

Parsing Citations Using Conditional Random Fields

Qing Zhang, MS, Yong-Gang Cao, PhD, Hong Yu, PhD
University of Wisconsin-Milwaukee, Milwaukee, WI

Abstract

With the increasing numbers of biomedical articles becoming open access, there is a greater need to develop natural language processing systems that take into account more than abstracts and work on full text. Citation plays an important role in both the rhetorical structure and the semantic content of these articles, and as such, has been of benefit to many text mining tasks, including information retrieval, extraction, summarization, and question answering.

We define a full citation as a citation that incorporates four or more of the following fields: Author (further separated by Surname-SN, GivenName-GN), Title, Source (i.e., journal, conference, or other source of publication), Volume (VOL), Pages (further separated by FirstPage-FPAGE, and LastPage-LPAGE), and Year.

In order for any text mining system to benefit from citation information, citations must be automatically identified and extracted from full-text articles and then the fields must be extracted from individual citations. However, this automation is not a simple task, and one of the greatest challenges being that citations appear in different formats.

Methods: We designed a large scale, corpus-based supervised machine-learning approach that applies conditional random fields (CRFs) to automatically segment a citation into its fields.

Data: We downloaded a large number of open-access articles from PubMed Central (PMC, <http://www.pubmedcentral.nih.gov/>). We found that for certain articles, PMC provides a choice of two formats: the parsed XML format from which the citation fields can be automatically extracted and the HTML format from which the original full citation can be automatically extracted, and we used only these articles for our study. We ensured that our data covered every journal in PubMed Central to ensure the generalizability of our data. We selected a total of 672 articles with an average of 41 citations per article. The 27,606 citations that resulted were used for training and evaluation.

Evaluation and Results: We performed 10-folder cross-validation. The overall precision of our system

is 0.9794, recall is 0.9796, and F1 is 0.9795. We implemented our model and the system can be accessed at

<http://www.cs.uwm.edu/~qing/projects/cithit/index.html>.

Acknowledgement: We are thankful for the support of 5R01LM009836-02 to Hong Yu and Lamont Antieau for proofreading.