Parsing Tool

Wilson Poulter
COMPSCI 9647B: Unstructured Data
Dr. Brent Davis
April 8, 2022

ParsCit: A Reference Parsing Tool

Introduction

ParsCit (a portmanteau of “parse” and “citation”) is a tool that can be used to extract the citational content from a document such as a research paper or a legal case. *ParsCit* finds not only what other documents have been cited in the original document, but also where these citations occur in the original document (Councill et al., 2008). In doing so, *ParsCit* automates what would otherwise be a very labour-intensive process on the massive collection of scientific and other scholarly articles published each year. This tool allows for libraries and other bibliographic services, such as *CiteSeer^X* or *Google Scholar*, to create citation networks, improve document similarity searches by adding citational similarity facets, and to compute researcher or journal impact indices such as the *h-index* (Hirsch, 2005), the *Journal Impact Factor* (Garfield, 1999), or the *Science Citation Index* (Garfield, 1972).

Capabilities and Mechanisms

The extraction of references is a difficult task: most scholarly documents are only available in pdfs that do not tag their references in a structured manner, and furthermore, there is no single standard for providing references. Even amongst these variations, writers are prone to errors when constructing their bibliographies. The task of citation analysis is to identify the other documents in the corpus that are cited within the current document, revealing relationship between the documents in a corpus, despite these difficulties. To do so, a citation parsing tool must distinguish not only the reference strings from other elements of the document, but also the elements of each reference string itself, such as the author, title, and year of publication. Lastly, the reference string should in many cases be related to another string in the document (i.e., the in-text citation) that provides the context for the reference.

ParsCit is a citation parsing tool that is capable of doing all of the above through supervised learning using a *conditional random field* (CRF) model (Councill et al., 2008). The CRF model was introduced to label sequential data, such as text strings, whose labels may show dependency on the labels that have come before it (Lafferty et al., 2001). A CRF model has a simple set-up: given an input space X, a label space Y, and training data $(x, y) \in X^* \times Y^*$, we want to construct a model that outputs $\hat{y} = \text{argmax}_y P(y \mid x)$ for an input sequence x. Furthermore, we constrain our model of $P(y \mid x)$ by stating that $P(y_i \mid x, y_j : i \neq j) = P(y_i \mid x, y_{i-1})$, that is, the probability of label y_i is only conditioned by x and the label y_{i-1} that came before it. From this, one can construct a set of feature functions $f_k(i, x, y_i, y_{i-1})$ and parameters λ_k so that $P(y \mid x) = \exp\left(\sum_{k,i} \lambda_k f_k(i, x, y_i, y_{i-1})\right)$. Based on this model, the maximum log-likelihood is optimized with respect to λ given the training data.

In *ParsCit*, the input space is given by word tokens that may include punctuation, the label space is given by thirteen common fields found in reference strings (e.g, author, title, year, etc.), and there are 23 total features that are observed, such as punctuation, location in the reference string, and whether the token matches an entry in a pre-defined dictionary of names, locations, publishers, etc. Thus, a feature f that represents known surnames may output 1 on input (i, x, y_i, y_{i-1}) if x_i is a known surname and y_i is an author label, while on the other hand, if y_i was a date label, then f may output 0.

To locate the reference strings, *ParsCit* relies on a UTF-8 encoding of the text and a set of heuristics to scan for the reference section of the document. First, it searches for labels such as "bibliography", "references", or "notes" in the texts. If such a label appears early in the text, *ParsCit* will continue to search for similar candidates. Once it has located the start of the section, it tries to locate the end of the section by similar means, finding the next section label that is unrelated to references, or the end of the document. The reference strings, once located, are segmented into their individual strings by using regular expressions built around parenthetical, bracketed, or numbered references (e.g., "[1]", "(1)", "1"). If these are not found, then other heuristics such as short lines of text, punctuation, or author names are used to segment them. Once this segmentation is complete, the CRF model can be deployed. The extracted metadata is then normalized based on its assigned label to fit a chosen standard. To find the citation contexts associated with each reference, a regular expression generated by the reference string is used to search the document for possible matches. The form of this regular expression depends on whether the reference was marked or unmarked. The output of *ParsCit* is given in an XML file that tags each part of the reference string with its appropriate label, and includes the context citations as well.

Related Methods

Citation analysis was first introduced by Eugene Garfield's *Science Citation Index* (Garfield, 1955), a source in which all scientific articles would be indexed so that each entry contains the indices of all other documents that reference the document at the current index. Though in 1955, Garfield's project would have to be completed manually using card catalogs, he theorized this process's automation early on (Garfield, 1965). While further development in pure citation analysis flourished in the years following Garfield's invention, like the creation of citation networks (Price, 1965), it was not until 1998 that automatic citation was introduced through *CiteSeer* (Giles et al., 1998).

At its genesis, *CiteSeer* could crawl the internet using various search engines and some heuristics to locate Postscript documents that likely contained research articles. To extract the metadata from citations found in these documents, *CiteSeer* used a set of detailed heuristics to identify labelling patterns in the citations. *CiteSeer* enabled users to easily navigate citation chains through links between various document's citations, and further developed a novel document similarity measure "common citation-inverse document frequency" (CCIDF), which conducts a Bag of Citations type-search on documents, but weighted according to inverse frequency in the corpus; thus, documents would be more similar if they cite a common article that is uncommonly cited within the whole corpus.

ParsCit marked an improvement upon *CiteSeer*'s citation extraction tools by using the CRF learning methods described in the section above rather than just pre-defined heuristics. Recent contributions have been made that expand this CRF model by eschewing pre-determined features, and instead use a bidirectional long short-term memory neural network on word and character embeddings on tokens, which are themselves then fed into either a soft-max or a CRF to predict the output labels for a sequence (Prasad et al., 2018). This model, called *Neural ParsCit* by its creators, performs better than *ParsCit* on multi-lingual documents. That said, the CRF module that the neural output is fed into remains vital to the accuracy of this model, as its removal tends to randomize labels on difficult tokens.

While *ParsCit* and its extensions can effectively extract the syntactic elements of citational information in research literature, it is important to recognize that not all citations should be treated equally, as citations may be used in one study to indicate either the support or the contrast of results from another study. An attempt to capture this nuance is made by the creators of *scite* (Nicholson et al., 2021), which is a tool that extracts the contextual information found at citations using a *SciBERT* architecture (Beltagy et al., 2019) to determine whether a reference is either (i) supporting, (ii) contrasting, or (iii) mentioning, based on training data that has been labelled with these classes by scientists. Although the effectiveness of this method is limited, especially on the contrasting class, this method invites a new wave of innovation in citation-analysis that is deeper than before.

Opportunities

Several opportunities present themselves with the continued development of citation-parsing tools, such as (i) improved architectures, (ii) content-based indices, and (iii) distributed infrastructures.

(i) Citation segmentation for humans is format rather than text-based. Further developing computer-vision to aid in citation segmentation would benefit the overall citation parsing architecture (Bhardwaj et al., 2017).

(ii) Given *scite*'s ability to provide context to references, future iterations of citation-based indices may take this into account by penalizing or rewarding contrasting vs. supporting citations, respectively.

(iii) The continued development of a universal and complete citational index through a network distributed over library servers that is available freely and globally (Lauscher et al., 2018).

References

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text [arXiv: 1903.10676]. *arXiv:1903.10676 [cs]*. Retrieved March 14, 2022, from http://arxiv.org/abs/1903.10676

Bhardwaj, A., Mercier, D., Dengel, A. & Ahmed, S. (2017). DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction [Series Title: Lecture Notes in Computer Science]. In D. Liu, S. Xie, Y. Li, D. Zhao & E.-S. M. El-Alfy (Eds.), *Neural Information Processing* (pp. 286–293). Springer International Publishing. https://doi.org/10.1007/978-3-319-70096-0_30

Councill, I. G., Giles, C. L. & Kan, M.-Y. (2008). ParsCit: An open-source CRF reference string parsing package. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 661–667.

Garfield, E. (1955). Citation Indexes for Science. *Science*, *122*(3159), 108–111.

Garfield, E. (1965). Can Citation Indexing be Automated? *Statistical Association of Methods for Mechanized Documentation, Symposium Proceedings, Washington 1964*, 189–192.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471

Garfield, E. (1999). Journal impact factor: A brief review. *Canadian Medical Association Journal*, *161*(8), 979–980.

Giles, C. L., Bollacker, K. D. & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. *Proceedings of the third ACM conference on Digital libraries - DL '98*, 89–98. https://doi.org/10.1145/276675.276685

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Lafferty, J., McCallum, A. & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

Lauscher, A., Eckert, K., Galke, L., Scherp, A., Rizvi, S. T. R., Ahmed, S., Dengel, A., Zumstein, P. & Klein, A. (2018). Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 109–118. https://doi.org/10.1145/3197026.3197050

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P. & Rife, S. C. (2021). Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, *2*(3), 882–898. https://doi.org/10.1162/qss_a_00146

Prasad, A., Kaur, M. & Kan, M.-Y. (2018). Neural ParsCit: A deep learning-based reference string parser. *International Journal on Digital Libraries*, *19*(4), 323–337. https://doi.org/10.1007/s00799-018-0242-1

Price, D. J. d. S. (1965). Networks of Scientific Papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510