Metadata Extraction from Scholarly Articles

Ricardo Candeias

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract. Modern digital libraries of scholarly materials depend on the availability of quality metadata and, when performed manually, metadata annotation can be one of the most important costs in building a digital library. This paper describes PaperCut, a system based on supervised machine learning for extracting metadata fields from scholarly articles, using a stack of Conditional Random Fields models for labeling token sequences in the text of the articles, as belonging or not to the considered metadata fields. Experimental results show that PaperCut compares well with other previously published methods.

1 Introduction

After several years of massive digitization activities, current digital libraries hold large collections of scholarly materials. Moreover, most of the scholarly materials nowadays distributed by publishers are digital-born PDF documents. Still, most of these PDFs were designed for human consumption, lacking the descriptive metadata that is essential in supporting the advanced search and retrieval functionalities that are expected of modern digital libraries. Thus, an ongoing challenge within the Digital Libraries research community relates to the automatic extraction of metadata information from source documents such as PDF scholarly articles. This task raises important research challenges, since programatically recovering metadata fields from scholarly articles requires a machine to understand the structure of the strings in the different parts of the documents.

In each PDF article is found a set of metadata fields (e.g., author, title, year of publication, title of the journal, etc.) that are represented as surface strings, with implicit clues such as text position, punctuation and font size that can be used to assist in recovering the encoded information. While interpreting these fields is often straightforward for human readers, the sheer diversity of different document templates makes this process difficult to automate.

In this paper we describe the implementation of PaperCut, a system for metadata extraction from scholarly articles that uses machine learned models within a stacking framework, based on the formalism of Conditional Random Fields (CRF). While several methods that use machine learning have been proposed for this and for related metadata extraction problems, our innovative contribution lies in devising a rich set of features for this problem and combining it with a CRF stacking approach specific for metadata extraction from scholarly

articles. Experimental results show that PaperCut compares well against other previously published proposals.

The rest of this paper is organized as follows: Section 2 presents previous works concerning with metadata extraction. Section 3 presents the proposed method, describing both the stacking-based framework and the training of CRF models. Section 4 presents the experimental validation of the proposed method, describing the evaluation protocol and the obtained results. Finally, Section 5 presents our conclusions and points directions for future work.

2 Related work

Metadata extraction from scholarly articles has been described as *the most difficult task performed by an automated digital library system for research papers* [1]. The task remains an important challenge for the Digital Libraries community, although several authors have proposed either machine learning approaches to address the task [5], or knowledge based approaches based on hand-tuned rules [8,6]. In this paper, we follow a supervised machine learning approach.

The work presented in this paper is closely related to a class of problems that are frequently referred to as sequence labeling tasks, which refer to the assignment of labels to sequences of observations. Such labeling problems are common in many different fields, including computational natural language processing (e.g., tasks of part-of-speech tagging or named entity recognition) and information extraction (e.g., the task of metadata extraction from reference strings or from entire documents). Many different models have been proposed to tackle the sequence labeling task, with Conditional Random Fields (CRFs) currently being one of the most effective formalisms [7]. In this paper, we use the Conditional Random Fields (CRFs) formalism to model the process of assigning labels, corresponding to metadata classes, to sequences of textual tokens, learning such models from labeled training data and latter performing inference over unseen data that can later be applied to unseen data.

In the context of digital libraries, CRF models have been used in numerous applications, most notably metadata extraction tasks. The choice of this formalism can be justified by the work of Peng and McCallum [12], in which CRF models were shown to perform better than Hidden Markov Models [14,16] or Support Vector Machines [5]. Examples of systems using CRFs include CiteSeerX [1], a search engine for scientific literature, ArnetMiner [17], an academic social network search system, and ParsCit [2], a reference string parsing software package which is also incorporated in CiteSeerX.

The work made in the context of the ParsCit project is indeed the most relevant related work to the context of this paper. Two specific tasks have been addressed in the context of this project, namely (i) extraction and parsing of reference strings, i.e. the text strings in the bibliography or reference section of a published work that refer to a unique previously published document [2], and (ii) logical structure parsing of scientific documents, i.e. identifying the hierarchy of logical components such as titles, authors, affiliations, abstracts and sections [9].

Both tasks were addressed through as supervised machine learning, using Conditional Random Fields as the underlying model. In this paper, we follow on the original ideas behind the ParsCit project, by proposing a unified approach for metadata extraction from scholarly papers which includes the parsing of reference strings together with the logical structure of the articles, also using the formalism of Conditional Random Fields, although with a richer set of features.

3 Metadata extraction from scholarly articles

We propose to address the problem of recognizing individual metadata fields in scholarly articles by transcoding what is essentially a chunking problem (i.e., the task of discovering the chunks of text that correspond to specific metadata fields) into a tagging problem (i.e., a task of assigning tags to each individual token in the text of the articles, according their belonging to a given metadata field or not), using the standard begin/in/out (BIO) encoding of chunkings as taggings. Thus, the tagging problem involves separate B, I and O tags for each of the the different types of chunks, i.e. for each different metadata field.

A discriminative model known as Conditional Random Fields offers an efficient and principled approach for addressing this sequence tagging problem [7,12]. In our approach, we learn such models from training data, using them to annotate the words given in a textual sequence according to the BIO tags, from which we then generate the chunks that correspond to specific metadata elements. A particular novelty introduced in this paper is the usage of different CRF models for specific logical segments of the scholarly papers, instead of a single model for the entire contents of a document. This will be further detailed in Section 3.1.

Conditional Random Fields are essentially undirected probabilistic graphical models (i.e., Markov networks) in which vertexes represent random variables and each edges represent a dependency between two variables, that are discriminatively-trained to maximize the conditional probability of a set of hidden tags $y =< y_1, \ldots, y_C >$ given a set of input tokens $x =< x_1, \ldots, x_C >$. This conditional distribution has the following form:

$$p_\Lambda(y|x) = \frac{1}{\sum_y \prod_{c=1}^{C} \phi_c(y_c, y_{c-1}, x, c; \Lambda)} \prod_{c=1}^{C} \phi_c(y_c, y_{c-1}, x, c; \Lambda) \qquad (1)$$

In the equation, ϕ are potential functions parametrized by Λ. Assuming ϕ_c factorizes a log-linear combination of features computed over the subsequence c, then $\phi_c(y_c, y_{c-1}, x, c; \Lambda) = exp(\sum_k \lambda_k f_k(y_c, y_{c-1}, x, c))$ where f is a set of arbitrary feature functions over the input, each having an associate model parameter λ_k. The feature functions can informally be thought of as measurements on the input sequence that partially determine the likelihood of each possible value for y_c. The parameter k represents the number of considered features and parameters $\Lambda = \{\lambda_k\}$ are a set of real-valued weights, estimated from labeled training data by maximizing the data likelihood function through stochastic gradient descent. Given a CRF model, finding the most probable sequence of hidden tags

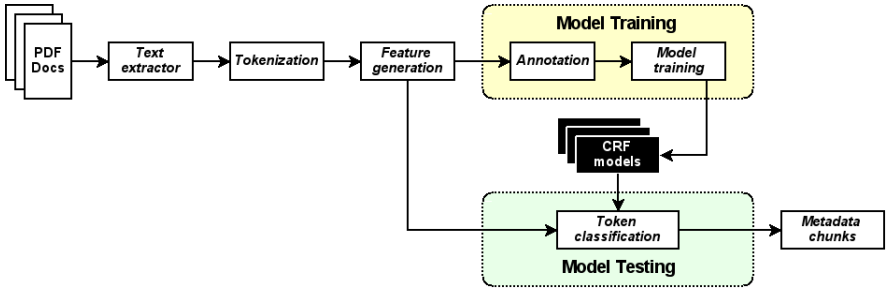

Fig. 1. The PaperCut approach to metadata extraction from scholarly papers.

given some observations can be made through the Viterbi algorithm, a form of dynamic programming applicable to sequence tagging. Typical applications are based on first-order chain CRFs, which make the first-order Markov assumption on the dependencies among y (i.e., the transitions between tags y depend only on the origin and destination). For more details about the formalism of CRF models, the reader should refer to the original paper by Lafferty et al. [7].

Besides the usage of CRF models, the proposed method for extracting metadata fields from PDF scholarly articles also involves a series of pre-processing operations. Figure 1 provides an illustration for the general approach.

The Portable Document Format (PDF) presents some problems to information extraction systems, mostly due to the fact that the order of textual objects in the PDF files does not always correspond to the reading order. Still, several tools for text extraction from PDF documents are nowadays available [3,4]. In our work, we used PdfToHtml[1] for extracting the text from the PDF files into XML representations. PdfToHtml is particularly interesting, since it supports the extraction of font and style information associated with the text.

After text extraction, the next steps relate to tokenization and token feature generation. The contents of each publication P are deterministically broken down into a sequence of tokens $\{p_1, p_2, \ldots, p_n\}$. This is made through the tokenization functionality available in the LingPipe[2] package. Next, each token is associated with a series of features, representing the evidence that will later be used for classifying the tokens as belonging or not to a particular metadata field (i.e., according to the values of these features, each token is to be assigned the correct label from a set of classes $C = \{c_1, c_2, \ldots, c_m\}$, representing begin/in/out positions for the considered metadata fields). Section 3.2 details our features, which encode information related to position, font style and textual contents.

Our approach for metadata extraction is based on supervised learning, thus requiring training data in the form of label class annotations for tokens described by features. We use ground-truth annotations (e.g., obtained with basis on the

[1] `http://sourceforge.net/projects/pdftohtml/`
[2] `http://alias-i.com/lingpipe/`

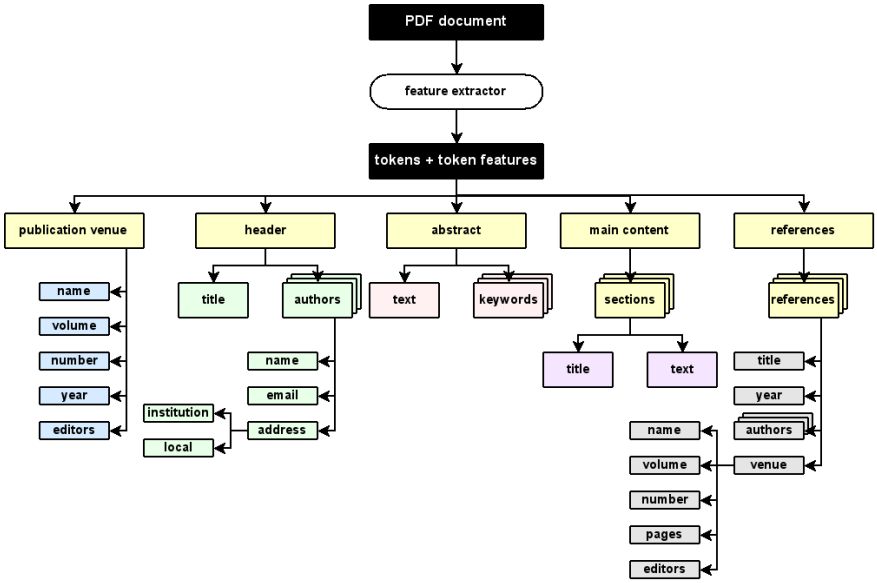

Fig. 2. Hierarchical structure and important metadata elements of scholarly papers.

LaTeX sources associated to academic articles in arXiv[3], and later manually revising the annotations) to learn a set of Conditional Random Fields (CRF) models, each responsible for annotating the tokens from a specific part of the document according to particular metadata fields. We used the linear-chain CRF implementation available through the LingPipe[4] package, in the development of the system. Section 3.1 details the training of these CRF models.

The trained CRF models can be used to assign the correct label to previously unseen tokens, with basis on the corresponding features. In the final step, after the tokens have been classified, we can recover the chunks of text corresponding to the individual metadata fields by taking the sequences of textual tokens annotated as belonging to the same metadata field.

3.1 Stacking of conditional random fields models

As previously stated, we use annotations revised by human experts to learn a set of CRF models, each responsible for annotating the tokens from a specific part of the document according to particular metadata fields. The general structure of a scholarly article, in terms of its main logical parts and important metadata elements, corresponds to the hierarchy given in Figure 2.

With basis on the hierarchical structure illustrated above, we propose to classify each token from a scholarly paper by using two distinct CRF models,

[3] http://arxiv.org/
[4] http://alias-i.com/lingpipe/demos/tutorial/crf/read-me.html

one corresponding to the logical structure of the paper (i.e., the five top different blocks shown in yellow over Figure 2) and another corresponding to the metadata fields that can be observed in the different logical blocks (i.e., the other colored blocks shown in Figure 2). Using this architecture based on stacking of multiple CRF models, instead of using a single CRF model considering a larger set of classes corresponding to the entire set of metadata elements, we hope to reduce the difficulty associated with model learning. Simpler models are less susceptible to data sparseness, thus requiring less training data. The training of the separate CRF models can also be made through data sets of different sizes, possibly from different origins, thus facilitating the gathering of training data.

Specifically, the first-level CRF model distinguishes between tokens that refer to (i) the publication venue of the paper, (ii) the header of the paper, (iii) the abstract and keywords of the paper, (iv) the individual sections that make up the content of the paper, and (v) the individual references that make up the bibliography of the paper. For each of the five classes considered in the first-level CRF model, we have a second-level CRF model that further segments the data. These models are as described bellow:

- A CRF model that analises the publication venue and distinguishes between the tokens that correspond to the name, volume, number, year and editors.
- A CRF model that analises the header of the article and distinguishes tokens that correspond to the title of the paper, the individual authors, the email addresses, the author institutions and the locations of these institutions.
- A CRF model that analises the abstract part of the paper and distinguishes tokens that belong to the text that composes the abstract from tokens that bellong to individual keywords associated to the paper.
- A CRF model that analises each of the sections of the paper and distinguishes tokens that belong to section titles from tokens that belong to the contents of the individual sections.
- A CRF model that analyzes each of the references in the paper and distinguishes tokens that belong to the title, publication year, individual authors, name of the journal, book or proceedings volume, the journal volume number, the issue number, the article's pages and the editors of the volume.

The different CRF models involve a set of common token features, which encode information relating to the textual contents, positions, and font styles. The full set of considered features is described in the next subsection.

3.2 The features involved in the conditional random fields models

The main advantage of CRF models over simpler approaches such as Hidden Markov Models (HMMs) comes from their modeling flexibility, which permits the feature functions to be complex, overlapping features of the input, without making additional assumptions on their inter-dependencies [7]. Taking inspiration from several previous works [12,2], we devised a rich set of features for the different layers of CRF models. The list of considered features is as follows:

- **Textual tokens :** The lowercased token identity for the target token.
- **Token size :** The length of a token.
- **Font size :** The font size for the target token, encoded as a numeric attribute which can take one of five different values and where lower/larger values correspond to smaller/larger font sizes.
- **Author name lexicon :** Wheter the target token appears in a list of author names which we compiled from sources such as the DBLP Computer Science Bibliography[5] or the Mathematics Genealogy Project[6].
- **Publication venue lexicon :** Wheter the target token appears in a list of names or abbreviations associated with popular conferences, journals and scientific publishers (e.g., tokens with a well-defined meaning such as ACM, IEEE, Elsevier, JCDL, SIGIR, TKDE, etc.) This lexicon was also compiled by us from sources such as the DBLP Computer Science Bibliography.
- **Character case :** Whether the target token is given in a capitalized, lowercased, uppercase or mixed-case form.
- **Dashes and dots :** Whether the target token contains at least one dash, one dot, or neither of the above.
- **Single character alone :** Whether the target token contains a single uppercase character or an initial such as *A.*.
- **Punctuation :** Whether the target token is a punctuation character.
- **Prefix :** Whether the target token begins with a specific pattern.
- **Suffix :** Whether the target token ends with a specific pattern.
- **Numeric :** Wheter the target token corresponds to a number.
- **Year :** Whether the target token corresponds to a sequence of four numbers, thus being likely to correspond to a year.
- **Month name lexicon:** Wheter the target token corresponds to the name of a month or a frequent abbreviation such as *Jan* or *Feb*.
- **Notes lexicon :** Wheter the target token appears in a lexicon containing words such as *appeared* or *submitted* that commonly appear in notes.
- **Abstract end lexicon :** Wheter the target token appears in a lexicon containing words such as *MSC, PACS* that are usually right after the end of the abstract.
- **Location lexicon :** Wheter the target token appears in a lexicon of location names, obtained from wikipedia.
- **Institution lexicon :** Wheter the target token appears in a lexicon containing words in different languages such as *institution* or *university*.
- **Email :** Wheter the token matches a regular expression for email addresses.
- **Containment of a digit :** Wheter the target token contains digits.
- **Absolute position :** The target token's absolute position in the text.
- **Position relative to abstract :** Whether the target token occurs before or after the first occurrence of the token with lowercased identity *abstract*.
- **Position relative to references :** Whether the target token occurs before or after the first occurrence of the token with lowercased identity *references*.

[5] `http://dblp.uni-trier.de/`

[6] `http://genealogy.math.ndsu.nodak.edu/`

- **Position relative to introduction :** Whether the target token occurs before or after the first occurrence of the token with lowercased identity *introduction*.
- **Position relative to keywords :** Whether the target token occurs before or after the first occurrence of the token with lowercased identity *keywords*.
- **Position relative to acknowledgments :** Whether the token occurs before or after the last occurrence of the token with lowercased identity *acknowledgments*.
- **Position relative to conclusion :** Whether the token occurs before or after the last occurrence of the token with lowercased identity *conclusion*.

The above features are computed for each token, as well as for the tokens within a 3-token window of the target token. Although first-order chain CRFs assume that the transitions between tags depend only on the origin and destination, the node features can be made to represent longer sequences or even global information relating to the tokens in the document. We also use a small set of edge features, combining some of the previous items with the previous token tag.

4 Evaluation experiments

The experimental validation of the system was based on a collection of 100 papers collected from arXiv, belonging to the domains of Mathematics or Computer Science, annotated according, annotated according to most of the metadata fields listed in Section 3.1 (all fields except the ones referring to publication venues, which almost never occurred in the considered papers). The annotations were first collected automatically from the information available in the LaTeX sources associated to the articles, and later manually revised. For the metadata fields specific to bibliographic references, we relied on a larger dataset that had already been used in previous information extraction experiments, namely the Cora[7] dataset that contained research paper citations with labeled segments according to the considered fields. Each token had thus a BIO annotation corresponding to its relation to a logical part of the document (i.e., a part corresponding to a description the header of the paper, the abstract and keywords, individual sections, and individual references) and a BIO annotation corresponding to the specific metadata fields that can be observed in the different logical parts. This data set was used for training and validation, using 10-fold cross validation and comparing the generated annotations against those proposed by the automated-method.

Table 1 and 2 gives the precision, recall and F_1 metrics for the individual metadata fields considered by the PaperCut system. The obtained results show that the proposed approach can extract the composing structure of an article, with an F_1 score over 97%. As it can be observed in terms of the model for extracting the individual fields of bibliographic references, the obtained results are comparable against those obtained at previous works that also had used the

[7] `http://www.cs.umass.edu/~mccallum/data.html`

	Token-based evaluation		
Large logical sections	Prec	Rec	F_1
Header	1.00	1.00	1.00
Abstract	0.98	0.98	0.98
Sections	0.99	0.99	0.99
References	0.99	0.97	0.98
Overall	0.97	0.98	0.97

Table 1. Logical sections token-based results.

Cora dataset, with an F_1 score over 82%, as shown in table 2. Overall, this model achieved some interesting results, except in some cases in which the low results may be attributed to the following reasons:

- Multiple different layouts that were associated to the papers used in the ground-truth collection;
- Failure by the tool responsible for extracting the PDF text, in terms of the recognition of the mathematical symbols, and accentuation, which led to a larger fragmentation of the text, thus losing the information of which tokens compose each line, and of which characters compose each token, that would be useful in terms of the section title and author name identification;
- The small size of the training dataset, which was caused by the difficulty in finding (i.e, caused by the lack of structure in the articles at arXiv that was required to use features like font), and labeling articles.
- In terms of the header model, due to the identification of institution and locations, these have shown a low result due to the name of these entities being written in the native language of the author, making the use of certain lexicon features unsuccessful in these cases.

5 Conclusions and future work

In this paper we presented PaperCut, a system for metadata extraction from scholarly articles that uses machine learned Conditional Random Fields (CRF) models within a processing framework based on stacking. While several methods for metadata extraction task have been proposed in the past, our contribution lies in devising a rich set of features for this problem, as well as in combining it with a framework based on stacking.

Despite the interesting results, there are also many ideas for future improvements. In my opinion, the main challenge is related to improving the validation mechanism, through the usage of a larger set of annotated scholarly articles, and by evaluating the system against articles from other scientific domains. We also have that the CRF models used in our experiments make the first-order Markov assumption on the dependencies among transitions, although it would be interesting to experiment with higher-order CRF models, capable of modeling transitions in longer token sequences [19]. The performance of our CRF models

	Token-based evaluation			Chunk-based evaluation		
Metadata fields	Prec	Rec	F_1	Prec	Rec	F_1
Title	0.87	0.97	0.91	0.69	0.68	0.68
Author Names	0.85	0.75	0.79	0.40	0.40	0.40
Author Emails	0.92	0.80	0.85	0.02	0.01	0.01
Institutions	0.80	0.73	0.76	0.01	0.02	0.02
Locations	0.87	0.74	0.78	0.03	0.05	0.07
Overall	0.86	0.80	0.82	0.21	0.22	0.22
Keywords	0.97	0.84	0.89	0.02	0.04	0.02
Text	0.99	0.99	0.99	-	-	-
Overall	0.98	0.92	0.94	-	-	-
Title	0.52	0.81	0.61	0.22	0.17	0.11
Text	0.99	0.99	0.99	-	-	-
Overall	0.76	0.90	0.80	-	-	-
Title	0.95	0.95	0.95	0.90	0.86	0.88
Year	0.97	0.96	0.97	0.92	0.76	0.84
Authors	0.98	1.00	0.99	0.98	1.00	0.99
Institution	0.90	0.90	0.90	0.83	0.71	0.77
Location	0.88	0.92	0.90	0.77	0.74	0.76
Book title	0.90	0.87	0.88	0.79	0.75	0.77
Journal	0.81	0.82	0.82	0.71	0.73	0.72
Editor	0.98	0.76	0.86	0.85	0.75	0.80
Overall	0.80	0.90	0.91	0.84	0.79	0.82
Overall	0.85	0.88	0.87	0.66	0.30	0.29

Table 2. Token-based and chunk-based evaluation results.

is also restricted by the availability of labeled training data, making it interesting to experiment with semi-supervised CRF models, capable of leveraging on small sets of hand-labeled data together with large amounts of unlabeled data [13].

The extracted metadata information also presents some important limitations. The proposed method is, for instance, not capable of associating the extracted e-mail addresses and institutions to the corresponding authors. For future work, it would be interesting to extend the proposed method in order to perform other related tasks besides the extraction of metadata from PDFs, such as citation matching and the normalization of author names, institutions and publication venues [11], the extraction of keywords from the textual contents [10], or the classification of article citations according to different semantic dimensions [18] (e.g., self-citations, citations to papers with similar approaches, citations to papers with contrasting approaches, etc.). Specifically in case of articles from the domain of mathematics, it would be interesting to combine the approach proposed here with advanced OCR techniques for recognizing formulas in the publications [15], since mathematical formulas can be used to support particularly interesting retrieval functionalities.

References

1. I. G. Councill, C. L. Giles, E. D. Iorio, M. Gori, M. Maggini, and A. Pucci. Towards next generation citeseer: A flexible architecture for digital library deployment. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, 2008.
2. I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the Language Resources and Evaluation Conference*, 2008.
3. H. D'ejean and J.-L. Meunier. A system for converting pdf documents into structured xml format. In *Proceedings of the 7th IAPR International Workshop on Document Analysis Systems*, 2006.
4. K. Hadjar, M. Rigamonti, D. Lalanne, and R. Ingold. Xed : a new tool for extracting hidden structures from electronic documents. In *Proceedings of the 1st International Conference on Document Image Analysis for Libraries*, 2004.
5. H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries*, 2003.
6. A. Ivanyukovich and M. Marchese. Unsupervised metadata extraction in scientific digital libraries using a-priori domain-specific knowledge. In *Semantic Web Applications and Perspectives*, 2006.
7. J. D. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.
8. S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 1999.
9. M.-T. Luong, T. D. Nguyen, and M.-Y. Kan. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems*, forthcoming.
10. T. D. Nguyen and M. Y. Kan. Keyphrase extraction in scientific publications. In *International Conference on Asian Digital Libraries*, 2007.
11. H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Proceedings of 15th Annual Conference on Neural Information Processing Systems*, 2002.
12. F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting*, 2004.
13. Y. Qi, P. Kuksa, R. Collobert, K. Sadamasa, K. Kavukcuoglu, and J. Weston. Semi-supervised sequence labeling with self-learned features. In *Proceedings of the 9th IEEE International Conference on Data Mining*, 2009.
14. K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*, 1999.
15. M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. Kanahori. INFTY : an integrated OCR system for mathematical documents. In *Proceedings of the ACM Symposium on Document Engineering*, 2003.
16. A. Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries*, 2003.

17. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.
18. S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2006.
19. N. Ye, W. S. Lee, H. L. Chieu, and D. Wu. Conditional random fields with high-order features for sequence labeling. In *Proceedings of 23rd Annual Conference on Neural Information Processing Systems*, 2009.