

CitHit – An Automatic Citation Recognition System

Qing Zhang, MS¹, Yong-Gang Cao PhD², Hong Yu, PhD^{1,2}
¹Department of Computer Science, ²Department of Health Sciences

University of Wisconsin-Milwaukee, Milwaukee, WI

Abstract: We report CitHit, a fully implemented citation recognition system that extracts and parses citations from biomedical articles (PDF format) (<http://www.cs.uwm.edu/~qing/projects/cithit/index.html>). CitHit implements both rule-based and machine-learning algorithms and performs with over 97.5% accuracy for recognizing the boundary of a citation and over 95% accuracy for parsing a citation into its fields.

Introduction

As open-access, full-text articles are increasingly available, and there are natural language processing challenges in mining knowledge from the full text. Citation is an important component of this knowledge as it connects different articles semantically. Automatically parsing citations will assist biomedical text mining.

We developed CitHit, an automatic citation recognition system, to recognize the citations in PMC articles in PDF format. The system extracts citations from articles, parses the full citations (discerning fields like author name, article title) and then maps them to the citation id in the full text. In addition, CitHit maps each citation to a unique PMID; by doing so, CitHit disambiguates citations that refer to different articles, and group those that correspond to the same. We applied the supervised machine-learning algorithm the conditional random fields to parse the full citation and reported over 95% accuracy.

System Architecture CitHit consists of four units: *Splitting*, *Spotting*, *Segmentation*, and *Mapping*. Splitting identifies the boundary of a full citation in a citation list. Spotting locates the mention of a citation in the full text. Segmentation parses a full citation to recognize fields, including author name, title, and source. Finally, the mapping unit assigns a PMID to the parsed citation. The system architecture is illustrated in Figure 1.

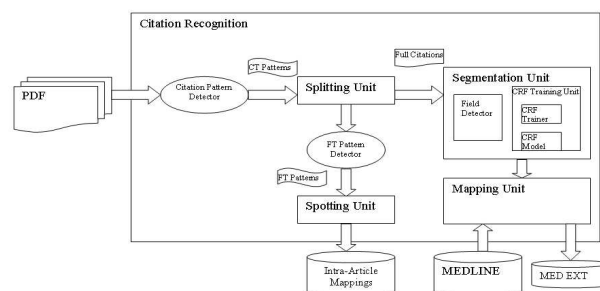


Figure 1 System Architecture

Evaluation Our preliminary evaluation results show an F-score of 0.975 for splitting unit, 0.778 for spotting unit, 0.979 for segmentation unit, and 0.896 for mapping unit.

Related Work The Science Citation Index (SCI) constructs citations and indexes scientific articles. However, the database is mostly built manually. CiteSeer¹ has a tool for automatic citation recognition. The tool, however, is not publicly available. ParsCit² is an automatic citation recognition tool. However, its overall performance is unknown.

Acknowledgement: The authors acknowledge support from the National Library of Medicine to Hong Yu, grant number 1R01LM009836-01A1, and Lamont Antieau for proofreading.

References

1. Giles CL, Bollacker KD, Lawrence S. CiteSeer: An automatic citation indexing system. In: *Proceedings of the third ACM conference on Digital libraries*. ACM New York, NY, USA; 1998:89-98.
2. Councill IG, Giles CL, Kan MY. ParsCit: An open-source CRF reference string parsing package. In: *Proceedings of LREC*. Vol 2008.; 2008.