Published in final edited form as:

Comput Biol Med. 2011 April ; 41(4): 190–194. doi:10.1016/j.combiomed.2011.02.005.

Parsing Citations in Biomedical Articles Using Conditional Random Fields

Qing Zhang, Yong-Gang Cao, and Hong Yu*

University of Wisconsin-Milwaukee, Milwaukee, WI, 53211, USA

Qing Zhang: qing@uwm.edu; Yong-Gang Cao: yonggang@uwm.edu

Abstract

Citations are used ubiquitously in biomedical full-text articles and play an important role for representing both the rhetorical structure and the semantic content of the articles. As a result, text mining systems will significantly benefit from a tool that automatically extracts the content of a citation. In this study, we applied the supervised machine-learning algorithms Conditional Random Fields (CRFs) to automatically parse a citation into its fields (e.g., Author, Title, Journal, and Year). With a subset of html format open-access PubMed Central articles, we report an overall 97.95% F1-score. The citation parser can be accessed at: <http://www.cs.uwm.edu/~qing/projects/cithit/index.html>.

Keywords

natural language processing; information extraction; citation parsing; citation indexing; conditional random fields; machine learning; biomedical text mining

1 Introduction

As more and more full-text biomedical articles become open-access, there is a great need to move beyond merely examining abstracts and to develop text mining approaches that apply to full-text articles. Citations are used ubiquitously in biomedical articles; for instance, we found an average of 34 citations for the 160,000 full-text biomedical articles in the TREC Genomics Track text collection [1]. Citations play important roles for both the rhetorical structure and the semantic content of the articles, and as such, citation information has shown to benefit many text mining tasks including information retrieval, information extraction, summarization, and question answering.

For example, citation indexing has been used to associate citing articles with cited ones, and the associations have been used to score Science Citation Index to measure the impact factors of scientific journals and articles [2]. Two articles can be considered as “related” if they share a significant set of co-citations and a study that incorporated this model has

© 2011 Elsevier Ltd. All rights reserved.

*Corresponding author. hongyu@uwm.edu.

Conflict of interest statement

There is no conflict of interest: ‘none declared’.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

shown to improve information retrieval [3]. The number of times a citation is cited in a paper may indicate its relevance to the citing paper [4, 5]. Citances (or the citation sentences) sometimes represent the condensed semantic content of the documents they identify [6, 7] and have been used to extract scientific fact [8] and for summarization [7]. In addition, citation indexing has been used to model the evolution of author and paper networks [9] and research collaboration [10].

In order for text mining systems to benefit from citation information, one must automatically identify citations from full-text articles and extract their fields, including *Author*, *Title*, *Journal* and *Year*. Citation parsing automatically parses a full citation into its fields. Co-authorship is frequent in the biomedical domain, and it is important for a citation parser to identify the information of each author, including given name and surname. Separating an author's surname from the given name will enable a text mining system to separate two different authors (e.g., "John Smith" and "Smith John") who share the same names.

Citation parsing is challenging because citations come with different formats that are rooted in either different requirements by different publishers or nonstandardized formats introduced by authors. The following examples illustrate some variations in citation format:

[Example 1] Yu, H and Lee M. 2006. Accessing Bioscience Images from Abstract Sentences. *Bioinformatics*. Vol 22 No. 14, pages e547–e556.

[Example 2] Hong Yu and Minsuk Lee. Accessing Bioscience Images from Abstract Sentences. *Bioinformatics*. Vol 22 No. 14, pages e547–e556. 2006.

[Example 3] Yu H, Lee H. 2006. Accessing Bioscience Images from Abstract Sentences. *Bioinformatics*: 22 (14), e547–e556.

First, the order of fields may vary. As shown in Examples 1 and 2, the publication year (2006) may appear before or after the title. There are also different ways to present publication volume and issues, for example, "Vol 22 No. 14" and "22 (14)" in Example 2 and 3. Author names also come with different format (e.g., "Yu H", "Hong Yu", "Yu, H" or "H. Yu"); Journal names vary as well. For example, an article that is published in "American Medical Informatics Association Fall Symposium" can be referenced as "AMIA," "Proceedings of AMIA", "Proceedings of the American Medical Informatics Association Fall Symposium," "Proceedings of the American Medical Informatics Association (AMIA) Fall Symposium," or "Proc AMIA" These variations pose a significant challenge for developing a natural language processing system that automatically parses citations into fields, since we can see from the example each field may have different formats.

In this work we report on the development of a natural language processing system that automatically parses a citation into its fields (i.e., *Author Given Name*, *Author Surname*, *Title*, *Source*, *Year*, *Volume*, *First Page*, and *Last Page*). Although citation parsing is not new, and tools for doing it have been widely used in CiteSeer, few tools are publicly available and little work has been evaluated for citation parsing in the biomedical literature.

2 Related Work

The Institute for Scientific Information (ISI) constructs citations and indexes scientific articles to produce multidisciplinary citation indices including the Science Citation Index (SCI). However, the database is mostly built manually. Similarly, HighWire Press (<http://highwire.stanford.edu>) and other open journal projects link citations across journal sites. It is unclear whether they provide tools to automatically recognize citations.

An *Autonomous Citation Indexing (ACI)* system automatically locates articles, extracts citations, identifies identical citations that occur in different formats, and identifies the fields

of citations and has been developed and implemented by CiteSeer [11], [12]. Citation parsing is a subtask of ACI.

Earlier work in citation parsing uses manually crafted rules and external databases. CiteSeer [11, 12] is such an application that is built upon a set of heuristic rules (e.g., the format of any citation is uniform within one article) and external databases of author names and journal names. They reported a performance of 80.2% for *Title*, 82.1% for *Authors*, and 44.2% for *Page Number*. Wei (2007) [13] reported a similar approach. Besagni and Belaid (2004) [14] also developed a rule-based system. Their system assigned each field a tag (e.g., *alphanumeric string* and *common word, capital initial*) that resembles a part-of-speech tag. Rules (e.g., “when author names are given, it is always at the beginning of the reference”) were used to detect each field. They evaluated their system on a dataset of 2,575 citations that came from 64 articles randomly selected from 140 journals, and reported that 75.9% references were completely parsed.

Powley and Dale (2007) [15] observed that the pattern of a citation mentioned in the body of a full-text article (e.g., “*Yu and Lee, 2006*”) is consistent with the citation pattern in the reference section, and accordingly they developed rules to capture the author name. They reported 92% precision and 100% recall, although the system works well only when the citation is presented by author name and year, not by other common patterns (e.g., the digit format (e.g., “[1]”). Other related work link citations to the same reference that differ in format [16].

Statistical and machine-learning approaches have been reported for citation parsing. Takasu (2003) [17] developed a Hidden Markov Model (HMM) for citation parsing and reported over 90% accuracy on an evaluation data of 1,575 citations.

BibPro [18] is a citation parser that is based on sequence alignment techniques. Specifically, they used sequence alignment tool BLAST to find the most similar citation fields for a citation, and subsequently parsed the citation and reported an accuracy of 97.68%.

Kramer et al (2007) [19] developed Probabilistic Finite State Transducers (PFSTs) for citation parsing. Based on the observation that the content of a citation field is independent of other fields, they built an FST model for each citation field. An evaluation on the Cora dataset that serves as a common benchmark for accuracy measurements yields a field accuracy of 82.6%.

Peng and McCallum (2006) [20] parsed citations with Conditional Random Fields (CRFs), and reported an F1-score ranging from 76.1% (for *Publisher*) to 99.4% (for *Author*), although their system is not made available to the public. ParsCit [21] is an open-source CRF-based citation parser that has been used by CiteSeer^x [4]. The model was reported to be trained on 200 reference strings sampled from computer science publications. Their performance is 95% for F1-score on Cora dataset.

Most systems (except for ParsCit and the system developed by Kramer et al [19]) described above were not available publicly. None of the systems developed was evaluated in the biomedical domain. Furthermore, none of the systems described above extracted author's surname and given name as we do in this study.

3 Methods

For citation parsing, we developed conditional random fields [22], which are probabilistic models that relate to the HMM. The CRF model has an advantage over the HMM in that it relaxes strong independent assumptions [23], and as a result, has shown to work well with

biomedical sequence data that frequently comes with variations. For example, research has found that CRFs performed the best in biomedical named-entity recognition tasks [24]. In the following, we first define the problem then describe our experiments on applying CRFs for citation parsing.

3.1 Problem Definition

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

3.2 Data

Creating training and testing data set that can be used for applying supervised machine learning is one of the key components of our work. We automatically extracted the training and testing data from articles in the PubMed Central (PMC). PMC is a free digital archive of biomedical and life sciences journals. As of February 2009, the PMC Open Access article collection has incorporated a total of 794 journals and 121,537 full-text articles. We found that the average number of articles per journal was 154, with standard deviation 1,302, the maximum number of articles per journal was 32,881, and the minimum was 1.

We found that for certain articles, PMC provides two formats of a full-text article: the parsed XML format from which we can extract the citation fields, and the HTML format from which we can extract the original full citation. Figure 1 and Example 4 show the two formats of a citation. The articles with both the XML and HTML representations were selected to extract citations to use as the data for this study.

[Example 4] Abbott NJ, Ronnback L, Hansson E (2006) Astrocyte-endothelial interactions at the blood&brain barrier. Nat Rev Neurosci 7:41–53

In order to ensure that our data was representative, we randomly selected 2% of the articles with the XML format from every journal. If a journal incorporated 49 articles or fewer, we randomly selected one article from that journal. Our selection resulted in a total of 2,988 articles, from which we were able to identify 672 articles that have the corresponding HTML citation format. Those 672 articles belong to 188 journals. We extracted full citations and their labeled fields from those 672 articles, which incorporated an average of 41 citations (Min 1, Max 333, and Standard Deviation 33). The total number of citations was 27,606, which were used as the tagged citation data for training and evaluation.

3.3 Condition Random Fields

We chose Mallet [25] as our CRF package for implementation. The overview of the citation parsing system is shown in Figure 2.

3.4 Learning Feature

We use Abner's [24] default features to train the CRF model. Abner's features capture the morphological characters of each token, for example, whether the token incorporates a capital letter, all capital letters, a digit, all digits, and other symbols (e.g., Roman and Greek characters and “-”), as well as the length of the token. We found that the features can be effective in discriminating different fields. For example, a combination of upper- and lower-case letters will lead to the detection of Author (“Hong”). More details about Abner features can be found in [26].

4 System and Evaluation

We used the 27,606 citations extracted from PMC for a 10-fold cross-validation to take the advantage of it: All observations are used for both training and validation, and each observation is used for validation exactly once. We split each citation into a series of tokens (each token contains a word or punctuation) and tagged them following the tag definition style used by the CoNLL shared task (<http://www.cnts.ua.ac.be/conll2002/ner/>). We defined the beginning of each field, inside the field, and OTHERS by “B-<fieldname>”, “I-<fieldname>” and “O”, respectively. For example, B-TITLE refers to the word at the beginning of a title. The tags of *FirstPage* and *LastPage* are FPAGE and LPAGE. Tages SN and GN denote *Surname* and *GivenName*.

We evaluated the test result for each field by recall, precision and F1-score (F1). With $F1 = (2 * Precision * Recall) / (Precision + Recall)$. Recall is the number of correctly predicted fields divided by the total number of annotated fields, and precision is the number of correctly predicted fields divided by the total number of predicted fields. These measures are per-entity.

4.1 Results

Table 1 shows the recall, precision, and F1-score of the 10-fold cross-validation. Figure 3 is the corresponding F1-score of each field and the overall score.

4.2 Error Analysis

In order to understand the source of errors, we examined those fields that were predicted to be wrong and summarized the patterns in them. For example, one pattern that our system wrongly predicted is that a word from the SOURCE field is a word from the TITLE field. We identified a total of 101 such patterns.

We found that 84.6% of the total number of errors fell into a total of 23 patterns. Figure 4 shows the number of instances of those patterns.

We further manually examined the source of errors for frequently-occurring patterns. We found that one source of error was introduced by the ambiguity of the period symbol “.”. For example, in SOURCE, the period not only represents the boundary of a source (e.g., “Nature Neuroscience.”), but also appears in the abbreviated version of journal titles (e.g., “Nat. Neurosci.”). The ambiguity resulted in some inconsistent tagging in the training data as well. Similarly, the ambiguity of the period applies to other fields, for example, the period that appears in the AUTHOR field (e.g., “Haynes, J. D.”).

Our system sometimes confuses SOURCE with TITLE. We speculate that this error was in part introduced by the fact that some full citations have a missing title, and some full citations do not separate the title from the source. For example, for “Colquhoun, D.; Sigworth, F.J. Fitting and statistics of single-channel records analysis Single-Channel Recording 1983.pp.191–263.”, the “Single-Channel Recording” is the source and directly follows the title.

5 Discussion

We reported in the related work that has shown different approaches for related tasks. We found that our approaches achieve the highest performance among all. We did not implement other approaches because it would be expensive for us to re-implement the systems.

Our system was trained in the bioscience domain and therefore, its performance remains to be evaluated when applying to other domains such as Math and Physics. On the other hand, the supervised machine-learning methods we developed are generalizable and the learning framework is robust, as long as training data is made available.

Although it has been explored little in existing work of citation parsing, different journals publish articles in different citation styles and such style information may be useful features to improve citation classification. However there are more than 6,000 journals in biomedical domain and therefore it would be a non-trivial task to incorporate such mapping into our supervised machine-learning framework. Furthermore, given a standalone article, it is not an easy task to extract its journal information and to obtain its journal style accordingly. Therefore, we did not pursue the citation style information in this study.

6 Conclusion and Future Work

In this study, we developed an open-source package that automatically parses a citation into its fields in full-text biomedical articles. We applied and implemented a supervised machine-learning system based on Conditional Random Fields (CRFs) for citation parsing and report 97.95% F1-score to parse a citation into a total of eight fields. Our results show that CRFs are efficient machine learning model for citation parsing. In future work we will integrate our citation parser into the framework of an autonomous citation indexing system.

Acknowledgments

The authors acknowledge the support from the National Library of Medicine to Hong Yu, grant number 1R01LM009836-01A1. We would like to thank Lamont Antieau for proofreading the manuscript. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the NIH views.

References

1. Hersh, W.; Cohen, AM.; Roberts, P.; Rekapalli, HK. TREC 2006 genomics track overview. The Fifteenth Text Retrieval Conference; 2006.
2. Garfield, E.; Merton, RK. Citation indexing: Its theory and application in science, technology, and humanities. Wiley; New York: 1979.
3. Tbahriti I, Chichester C, Lisacek F, Ruch P. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library. International Journal of Medical Informatics. 2006; 75:488–495. [PubMed: 16165395]
4. Voos H, Dagaev KS. Are All Citations Equal? Or, Did We Op. Cit. Your Idem? Journal of Academic Librarianship. 1976; 1:19–21.
5. Herlach G. Can Retrieval of Information from Citation Indexes Be Simplified? Journal of the American Society for Information Science. 1978; 29:308–10.
6. Garfield, E. Can citation indexing be automated. Statistical Association Methods for Mechanized Documentation, Symposium Proceedings; Washington. 1964. p. 189-192.
7. Schwartz, AS.; Hearst, M. Summarizing key concepts using citation sentences. Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL; 2006. p. 134-135.
8. Nakov, P.; Schwartz, A.; Hearst, M. Citances: Citation sentences for semantic analysis of bioscience text. Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics; 2004.
9. Borner, K.; Maru, JT.; Goldstone, RL. The simultaneous evolution of author and paper networks. National Acad Sciences; 2004.
10. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications. 2002; 311:590–614.

11. Giles, CL.; Bollacker, KD.; Lawrence, S. CiteSeer: An automatic citation indexing system. Proceedings of the third ACM conference on Digital libraries; New York, NY, USA: ACM; 1998. p. 89-98.
12. Lawrence S, Giles CL, Bollacker K. Digital libraries and autonomous citation indexing. IEEE computer. 1999
13. Wei, W.; King, I.; Lee, JH. Bibliographic Attributes Extraction with Layer-upon-Layer Tagging. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference; 2007;
14. Besagni, D.; Belaid, A. Citation recognition for scientific publications in digital libraries. Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop; 2004; p. 244-252.
15. Powley, B.; Dale, R. Evidence-based information extraction for high accuracy citation and author name identification. Proceedings of RIAO 2007: the 8th Conference on Large-Scale Semantic Access to Content; 2007.
16. Pasula H, Marthi B, Milch B, Russell S, Shpitser I. Identity uncertainty and citation matching. Advances in Neural Information Processing Systems. 2003;1425–1432.
17. Takasu, A. Bibliographic attribute extraction from erroneous references based on a statistical model. Digital Libraries, 2003. Proceedings. 2003 Joint Conference; 2003; p. 49-60.
18. Chen, CC.; Yang, KH.; Kao, HY.; Ho, JM. BibPro: A Citation Parser Based on Sequence Alignment Techniques.
19. Kramer, M.; Kaprykowski, H.; Keysers, D.; Breuel, T. Bibliographic meta-data extraction using probabilistic finite state transducers. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference; 2007;
20. Peng F, McCallum A. Information extraction from research papers using conditional random fields. Information Processing and Management. 2006; 42:963–979.
21. Councill IG, Giles CL, Kan MY. ParsCit: An open-source CRF reference string parsing package. Proceedings of LREC. 2008
22. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning; 2001. p. 282-289.
23. Wellner, B.; McCallum, A.; Peng, F.; Hay, M. An integrated, conditional model of information extraction and coreference with application to citation matching. Proceedings of the 20th conference on Uncertainty in artificial intelligence; Virginia, United States: AUAI Press Arlington; 2004. p. 593-601.
24. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Oxford Univ Press; 2005.
25. McCallum, AK. MALLET: A Machine Learning for Language Toolkit. OpenURL. 2002. <http://mallet.cs.umass.edu> website
26. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets. Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA); 2004. p. 104-107.

Biographies

Qing Zhang is currently a PhD student of Department of Computer Science, University of Wisconsin-Milwaukee, USA. Her research interests include machine learning, text mining, and information retrieval. She received a BS and MS from Beijing University of Aeronautics and Astronautics, Beijing, China. She is a student member of AMIA.

Yong-Gang Cao is currently a research associate in University of Wisconsin-Milwaukee. He was an associate research in Microsoft Research Asia. He received his PhD from Beihang University in 2006. His research interests include machine learning, text mining, information retrieval, and nature language processing.

Hong Yu is currently an associate professor of University of Wisconsin-Milwaukee. She received her PhD from Columbia University, USA. Her research interests include text-mining, knowledge representation, and user-centric systems.


```

<ref>
  <authorGroup>
    <author>
      <surName>Abbott</surName>
      <givenName>RD</givenName>
    </author>
    <author>
      <surName>Roushock</surName>
      <givenName>L</givenName>
    </author>
    <author>
      <surName>Shannon</surName>
      <givenName>E</givenName>
    </author>
  </authorGroup>
  <articleTitle>Anticrypte-endothelial interactions
    at the blood-brain barrier</ArticleTitle>
  <Year>2006</Year>
  <PubId></PubId>
  <CitationId></CitationId>
  <isMatched>false</isMatched>
  <sourceMed>Dev Neurosci</source>
  <volume>7</volume>
  <epage>41</epage>
  <ipage>53</ipage>
</ref>

```

Figure 1.
Gold standard XML.

Figure 2.
The overview of the citation parsing system.

Figure 3.
The F1-score of each field.

Figure 4.
The number of wrongly predicted fields by pattern.

Table 1

Recall, Precision, and F1-score for automatic citation field recognition

	Precision	Recall	F1
Title	95.46%	95.24%	95.35%
Source	87.22%	87.03%	87.13%
Year	99.64%	99.77%	99.71%
Surname	99.51%	99.77%	99.64%
GivenName	97.94%	97.97%	97.96%
Volumn	99.15%	99.32%	99.24%
FirstPage	99.66%	99.44%	99.55%
LastPage	99.91%	99.40%	99.65%
OVERALL	97.94%	97.96%	97.95%