A New Dataset for Fine-Grained Citation Field Extraction

Sam Anzaroot ANZAROOT@CS.UMASS.EDU
Andrew McCallum MCCALLUM@CS.UMASS.EDU
University of Massachusetts Amherst, Amherst, MA 01003

Abstract

Citation field extraction entails segmenting a citation string into its constituent parts, such as title, authors, publisher and year. Despite the importance of this task, there is a lack of well-annotated citation data. This paper presents a new labeled dataset for citation extraction that, in comparison to the previous standard dataset, exceeds four-times more data, supplies detailed nested labels rather than coarse-grained flat labels, and is derived from four different academic fields rather than one. We describe our new dataset in detail, and provide baseline experimental results from a state-of-the-art extraction method.

1. Introduction

Building tools that collect and organize research literature can provide insight into the landscape and process of science, and help individual researchers be more efficient. For example, analysis of citation graphs between papers can enable automated clustering for discovering trends in scientific sub-communities and can assist researchers in finding related work.

Sometimes such bibliographic data is provided in prestructured form, but often the case that data is supplied only in unstructured full text. In the unstructured case, reference sections of papers must be located, citations segmented from each other, citation fields must be extracted from within each citation, and the citations much be disambiguated. In this paper we concern ourselves with citation field extraction. Many citations include fields such as multiple author names (first, middle and last), paper title, journal name, volume, number, publisher, and year. Some also include publication status, web address, organization names, thesis indicators, postal addresses, and indication of publication language. Effective analysis requires extracting these fields accurately. Although the task may seem straightforward, truly high-accuracy citation field extraction has been elusive. Real-world citation strings are replete with wide variety and odd exceptions to common preconceptions about their simplicity. This irregularity makes rule-based methods brittle, and machine learning methods have become the tool of choice for citation field extraction. However, high-accuracy machine learning typically requires substantial labeled data.

The most widely used labeled data in citation field extraction is the CORA Field Extraction dataset (Seymore et al., 1999). Unfortunately it has numerous weaknesses. The dataset is small, containing only 500 citations. It has labels only for coarse-grained fields; for example it has a monolithic authors field, not labels indicating separate authors, nor first, middle and last names. Finally the dataset consists of citations only from within the field of computer science.

This paper presents a new labeled dataset for citation field extraction and provides baseline experimental results from a state-of-the-art method. The new data set contains over 1800 citations gathered from across physics, mathematics, computer science, and quantitative biology—all labeled with both fine-grained fields and coarse-grained field agglomerations.

A state-of-the-art linear-chain conditional random field trained and tested on subsets of this data achieves 95% token-level F1 and 91% field-level F1. In ongoing work we are developing more advanced methods with higher accuracy.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

2. Current state-of-the-art citation extraction techniques

The oldest method for citation field extraction is manual creation of logical rules (Jewell, 2000; Giles et al., 1998; Ding et al., 1999). For example (Ding et al., 1999) matches input citation strings against hand-designed patterns, which were devised by human analysis of citation from 43 journal papers. Such rule-based systems are not typically resilient to the wide variety of citation styles and formatting exceptions because the relatively small number of human generated patterns do not have broad coverage. Furthermore, these systems are difficult to transfer to new domains since new patterns must be manually devised for each variation in citation style.

Machine learning alleviates both these issues by learning weights that combine the evidence of many high coverage features, and by automating the creation of new models from domain-specific labeled training data.

Hidden Markov models (HMMs) are a commonly used in machine learning for segmentation and labeling tasks. HMMs provide a generative model over pairs of input variables and label variables in a sequence. In biliometrics, HMMs were originally shown to be useful for the task of extracting various fields (such as title, author, institution, abstract, etc) from the headers of research papers (Seymore et al., 1999). HMMs were later used for citation field extraction trained on the CORA dataset (Hetzner, 2008).

In many information extraction tasks conditional random fields (CRFs) (Lafferty et al., 2001) have replaced HMMs since CRFs offer more flexibility in the design of input features. CRFs were first applied to citation field extraction by (Peng & McCallum, 2004), who showed them to provide higher accuracy than HMMs on supervised citation field extraction, improving macro token level F1 for citation extraction on CORA from HMM's 86.6 to 91.5.

Others researchers have replicated the result of (Peng & McCallum, 2004) on the CORA dataset (Councill et al., 2008). However, when a model trained on CORA is evaluated on a new set of randomly-sampled computer science citations, the F1 score drops significantly (Councill et al., 2008). Our own experimental results described below paper confirm this outcome. We conclude that the CORA dataset is not representative of natural variability in citation style—even within the field of computer science.

Other research efforts address the tasks of citation field extraction and disambiguation jointly (Poon & Domingos, 2007; Singh et al., 2009). Here disambiguation refers to clustering citations that refer to the same paper. By modeling both segmentation and disambiguation together these systems strive to avoid cascading errors and to share strength between the tasks. For example, joint inference can improve segmentation by using information from one unambiguous title field to help infer accurate extraction within a more ambiguous coreferent citation string.

Two citation datasets have been used for such joint inference: (1) the CORA Coreference dataset (distinct from the CORA Field Extraction dataset) (Bilenko & Mooney, 2003) and the CiteSeer dataset (Lawrence et al., 1999). Each contains both annotated field extractions and disambiguation information. CORA Coreference and CiteSeer contain 1295 and 1563 citations respectively. Despite having field extraction information, they are not good datasets for training supervised field extraction classifiers since the citations within each dataset have little variability. For instance, the 1295 citations in CORA coreference are variants of only 134 papers, while CiteSeer contains citations from 785 research papers in only four subdomains of artificial intelligence.

Joint modeling has been explored in two frameworks: Markov logic networks (Poon & Domingos, 2007) and imperatively-defined factor graphs (Singh et al., 2009). Both of these frameworks use factor graphs with joint factors to allow bi-directional information to flow between the two tasks. The later more tightly couples field extraction and disambiguation by employing factors that use the intermediate output of the field extraction, and shows a 0.21 absolute improvement in field-level F1 on field extraction.

The CORA Coreference dataset is segmented even more coarsely than the Cora Citation Field Extraction dataset. For example, not only is it missing separation among authors, but it collapses many fields into a fused "venue" field. Both datasets contain citations exclusively from within computer science.

Our new UMass Citation Field Extraction Dataset offers advances by (1) providing both coarse- and fine-grained labeled field segmentation, (2) including citation data from multiple scientific disciplines, and (3) assembling significantly more data.

3. Dataset

In May 2012 we collected 5,000 research papers in PDF format from ArXiv.org, comprising 1,250 papers each from its sections on physics, mathematics, computer science and quantitative biology. The papers represent a variety of formats and styles, including journal pre-prints, conference papers and technical reports. Text and layout information were extracted using our custom-improved pdf2text system. Five citations per PDF were then manually extracted from 1200 of those papers, resulting in 6,000 unlabeled citation strings. Of these, 1829 citation strings have been labeled to date.

Each of these citation strings is labeled in a hierarchical manner, demarcating both coarse-grain labeled segments, as well as fine-grain labeled segments within. The coarse-grained segment labels are: ref-marker, authors, title, venue, date and ref-id; the list of fine-grained segment labels (as well as descriptions of both) is given below.

The coarse-grain segment labels are:

ref-marker A marker in the citation for referencing the citation in the paper.

authors A list of the authors in a citation.

title The title of a citation.

venue Description of where the cited information was published, including volume and page information, editors, etc.

date The date the cited work was published.

reference-id Any extra global document identifier, such as arxiv.org ids, or DOIs.

A venue label may contain the following fine-grain segment labels labels:

note Any plain text note about the citation, For example, list of thesis supervisors, book distributors, or the text “and references therein”.

web A web address mentioned in an citation.

status The current status of the publication, e.g. in preparation, submitted, accepted, revised, available.

language Information about the language of the cited work. Can be either that the referenced item is in a specific language (e.g. in French) or translation information (e.g. translated by John Smith).

booktitle The name of a book or conference proceedings in which an article is published.

date The date the venue of cited work was published.

address The location of a conference, or of a publisher.

pages The pages on which the article appears in book or proceedings.

organization The sponsoring organization of a conference.

volume The volume of journal or conference in which the cited work appears.

number Issue number of the article.

publisher The publisher of the journal, conference, book etc.

editor The list of editors who edited the journal.

tech The words describing the tech report or type of unpublished material with possible tech report number, e.g. Unpublished manuscript, ArXiv e-prints, preprint, Personal Communication.

institution Organization that publishes the tech report.

series The name of the series in which the book being cited is published.

chapter The chapter in the book the citation is referencing.

thesis The part of the citation mentioning that the cited work is a thesis, e.g PhD Thesis.

school The university that published the thesis.

department The department that published the thesis.

Both editor labels and author labels may contain **person** labeled segments, which contain one person’s name. The person segment can then include the following labeled sub-segments:

person-first A first name or initial of a person.

person-middle A middle name or initial of a person.

person-last A last name of a person.

person-affix An affix of a person, e.g. Jr., Sr.

A date segment may also include **year** and **month** labeled sub-segments.

This dataset is available at http://iesl.cs.umass.edu/data/umasscitationfield

Two example hierarchical labeled citations (both from physics) can be seen in Figures 1 and 2. These examples illustrate a common practice in physics citations: they do not contain the title of the paper, only volume, page number, and year following the journal name. Thus all the identifying information is contained in a venue section. In rare cases, there are multiple venue sections in one citation. The citation in Figure 2 depicts the citation of an article contained in a book. Such examples often contain more variability in formats. In this case, the article in the book is referenced only by the name of the author. In other cases, the page number, volume number, and article number are included as well.

4. State-of-the-art model

Here we apply a common state-of-the-art information extraction method—linear-chain conditional random fields—to our new dataset in order to provide baseline experimental results and corresponding error analysis.

4.1. Linear-chain conditional random fields

Conditional random fields (CRFs) are undirected graphical models trained to maximize the conditional probability of a set of target random variables Y given input variables, X. (Lafferty et al., 2001). The structure of the graphical model encodes the selected dependencies among variables. In a linear-chain CRF the target variables are connected in series, indicating a Markov assumption on the sequence, $y_1, y_2, ...y_N$.

For citation field extraction, the Y variables correspond to a sequence of field labels, one variable for each word or punctuation token in the citation. The X variables are features taken from the corresponding token and its context.

The parameters of the model are estimated by maximum likelihood, using the L-BFGS method of quasi-Newton optimization. Training-time inference is performed by "forward backward" belief propagation. Test-time inference is performed by Viterbi max-product belief propagation.

4.2. Labeling Schema

Our dataset's hierarchical class labels are transformed into single discrete variable values by conjoining the names of nested labeled regions. For example, a person's first name within an author segment has label value author/person/person-first, representing that it has a person-first label, within a person label, within an author label.

We represent the input with classes of features from both (Peng & McCallum, 2004) and (Councill et al., 2008). Our features are:

Word Features the word itself; the word (lowercased and not) with digits replaced by either "YEAR" or "NUM"; the first three characters in the word; does the word match "pages," "pp" or similar variants; does the word consist of a single character.

Case is the word capitalized; does it consist of all capital letters; a single capital letter; a capital letter followed by a period.

Numeric is the word a number; a number enclosed in parenthesis; does it contain a digit; end with digit.

Punctuation is the word is a punctuation mark; does it contain period; contain a dash.

Regular Expressions does the word match a regular expression indicating two numbers separated by a dash; indicating an email addresses; a website URL.

Counts the number of digits in the word and the number of alphanumeric characters in the token.

Location the relative location of word in citation, with a bin size of 12.

Possible Editor does the word "editor" (or variants thereof) appear within 10 words of the token.

Lexicons is the word in lexicons of author names, venues, and month names.

The lexicons are gathered from existing databases of author names and venues, including DBLP and a large collection of BibTeX records. The lexicons phrases that are matched against the citation word sequence up to length 14. The lexicon feature encodes whether the word is at the beginning, middle, or end of the phrase matched into the lexicon. Note that our current lexicons are predominately from resources within computer science.

Each word also contain binary conjunctions among all the word's features, as well as between the token's features and the features two tokens before and after the token.

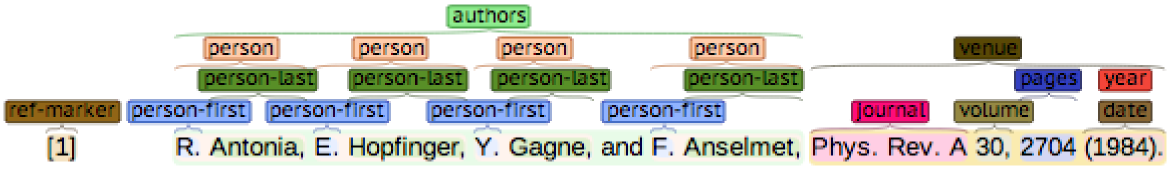

Figure 1. Example citation labeled in UMass dataset.

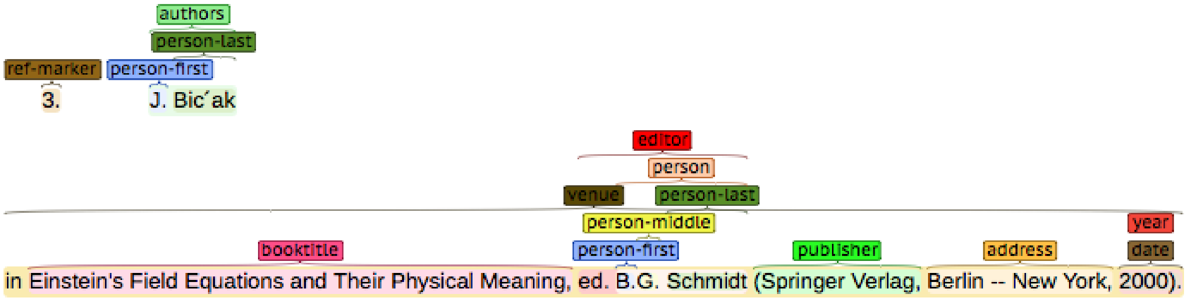

Figure 2. Example of citation of an article in a book.

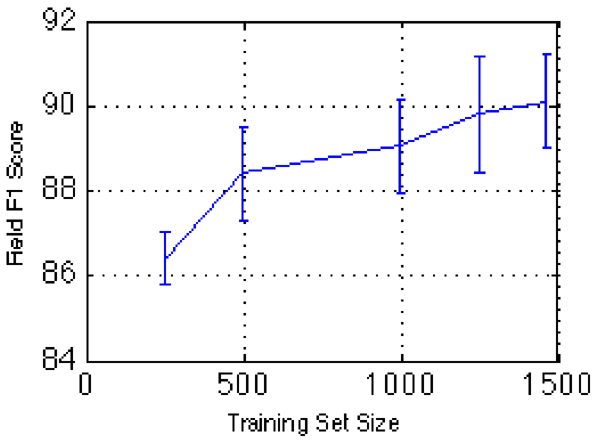

Figure 3. Performance of field level f1 trained with increasing amounts of data. Error bars are 1 sttddev.

5. State-of-the-art results on UMass dataset

5.1. Evaluation

We split the data 80/20 into training and testing data with 1463 and 366 citations respectively. The split is performed on the dataset-provided sequence of citations, in which the articles are ordered randomly, but the citations from one paper occur together. We report field-level and token-level F1 scores for every label. (F1 is the harmonic mean of precision and recall.) Token-level F1 is based on the number of individual word tokens that are given the correct label. Field-level F1 is based on the number of fields (such as title and publisher that are segmented and labeled perfectly; there is no partial credit for segment boundaries that close but not perfect.

In addition, we experiment with the effect of training set size on the overall field level f1 score by evaluating the test set performance with models estimated from increasing amounts of training data.

5.2. Results

Token-level and field-level results can be seen in table 2. Top level labels by themselves refer only to portions of the citation that are part of the coarse labels but not part of any fine-grain label, for example, the words "and" or "et. al." in the authors field and "In" in the venue field.

Figure 3, shows field-level F1 increasing as the amount of training data increases. Note that performance rises from approximately 86% with 300 training citations to 91% at about 1500 training citations. Note that previous publications on citation field extraction presented results of training on as few as 350 citations; thus we argue that this earlier work was operating in a strikingly data-poor environment.

In order to determine the ability of models trained

Table 1. Token level f1 with classifiers trained on Cora (C) and UMass (U) testing on their own testing sets and each other's testing sets. First block contains easily mapped fields.

Label	C on U	U on C	Reduction	U on U	C on U	Reduction
Author	99.40	97.41	1.99	96.31	95.28	1.04
Date	98.90	97.52	1.38	94.29	93.30	0.99
Journal	91.30	80.07	11.23	93.39	79.13	14.26
Location	87.20	93.70	-6.50	98.13	73.17	24.96
Pages	98.60	98.31	0.29	97.34	88.40	8.94
Publisher	76.10	85.36	-9.26	82.70	70.48	12.22
Title	98.30	97.32	0.98	95.13	91.77	3.36
Average			0.02			9.40
Booktitle	93.70	86.50	7.20	48.70	30.72	17.98
Volume	97.80	88.55	9.25	94.77	81.77	13.00
Tech	86.70	20.00	66.70	90.00	40.54	49.46
Note	80.80	66.67	14.13	72.41	37.78	34.64
Editor	87.70	69.84	17.86	68.75	35.04	33.71
Institution	94.00	28.00	66.00	75.00	36.17	38.83
Average All			13.94			19.49

on our dataset to generalize—especially in comparison to previous datasets—we compare models trained on CORA and trained on our dataset evaluated on the opposing datasets' test set. To accomplish this we map the labels of our dataset to the smaller set of CORA's labels. Note that some mappings are ambiguous; for instance, a volume in a series contained in the CORA dataset would be labeled as part of a booktitle, whereas it's a volume contained in a venue in the new dataset. In order to make an effective comparison we separate the labels into those whose mappings are ambiguous, and those that are not. As can be seen in table 1, on the unambiguously mapped labels, the model trained on our dataset has a small average reduction in F1 on CORA testset; however the model trained on CORA averages a 9.40 token-level F1 reduction on our new dataset's test set.

5.3. Error Analysis

As seen in table 2, author names are reliably extracted. Among the most important fields, booktitle and series have low scores.

Booktitles are difficult for many reasons. For example, they sometimes abut the title field in the citation, and can be mistakenly interpreted as a continuation of the title. Consider the citation string "Deformations of maps , Algebraic Curves and Projective Geometry." The booktitle begins after the comma and the title of the article precedes it, but our model labels the entirety as one title field. Additionally, often there is not enough context to distinguish between journal names and booktitles.

Sometimes there is also insufficient local context to distinguish between a book, in which the book's name should be labeled title, and an article in a book, in which the book's name should be labeled booktitle (and the article title labeled title). Consider the following example (in which the labeling of the editor is elided for brevity):

> ...[Monogenic forms on manifolds,]$_{\textbf{title}}$ in Z. Oziewicz et. al. (Eds.), [Spinors, Twistors, Clifford Algebras and Quantum Deformations,]$_{\textbf{title}}$ Kluwer Academic Publishers, 1993, [159-166.]$_{\textbf{pages}}$

Here the model has incorrectly labeled "Spinors , Twistors , Clifford Algebras and Quantum Deformations" a title, where it should be labeled a booktitle. Context that could have helped avoid this error—namely that there is already another title field in the citation, and that a limited range of pages is given—are not available given the Markov indepedent assumption in the linear-chain CRF. (For this reason and others are are currently researching extraction models that can leverage more global dependencies.)

Note that our multiple editor sub-fields provide significantly more detail than the CORA dataset, which have one label for the entire editor section. We have segments for each individual editor, as well as first, middle and last names for each editor, as seen in this example:

Table 2. Field & Token f1, precision and recall. Higher levels of the hierarchy are replaced by letters denoting the label.

Label	field f1	Precision	Recall	token f1	Precision	Recall
AUTHORS	96.13	94.05	98.31	98.00	97.78	98.21
A-P-PERSON-AFFIX	40.00	50.00	33.33	80.00	100.00	66.67
A-P-PERSON-FIRST	95.05	92.31	97.96	97.67	96.32	99.05
A-P-PERSON-LAST	95.20	92.58	97.97	97.69	96.32	99.09
A-P-PERSON-MIDDLE	92.84	89.72	96.19	96.23	95.22	97.26
DATE-YEAR	90.91	87.72	94.34	92.61	92.16	93.07
REF-MARKER	97.64	96.42	98.90	99.69	100.00	99.38
REFERENCE-ID	87.10	87.10	87.10	96.10	100.00	92.50
TITLE	87.07	84.96	89.30	97.09	95.13	99.14
VENUE	48.00	60.00	40.00	52.17	60.00	46.15
V-ADDRESS	85.71	94.29	78.57	92.11	98.13	86.78
V-BOOKTITLE	41.86	42.86	40.91	55.56	48.70	64.66
V-CATEGORY	0.00	100.00	0.00	0.00	100.00	0.00
V-CHAPTER	0.00	100.00	0.00	0.00	100.00	0.00
V-DATE-MONTH	62.50	50.00	83.33	87.50	77.78	100.00
V-DATE-YEAR	92.82	91.38	94.31	96.17	95.02	97.35
V-DEPARTMENT	100.00	100.00	100.00	100.00	100.00	100.00
V-EDITION	61.54	80.00	50.00	54.05	100.00	37.04
V-EDITOR	60.61	52.63	71.43	67.86	59.38	79.17
V-E-P-PERSON-FIRST	69.57	72.73	66.67	75.00	81.82	69.23
V-E-P-PERSON-LAST	72.00	81.82	64.29	70.59	75.00	66.67
V-E-P-PERSON-MIDDLE	72.73	80.00	66.67	72.73	80.00	66.67
V-INSTITUTION	30.77	100.00	18.18	26.67	100.00	15.38
V-JOURNAL	91.37	87.89	95.13	95.52	93.39	97.75
V-LANGUAGE	0.00	0.00	0.00	20.00	100.00	11.11
V-NOTE	0.00	100.00	0.00	0.00	100.00	0.00
V-NUMBER	74.07	75.00	73.17	88.89	91.95	86.02
V-ORGANIZATION	66.67	50.00	100.00	44.44	28.57	100.00
V-PAGES	94.51	91.81	97.36	98.45	97.34	99.58
V-PUBLISHER	77.23	76.47	78.00	87.93	82.70	93.87
V-REFERENCE-ID	72.73	80.00	66.67	81.08	88.24	75.00
V-SCHOOL	0.00	0.00	100.00	0.00	0.00	100.00
V-SERIES	25.00	40.00	18.18	25.00	58.82	15.87
V-STATUS	36.36	50.00	28.57	57.14	72.73	47.06
V-TECH	57.14	72.73	47.06	72.00	90.00	60.00
V-THESIS	0.00	100.00	0.00	0.00	100.00	0.00
V-VOLUME	93.91	91.61	96.32	95.90	95.34	96.46
V-WEB	0.00	100.00	0.00	0.00	100.00	0.00
OVERALL	91.16	90.17	92.16	94.79	94.08	95.50

[In [[M.]$_{\textbf{first}}$ [R.]$_{\textbf{middle}}$ [Farrally]$_{\textbf{last}}$]$_{\textbf{person}}$ & [[A.]$_{\textbf{first}}$ [J.]$_{\textbf{middle}}$ [Cochran]$_{\textbf{last}}$]$_{\textbf{person}}$ (Eds.) ,]$_{\textbf{editor}}$.

6. Conclusion and Future Work

This paper describes a new dataset for citation field extraction that is larger, more fine-grained and more varied across areas than existing widely-used datasets. We show that machine learning for citation field extraction is not data-saturated, in that more training data continues to improve model performance. For this reason we plan to continue labeling additional data, and will release augmented versions of this dataset in the future.

Through error analysis we also show the limitations of the Markov dependencies in linear-chain CRFs, and are currently developing models with more expressive dependency structure. We are also experimenting with methods of semi-supervised learning to leverage the vast quantities of readily available citation data. We also plan to release large quantities of unlabeled citation strings to support further research in semi-supervised methods on this data.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and the University of Massachusetts and in part by in part by IARPA via

DoI/NBC contract #D11PC20152. The U.S. Government is authorized to reproduce and distribute reprint for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the US government.

References

Bilenko, Mikhail and Mooney, Raymond J. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pp. 39–48, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. doi: 10.1145/956750.956759. URL http://doi.acm.org/10.1145/956750.956759.

Councill, Isaac G, Giles, C Lee, and Kan, Min-Yen. Parscit: An open-source crf reference string parsing package. In *Proceedings of LREC*, volume 2008, pp. 661–667. European Language Resources Association (ELRA), 2008.

Ding, Ying, Chowdhury, Gobinda, Foo, Schubert, et al. Template mining for the extraction of citation from digital documents. In *Proceedings of the Second Asian Digital Library Conference, Taiwan*, pp. 47–62, 1999.

Giles, C. Lee, Bollacker, Kurt D., and Lawrence, Steve. Citeseer: an automatic citation indexing system. In *International Conference on Digital Libraries*, pp. 89–98. ACM Press, 1998.

Hetzner, Erik. A simple method for citation metadata extraction using hidden markov models. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pp. 280–284. ACM, 2008.

Jewell, Michael. Paracite: An overview., 2000. URL http://paracite.eprints.org/docs/overview.html.

Lafferty, John D., McCallum, Andrew, and Pereira, Fernando C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

Lawrence, Steve, Giles, C. Lee, and Bollacker, Kurt D. Autonomous citation matching. In *Proceedings of the third annual conference on Autonomous Agents*, AGENTS '99, pp. 392–393, New York, NY, USA, 1999. ACM. ISBN 1-58113-066-X. doi: 10.1145/301136.301255. URL http://doi.acm.org/10.1145/301136.301255.

Peng, Fuchun and McCallum, Andrew. Accurate information extraction from research papers using conditional random fields. In Susan Dumais, Daniel Marcu and Roukos, Salim (eds.), *HLT-NAACL 2004: Main Proceedings*, pp. 329–336, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Poon, Hoifung and Domingos, Pedro. Joint inference in information extraction. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, AAAI'07, pp. 913–918. AAAI Press, 2007. ISBN 978-1-57735-323-2. URL http://dl.acm.org/citation.cfm?id=1619645.1619792.

Seymore, Kristie, McCallum, Andrew, Rosenfeld, Roni, et al. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, pp. 37–42, 1999.

Singh, Sameer, Schultz, Karl, and Mccallum, Andrew. Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pp. 414–429, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04173-0. doi: 10.1007/978-3-642-04174-7_27. URL http://dx.doi.org/10.1007/978-3-642-04174-7_27.