A Comparison of Layout based Bibliographic Metadata Extraction Techniques

Michael Granitzer
University of Passau
Passau, Germany
Michael.Granitzer
@uni-passau.de

Maya Hristakeva
Mendeley Ltd.
London, UK
maya.hristakeva
@mendeley.com

Robert Knight
Mendeley Ltd.
London, UK
robert.knight
@mendeley.com

Kris Jack
Mendeley Ltd.
London, UK
kris.jack
@mendeley.com

Roman Kern
Knowledge Management
Institute
Graz University of Technology
Graz, Austria
rkern@tugraz.at

ABSTRACT

Social research networks such as Mendeley and CiteULike offer various services for collaboratively managing bibliographic metadata. Compared with traditional libraries, metadata quality is of crucial importance in order to create a crowdsourced bibliographic catalog for search and browsing. Artifacts, in particular PDFs which are managed by the users of the social research networks, become one important metadata source and the starting point for creating a homogeneous, high quality, bibliographic catalog. Natural Language Processing and Information Extraction techniques have been employed to extract structured information from unstructured sources. However, given highly heterogeneous artifacts that cover a range of publication styles, stemming from different publication sources, and imperfect PDF processing tools, how accurate are metadata extraction methods in such real-world settings? This paper focuses on answering that question by investigating the use of Conditional Random Fields and Support Vector Machines on real-world data gathered from Mendeley and Linked-Data repositories. We compare style and content features on existing state-of-the-art methods on two newly created real-world data sets for metadata extraction. Our analysis shows that two-stage SVMs provide reasonable performance in solving the challenge of metadata extraction for crowdsourcing bibliographic metadata management.

Categories and Subject Descriptors

H.3.1 [**H.3.1 Content Analysis and Indexing**]: [Metadata Extraction]; I.2.7 [**Natural Language Processing**]: Text Analysis

WIMS'12 , June 13-15, 2012 Craiova, Romania

Keywords

Metadata Extraction, Research Papers, Layout features, Bibliographic Metadata

1. INTRODUCTION

Managing bibliographic metadata has always been subject to central authorities like libraries and publishers. However, with the advancement of social research networks like Mendeley[1] and social bookmarking tools like CiteULike[2], metadata management is becoming more and more decentralized. Decentralized metadata management requires intelligent tools in order to reach a high level of metadata quality that in turn permits good retrieval quality and supports a good user experience. This is particularly true for creating a consistent bibliographic catalog of tens of millions of bibliographic entries for search and browsing. Given that users manage their publication artifacts (i.e. the PDF of a publication) via social research networks, the combination of automated metadata extraction and robust de-duplication techniques permits the automatic assignment of correct bibliographic data to an uploaded PDF.

Supervised metadata extraction techniques can increase metadata quality and support crowdsourced bibliographic metadata management. The most popular machine learning techniques for extracting bibliographic data are Support Vector Machines (SVMs) [4, 5] Hidden Markov Models (HMMs) [10], and Conditional Random Fields (CRFs) [1, 9]. SVMs have been successfully used in practice for extracting bibliographic metadata by large scale systems such as CiteSeer[3] and Mendeley. Both CiteSeer and Mendeley utilize the algorithm and selected feature set presented by Han et al in [4, 5]. Previous work has also shown that CRFs perform better than HMMs. This difference has been primarily attributed to the fact that CRFs cope better with arbitrary, dependent features and joint inference over entire sequences [9].

[1] `http://www.mendeley.com`
[2] `http://www.citeulike.org/`
[3] `http://http://citeseerx.ist.psu.edu/`

However, previous work mostly focuses on evalutaion using a rather small data set (see [10]). A detailed comparision of the available state-of-the-art algorithms and systems on large, real world data available in social research networks (e.g. documents with extraction errors, pre-prints, different front matters and varying disciplines) is yet missing. Also, previous work uses syntactic and semantic features, but ignores layout information. One exception to this is [7], where layout features have been used to recover research paper structures. However, through their use of a commercial system, results are hard to reproduce.

Therefore, our work provides a comparison of different bibliographic metadata extraction methods from academic research papers on two noisy real-world data sets. We investigate the accuracy of existing metadata extraction systems namely ParsCit [1], the Mendeley Desktop [4] and our own Layout-based CRF. We analyze (i) their variance across scientific domains and corresponding domain-adaptation abilities, (ii) the impact of domain-independent layout features and (iii) the influence of post-processing heuristics. Our work contributes to the field by:

- providing a detailed comparison of today's state-of-the-art methods for metadata extraction;
- revealing cross-domain properties for the three investigated system as well as the impact of layout features;
- identifying the need for good post-processing heuristics in order to achieve reasonable results for extracting authors;
- conducting experiments on two new real-world data sets.

As a result of our analysis, the problem of metadata extraction for crowdsourced bibliographic metadata management can be solved best by a two-stage SVM approach combined with well engineered post-processing heuristics. Our work extends our previous work on comparing the Mendeley Desktop with ParsCit (see [3]) on a smaller dataset by adding layout based Conditional Random Felds, layout based features and by conducting all experiments on one additional dataset.

2. PROBLEM DEFINITION

Let **A** be the set of artifacts (e.g. PDF documents) and let **MD** be the set of metadata records. Each record contains several fields, denoted via the superscript, i.e.md_i^{title} denotes the title of the i-th metadata record. According to those two sets, the following properties can be identified:

- A metadata md_i^x is textually completely represented in artifact a_i, denoted as $md_i \preceq a_i$ (i.e. Title)
- A metadata md_i^x is textually partially represented in artifact a_i, denoted as $md_i \prec a_i$ (i.e. Author, Publishing media)
- A metadata md_i^x is textually not represented in artifact a_i, denoted as $md_i \perp a_i$ (i.e. Genre, field)

[4]http://www.mendeley.com/download-mendeley-desktop/

In this work we concentrate on the cases of metadata being partially and completely represented in artifacts, i.e. $md_i \prec a_i$ and $md_i \preceq a_i$ and formalize the following goal:

DEFINITION 1. *Let a_i be an artifact and md_i the associated metadata. Our goal now is to develop a function $extract(a_i) \rightarrow \hat{md}_i$, which given an artifact a_i automatically extracts the contained metadata $\hat{md}_i$. The error between extracted and the given metadata should be minimized, i.e. $error(\hat{md}_i, md_i) \rightarrow min$.*

3. METHODOLOGY AND ALGORITHMS

Most approaches to metadata extraction utilize a token based approach (e.g. decomposing artifacts into words). For each token, supervised classification models predict whether the token belongs to a particular metadata field or not. Accuracy and efficiency usually depend on: (i) the selected classification model; (ii) the feature sets used for describing tokens; (iii) the heuristics applied to resolving ambiguous annotations; and (iv) external knowledge encoded in the form of gazetteer lists or grammars.

Two different classification models can be distinguished: Sequence Labeling and standard Text Classification. Standard text classification utilizes a vector representation of syntactic and semantic context features surrounding a token, but ignores previous and subsequent classification decisions. Due to their ability to generalize well from sparse feature sets SVMs are often used as base classifiers[4]. Recent applications to bibliographic metadata extraction showed promising results with precision and recall values around 0.9%.

Sequence Labeling techniques like CRFs and HMMs are considered to be more powerful for Natural Language Processing and Information Extraction tasks due to their ability to take previous classification decisions into account. To keep the complexity of the model small, only the direct predecessor in sequences is generally considered, resulting in so-called linear-chained models [6].

Besides the classification algorithm, features used during training and classification greatly influence the accuracy and extraction quality. Previous work mostly considered features covering syntactic and semantic token information [4, 1], such as orthography, morphology and POS tagging. Dictionaries containing data such as publisher names, author forenames and place names have also been used to introduce external domain knowledge features. Although it seems to be promising to use layout information (e.g. font size and text positioning) such features have hardly been used for metadata extraction. In contrast to dictionary based features, layout features, if properly encoded, can be seen as domain independent (i.e. independent of the research field).

After tokens have been labeled, post-processing disambiguates multiple, potentially conflicting labelings. Post-processing steps usually rely on some sort of domain specific heuristic like for example that authors have to occure after the title and not before. Such heuristics can range from simple rules to well-formed grammars. Post-processing becomes increasingly important in environments with noisy data. Although it is more considered as an engineering task, incorporating a-priori knowledge before or after applying supervised

learning techniques can greatly influence accuracy and practical applicability.

In the following we present three systems for extracting bibliographic metadata, namely ParsCit, the Mendeley Desktop and a linear-chained CRF. The systems differ in terms of classification models, feature sets, heuristics and PDF extractors employed. These characteristics will allow us to evaluate these different properties on real-world data set. Table 1 provides an overview on the systems' properties, where Mendeley Desktop can be seen as most general and the layout-based CRF as most specialized.

3.1 Layout-based Conditional Random Fields

Sequence labeling techniques require a set of token sequences $\mathbf{S}_i$, where each token sequence is a sequence of tokens t and features f occurring in an artifact. Each token is annotated with a label to be predicted for a particular sequence. In our case the metadata field corresponds to that label. Training takes a set of annotated sequences to learn a corresponding model. The model is used afterwards for automatically annotating new, unknown sequences. Therefore, we converted artifacts $a_i \in \mathbf{A}$ into a sequence of d-tuples $s_{i,j} = (t_{i,j}, f_{1,i,j}, \ldots, f_{d-2,i,j}, field_{i,j})$ representing the token, the features and the label. Here $t_{i,j}$ represents the textual representation of token j in artifact a_i (words, bi-words etc.), $f_{l,i,j}$ represents the l^{th} feature associated with that token and $field_{i,j}$ the metadata field of that token (i.e. author, title or regular text). Type information differs whether a token marks the beginning of a field or its interior. In accordance with the CoNLL format, we used the prefix "B-" to denote the beginning of a metadata field, "I-" the interior parts of a metadata field and "O" to denote tokens without any field association.

In order to obtain the label $field_{i,j}$ of a sequence tuple, fuzzy string matching with the original metadata, e.g. the author and title strings in the bibliographic metadata, has been used. Basically, all fields of the original metadata md_i are converted into token sequences $\mathbf{S}_{md_i}$. These token sequences are matched against the token sequence $t_{i,j}$ of the artifact $a_i \in \mathbf{A}$. All sub sequences of tokens in the artifact where less than 20% of the tokens have been added, updated or removed with regards to the metadata token sequence $\mathbf{S}_{md_i}$ are annotated with label $field_{i,j}$. Ambiguous annotations, where two fields overlap on a token sequence, have been resolved by taking the longer metadata field as the correct label. Although we do not have the label information, this fuzzy matching procedure allows us to obtain annotations for training which are correct with high probability. As it has been shown in related research on the topic of Open Information Extraction, such heuristics and fuzzy matching strategies for identifying a probably correct ground truth allow to achieve good extraction accuracy [11].

Features are generated from layout information obtained during PDF analysis. In particular, for each token we used *orthography, font family, font size, font width, variance of the font width, average spacing between characters, x-position, y-position* as well as the *gradients of x and y positions* as features. Real-valued features have been made discreet for two reasons. First, the CRF implementation used did not allow real-valued features and second, in order to introduce document-relative features (e.g. font sizes of a token relative to the average font-size used in the document). Hence, the hypothesis is that features relative to a document generalize better over different publishing styles and provide more robust estimates against parsing errors. Rounding and binning into 7 bins have been used for feature discretization. Absolute values are used as absolute features while rank estimates are used as relative features. In total we obtained 34 different features.

CRFs do not only take the features of the token to be classified into account, but also features of tokens before and after that token [6]. In order to keep the feature space tractable, we restricted previous and subsequent features to a window of ± 2 tokens. Hence, annotating token $t_{i,t}$ takes features $f_{\cdot,i,t-2}, f_{\cdot,i,t-1}, f_{\cdot,i,t}, f_{\cdot,i,t+1}, f_{\cdot,i,t+2}$ into account. We do not combine features from different positions (e.g. $f_{1,i,t-2} + f_{1,i,t-1}$), with the exception of orthographic features.

Still, we obtained a very high (around 10^5-10^6) number of different features. Hence, we used the stochastic gradient descent training algorithm [2] implemented in the *crfsgd*[9] framework in order to scale our approach up on the data set. We used the standard parameters as suggested in [2].

3.2 ParsCit Metadata Extraction

ParsCit is one of today's metadata extraction forerunners and is based on CRFs that are tailored to the computer science domain. It employs the CRF++ implementation[10]. ParsCit already contains trained models and uses token identity, orthographic case, punctuation, numbers, locations and several dictionaries as features (see [1]). To have a fair comparison with previous experiments, we used ParsCit as it is and did not do any re-training on our data set. Since ParsCit does not have its own PDF to text converter, we used PDF Box[11] as in our own CRF approach to generate corresponding test examples.

3.3 Mendeley Desktop

The metadata extraction algorithm used by Mendeley Desktop relies on a two-stage SVM method as outlined in [4]. It treats metadata extraction from header text as a multi-class classification problem using SVMs. The idea is to first classify each line of the header text into title, author or other (e.g. multi-author) classes using text and formatting features. The line classification is then improved by using contextual information such as the predicted class labels of the neighboring lines assigned in the previous step. Finally, multi-author lines are segmented into a list of individual author names. This is done using a simple recursive descent parser which assumes that the line conforms to a simple punctuation-based grammar. In addition, the algorithm feature set is based on [4, 5] and uses: (i) character-level features; (ii) dictionary/word-list features (e.g. academic titles); (iii) layout features; (iv) independent line features (e.g. number of words on line); and (v) contextual line features

[9] `http://leon.bottou.org/projects/sgd`
[10] `http://crfpp.sourceforge.net/`
[11] `http://pdfbox.apache.org/`

System	Classification Model	Features	Heuristics	PDF Extractor
ParsCit	CRF++[5]	syntactic, semantic and dictionaries	author name normalization	PDF Box[6]
Mendeley Desktop	two-stage SVM [4]	syntactic, semantic, dictionaries and layout features	recursive descent parser	PDFNet [7]
Layout-based CRF	Linear Chain CRF trained with stochastic gradient descent [2]	layout features and orthographic features	none	PDF Box[8]

Table 1: Comparison of System Properties evaluated in our experiments

(e.g. font size relative to previous and next line).

The algorithm is implemented using the libsvm library [12] to train the SVM classifiers with an RBF kernel. In contrast to ParsCit and the layout-based CRF, Mendeley Desktop uses PDFNet [13] to extract text from imported pdf documents. The sequence annotation in the training set uses fuzzy string matching in combination with Levenshtein distance.

4. DATA SETS

Previous work on metadata extraction from academic research papers focused on the Computer Science domain and used rather small data sets with no layout information [4, 10]. In order to take advantage of layout information and to consider noise resulting from PDF to text conversion, we created two real-world test data sets: the e-prints; and the Mendeley data sets. The e-prints data set has been used during development, while the Mendeley data set has been used primarily for validation. Both reflect real-world data obtained from existing archives and social research networks. Fuzzy matching (see above) has been used to generate the ground truth. To keep the data set as clean as possible, we excluded artifacts with ambiguous annotations (e.g. overlapping fields) or incomplete metadata.

4.1 e-prints Data Set

Our first data set, the *e-prints Data Set*, has been created by crawling the e-prints RDF Repository provided by the RKB-Explorer project[14] and downloading all available pdfs. From this data set, we took all journals and conferences that have more than 10 assigned PDFs resulting in 2,452 PDFs plus metadata. For our experiments, the data set has been narrowed down to three groups of presumably similar publication styles by using regular expression patterns on the journal names. Three groups have been created: Physical Reviews (215 publications); the British Medical Journal (138 publications); and all publications belonging to IEEE (344 publications). The data set, including preprocessed, layout and sequence annotated data, can be downloaded from `http://team-project.tugraz.at/the-project/results/`.

Manual inspection of the data revealed particularly challenging aspects like (i) multiple articles contained in one PDF; (ii) front matters from institutional repositories making metadata occurrences more frequent but less consistent in style; and (iii) mismatching Metadata due to incorrect or abbreviated metadata fields (e.g. abbreviated title and forenames).

4.2 Mendeley Data Set

The data set consists of 20,672 publications sampled from the 20 million pdfs available in the Mendeley research network. The sample was chosen to reflect different sources and research fields, namely Nature (1,399 publications, 7 journals) as general science literature, ACM (266 publications, 11 journals) and IEEE (1,481 publications, 48 journals) as computer science literature, BMC (7,572 publications, 9 journals) as bio-medical literature, Physical Reviews (8,815 publications, 5 journals) as Physics literature as well as arXiv (1,006 publications, 1 journal) as an Open Access publisher with mixed literature. In total the data set consists of 81 different journals and proceedings with each journal having more than 20 artifacts.

Manual inspection of the training data showed that journals included in the groups for Physical Reviews and BMC are most consistent in style, while Nature, ACM and IEEE are least consistent in style. Over all groups multi-column and single column journals are mixed as well as pre-prints, together with published versions. Compared to the e-prints data set the Mendeley data set contains less noise.

5. RESULTS

In evaluating the results, we estimated precision ($Prec$) and recall (Rec) for author (subscript a) and title (subscript t) fields. An extracted metadata is considered as correctly identified if it has a Levenshtein similarity higher than 0.7 to the original metadata. We chose 0.7 since PDF extraction errors in titles hinder string equality (e.g. titles including chemical compounds). We used 5-fold cross-validation on the e-prints data set and 3-fold cross-validation on the Mendeley data set to obtain the precision/recall figures as shown in tables 2 and 3.[15] Furthermore, we give two box blots to illustrate the variance of extraction performance among different journals. Figure 1 compares the overall performance of the systems, while Figure 2 compares the performance over different groups, experiments and extracted fields in detail.

[12] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[13] `http://www.pdftron.com/pdfnet/`
[14] http://eprints.rkbexplorer.com/

[15] Due to the size of the Mendeley data set, 5 folded cross validation created a too large training set for the layout based CRF.

Experiment	$Prec_a$	Rec_a	$Prec_t$	Rec_t
$CRF^{absolute}_{Layout}$				
BMJ	0.52	0.21	0.77	0.72
IEEE	0.59	0.49	0.62	0.87
Physical Reviews	0.38	0.21	0.74	0.87
CRF^{all}_{Layout}				
BMJ	0.52	0.21	0.80	0.76
IEEE	0.61	0.49	0.63	0.85
Physical Reviews	0.36	0.16	0.83	0.87
ParsCit				
BMJ	0.30	0.21	0.29	0.21
IEEE	0.60	0.48	0.82	0.82
Physical Reviews	0.87	0.48	0.84	0.81
$Mendeley$				
BMJ	0.27	0.20	0.29	0.26
IEEE	0.53	0.27	0.73	0.67
Physical Reviews	0.78	0.41	0.81	0.54

Table 2: Results on the e-prints Data Set. Detailed results can be found as spreadsheet under [8]

Notation in Experiments.

Experiments for the layout based CRF are labeled as CRF_{layout}. The superscript "all" denotes that relative and absolute features have been used while the superscript "absolute" denotes the absence of relative features. ParsCit has not been re-trained, which allows us to estimate the cross-domain performance of a system perfectly tuned to Computer Science (see [1]). Mendeley Desktop allowed us to re-train the model using fuzzy sequence matching. Results are denoted as $Mendeley$ with subscript "trained" for re-training on the data sets. The original Mendeley Desktop comes with a model trained on around 1,000 publications from very different domains including physics and biology.

5.1 Results on the Development Corpus

Results on our development corpus, the e-prints data set, show that for title extraction layout features achieve stable performance over all publishers and achieve comparable results with the existing approaches. The pre-trained models from Mendeley and ParsCit performed particularly poorly on the medical domain. Hence, without retraining neither systems are able to adapt to different domains (i.e. research fields).

Title extraction seems to be less challenging than author extraction. In order to boost the CRF_{Layout} further we added text tokens as lexical features. While this improved author extraction between 4-13%, it had a less significant impact for title extraction (between 1-6%). We also conducted experiments using only font and position features individually, but combining them yielded the best results. Results on these runs are omitted for space reasons. In order to see the impact of only layout features, we did not use any lexical feature for runs on the Mendeley data set.

Experiments on the development corpus also showed that relative feature encoding did not provide much benefit for the CRF, which seems to be counter intuitive. Moreover, author extraction performance varies largely among the systems. Layout features play a role (compare the CRF with ParsCit and Mendeley on BMJ in table 2), but cannot outperform well-tuned heuristics (Compare CRF with ParsCit and Mendeley on IEEE and Physical Reviews in table 2)

Overall, extraction performance is acceptable for titles in computer science and physics papers, but not so strong for author recall in all groups. However, systems tend to be optimized for the corpora under which they have been developed and therefore CRF_{Layout} results maybe over estimates. Hence, we focus on validating these findings on the Mendeley data set and report the results in the following subsections.

5.2 Results on the Mendeley Data Set

5.2.1 Title vs. Author Extraction

Similar to the development corpus, the title extraction problem can be solved with higher accuracy than author extraction one. While title extraction can be considered as fairly good, this is not true for author extraction. Recall figures are especially low. On average, author extraction recall ranges between 0.2 for CRF_{Layout} and 0.8 for $Mendeley_{trained}$ with similar variance over the different journals (compare Figure 1). Training Mendeley improves author extraction recall and reduces recall-variance among journals, but has negligible effects on author extraction precision. The post-processing heuristic may be responsible for this result. Both ParsCit and Mendeley apply post-processing which has, compared to $CRF - layout$, a high impact on the precision of author extraction on the Mendeley Data set. We take this as a hint on using post-processing heuristic to improve the extraction of structured metadata like names.

5.2.2 Layout features

Although CRF_{Layout} achieved reasonable performance on the e-prints data set, especially precision values for title extraction are significantly lower than for ParsCit and Mendeley on the Mendeley data set. Also, layout features alone are insufficient for author extraction, which is particularly shown by the very low recall figures.

All runs with layout features show higher variance across journal groups compared to Mendeley or ParsCit. This might depend on the larger heterogeneity in styles per journal group compared to the e-prints data set, which is especially true for Nature (see Figure 2 bottom-left). Nature has large variations in style across the different journals and hence CRF_{Layout} is hardly able to learn a good model.

5.2.3 Classification Models

Although CRFs maintain more powerful information extraction models, the two-stage SVM outperforms CRFs on metadata extraction. Clearly CRF_{Layout} lacks syntactic and semantic features, but ParsCit did not perform better than the standard Mendeley model. Comparing the standard Mendeley model and ParsCit on the Computer Science domains shows equal performance on IEEE papers and better results for Mendeley on ACM. Hence, ParsCit seems to be highly specialized on IEEE.

Similarly, re-training Mendeley's two-stage SVM shows an

Experiment/Group	$Prec_a$	Rec_a	$Prec_t$	Rec_t	Experiment/Group	$Prec_a$	Rec_a	$Prec_t$	Rec_t
Parscit	**0.77**	**0.50**	**0.75**	**0.69**	*Mendeley*	**0.79**	**0.54**	**0.86**	**0.81**
ACM	0.77	0.53	0.78	0.78	ACM	0.86	0.61	0.90	0.85
arXiv	0.89	0.58	0.90	0.61	arXiv	0.83	0.50	0.91	0.68
BMC	0.74	0.32	0.26	0.22	BMC	0.94	0.86	0.76	0.74
IEEE	0.75	0.55	0.87	0.81	IEEE	0.73	0.47	0.86	0.81
Nature	0.86	0.33	0.45	0.40	Nature	0.91	0.60	0.91	0.84
Physical Reviews	0.89	0.42	0.88	0.53	Physical Reviews	0.84	0.43	0.97	0.90
CRF^{all}_{Layout}	**0.50**	**0.21**	**0.73**	**0.73**	$CRF^{absolut}_{Layout}$	**0.48**	**0.20**	**0.72**	**0.72**
ACM	0.43	0.14	0.73	0.66	ACM	0.38	0.14	0.66	0.68
arXiv	0.53	0.27	0.75	0.77	arXiv	0.52	0.26	0.76	0.78
IEEE	0.52	0.24	0.78	0.79	IEEE	0.51	0.23	0.79	0.78
Nature	0.48	0.08	0.34	0.42	Nature	0.48	0.08	0.35	0.43
$Mendeley_{trained}$	**0.81**	**0.62**	**0.94**	**0.91**					
ACM	0.84	0.69	0.88	0.84					
arXiv	0.84	0.63	0.94	0.92					
BMC	0.94	0.91	0.96	0.95					
IEEE	0.75	0.54	0.94	0.91					
Nature	0.91	0.67	0.94	0.92					
Physical Reviews	0.85	0.59	0.99	0.98					
all groups	0.81	0.61	0.93	0.91					

Table 3: Extraction results on the Mendeley Data Set for the different systems. Bold rows show the average performance of the system, while the other rows show the performance on a particular groups.

impressive precision and recall improvement for title extraction and a good recall improvement for author extraction. This is especially true for BMC and arXiv. With an average precision of 0.94 and an average recall of 0.91 Mendeley provides satisfactory accuracy. Also, training Mendeley on all journals independent of the group to which they belong does not lead to poorer performance (see table 3 last line "all groups"). Hence, the SVMs provide enough model complexity to scale in terms of data set complexity.

Also variances among journal groups are lower than for ParsCit for title extraction and equal for author extraction. Hence, Mendeley's two stage SVM is reliable across groups. One reason for the more reliable SVM results may lie in the illposed line breaks in sequences extracted from PDFs as follows: Since every line break marks the beginning of a new sequence, sequences belonging to the same metadata field but ranging over several lines are broken apart. This happens especially for longer titles or formats with one-column titles like "Nature". While the SVM context model can recover from such bad splits, CRFs cannot. CRFs do not consider labeling information from previous sequences and hence may more easily fail in finding bad splits.

SVMs also outperform CRFs in terms of training time. Although we used stochastic gradient descent, the fastest training method currently known for CRFs, we could not conduct experiments on larger corpora, i.e. BMC and Physical Reviews, using our current feature setting. Also on a reduced feature set, training time has been inferior compared to Mendeley.

5.2.4 Domain Adaptation

For evaluating domain adaptation, we organized the data set in journal groups belonging to different research fields. The boxplots in Figure 2 show our experiments factored by those journal groups. Author extraction tends to be domain independent for ParsCit and Mendeley. In particular, the precision values achieved are stable across groups and have very low variance. The explanation for this result most probably lies in the semantic features that exploit dictionaries of person names and the post-processing of author names utilized by both Mendeley and ParsCit.

With respect to title extraction, performance varies considerably across domains. ParsCit's domain focus on Computer Science becomes clear (see Figure 2 bottom right). Precision and recall are very high for ACM and IEEE, but low for all other groups. Although Mendeley's standard model has been trained on only 1,000 PDFs from varying research fields, it shows surprisingly good performance and stability across groups. A performance that can be improved by training Mendeley on this data set (see Figure 2 bottom middle). Nevertheless, comparing ParsCit to Mendeley seems to support the hypothesis in favor of generating more heterogeneous than homogeneous data sets for bootstrapping metadata extraction.

6. CONCLUSION

In our work, we compared three different systems for extracting bibliographic metadata from real-world PDF artifacts. Together with strong de-duplication techniques, we consider the two-stage SVM approach combined with well-engineered heuristics as good solution for crowdsourced bibliographic metadata management.

Some properties and insights point to interesting future research directions. First, extending Conditional Random Fields to recover from imperfect sequences by taking labelling information from previous sequences into account seem to be an interesting idea for real-world, noisy data. Second, layout features showed potential. Although they performed below expectation, their domain independence

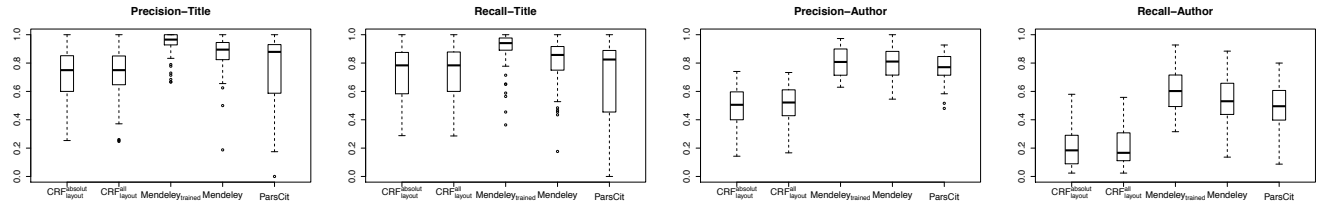

Figure 1: Boxplots on the precision/recall values for every system on the 82 journals of the Mendeley data set

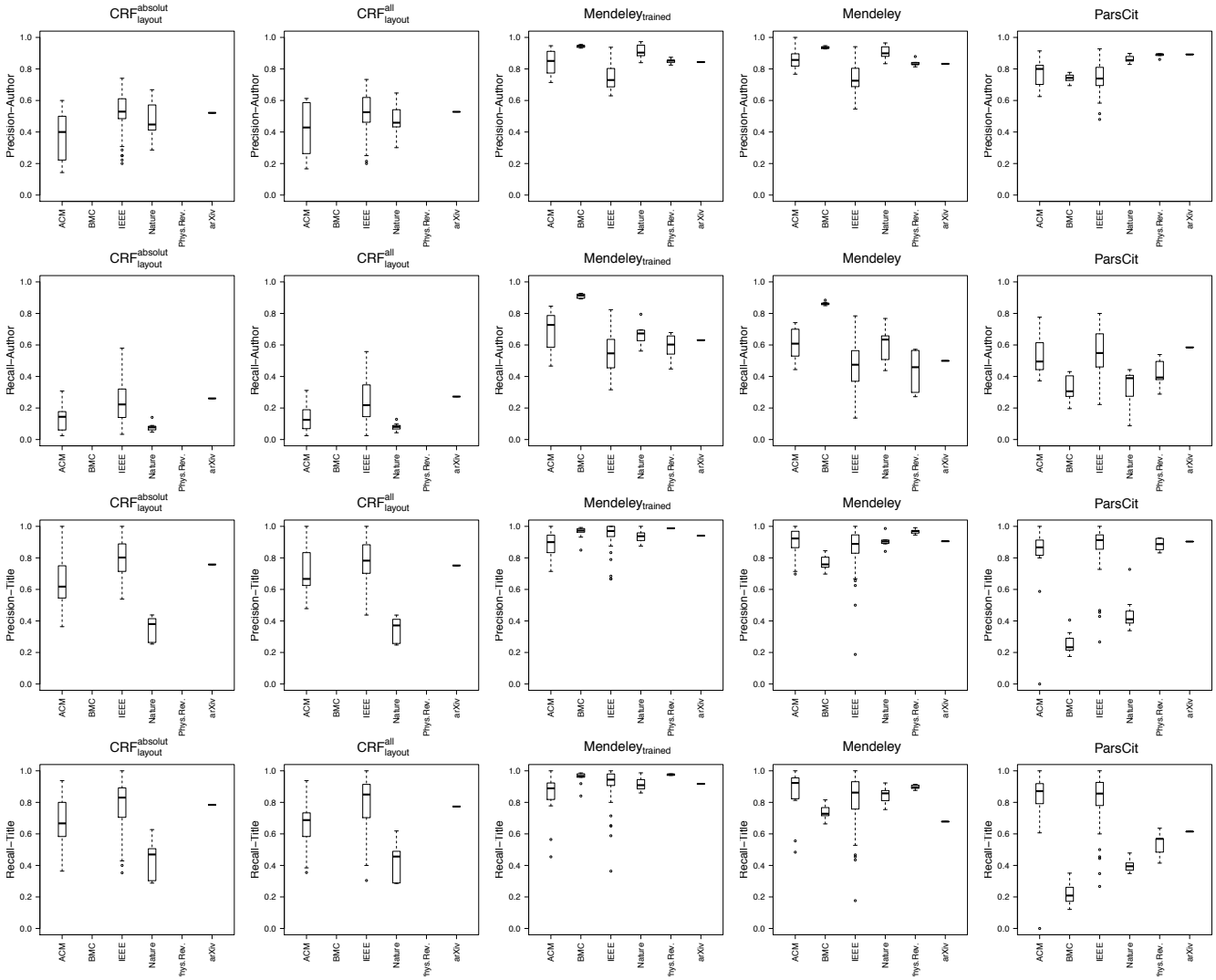

Figure 2: Boxplots on the precision/recall values for every journal group split up by the factors system and metadata-field

makes them interesting. Hence we will investigate clustering techniques to create classes of more homogeneous style information. We will apply these extended techniques to more complex extraction tasks like identifying equations, tables and figure.

Acknowledgement

This work has been funded by the European Commission as part of the TEAM IAPP project (grant no. 251514) within the FP7 People Programme (Marie Curie). The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

7. REFERENCES

[1] *ParsCit: An open-source CRF Reference String Parsing Package*. European Language Resources Association, 2008.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[3] M. Granitzer, M. Hristakeva, R. Knight, and K. Jack. A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management. In *Proceedings of the 27th Symposium On Applied Computing (poster)*, page to appear. ACM New York, NY, USA, 2012.

[4] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL'03*, pages 37–48, 2003.

[5] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles, and X. Zhang. Rule-based word clustering for document metadata extraction. In *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05*, page 1049, New York, New York, USA, 2005. ACM Press.

[6] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[7] M. Luong, T. Nguyen, and M. Kan. Logical structure recovery in scholarly articles with rich document features. *Journal of Digital Library Systems. Forthcoming*, 2011.

[8] G. Michael. Results and raw experimental data on the e-prints data set for a comparison of metadata extraction techniques for crowed-source bibliographic metadata management. `http://goo.gl/WMfU9`, 2011.

[9] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language TechnologyConference*, pages 329–336. HLT-NAACL04, 2004.

[10] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *Proceedings of AAAI 99 Workshop on Machine Learning for Information Extraction*, pages 37–42, 1999.

[11] F. Wu and D. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics, 2010.