Rule Management for Information Extraction from Title Pages of Academic Papers

Atsuhiro Takasu[1] and Manabu Ohta[2]
[1]*National Institute of Informatics, Tokyo, Japan*
[2]*Okayama University, Okayama, Japan*

Keywords: Digital Library, Document Understanding, Information Extraction, CRF.

Abstract: This paper discusses the problem of managing rules for page layout analysis and information extraction. We have been developing a system to extract information from academic papers that exploits both page layout and textual information. For this purpose, a conditional random field (CRF) analyzer is designed according to the layout of the object pages. Because various layouts are used in academic papers, we must prepare a set of rules for each type of layout to achieve high extraction accuracy. As the number of papers in a system grows, rule management becomes a big problem. For example, when should we make a new set of rules, and how can we acquire them efficiently while receiving new articles? This paper examines two scores to measure the fitness of rules and the applicability of rules learned for another type of layout. We evaluate the scores for bibliographic information extraction from title pages of academic papers and show that they are effective for measuring the fitness. We also examine the sampling of training data when learning a new set of rules.

1 INTRODUCTION

Information extraction is an important technology in utilizing documents. It helps to extract various kinds of metadata and to provide users with rich information access. For example, bibliographic information extraction from academic papers is useful to create or reconstruct metadata. It can be used for linking identical records stored in different digital libraries as well as for faceted retrieval. Although many researchers have studied bibliographic information extraction from papers and documents (e.g., (Takasu, 2003; Peng and McCallum, 2004; Councill et al., 2008)), it is still an active research area, and several competitions have been held [1].

For accurate information extraction, researchers have developed various rule-based methods that can exploit both logical structure and page layout. Document systems such as digital libraries usually handle various types of documents. Because the rules should be tailored for each type of document, formulating them requires effective and efficient methods. Rule management becomes harder as a system grows and contains larger numbers of articles with more variable layouts. For example, when receiving a set of articles, We must determine whether we should make a new set of rules or whether we can apply existing rules as in transfer learning (Pan and Yang, 2010). In addition, the rules should be properly updated because the layout of documents may sometimes change. To maintain such document systems, we require a rule management facility that can measure the fitness of rules and recompile rules when required.

We have been developing a digital library system for academic papers (Takasu, 2003; Ohta and Takasu, 2008). We are especially interested in extracting bibliographic metadata such as authors and titles. In previous studies, we applied a conditional random field (CRF) (Lafferty et al., 2001) to analyze and extract bibliographic information from title pages of academic papers. In these studies, we observed that rule-based methods can extract metadata with high accuracy, but we required multiple sets of rules and chose one according to page layout. In other words, we can obtain enough homogeneously laid out pages to learn a CRF that can analyze the pages with high accuracy for the task of metadata extraction from academic pages.

The use of multiple sets of rules requires rule management functions, such as choosing the appropriate set of rules for a particular document and deciding when to make a new set of rules for a change of page

[1] http://www.icdar2013.org/program/competitions

Takasu A. and Ohta M..
Rule Management for Information Extraction from Title Pages of Academic Papers.
DOI: 10.5220/0004827204380444
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 438-444
ISBN: 978-989-758-018-5

layout or a new page layout. This leads us to the study of managing rules for page layout analysis and bibliographic information extraction from title pages of academic papers. For this task, we first propose a method that uses two statistical measures calculated using CRF. Then, we examine their effectiveness for evaluating the fitness of a CRF for a particular page layout using three kinds of academic journals. The experimental results show that the measures decrease significantly when a CRF is applied to the title page of a journal that is different from the one used for learning the CRF. This result indicates that the statistical measures are effective for detecting page layout changes.

2 PROBLEM DEFINITION

There are several kinds of information extraction tasks for academic papers such as mathematical expressions, figures, and tables (Wang et al., 2004). This paper addresses bibliographic information extraction (Peng and McCallum, 2004), which is one of the fundamental tasks. This paper focuses on title page analysis, where we extract bibliographic information such as title, authors, and abstract from a title page. Figure 1 depicts an example of a title page. As shown in the figure, the task of bibliographic information extraction from the title page is to extract the red rectangles shown and to apply labels, such as "title" and "authors", to them.

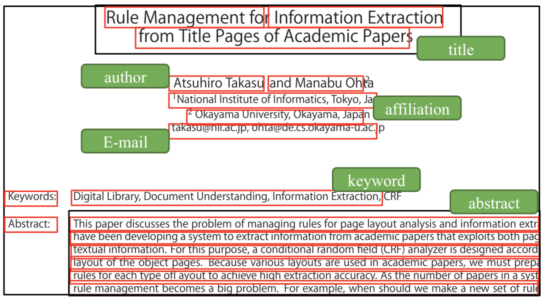

Rule Management for Information Extraction
from Title Pages of Academic Papers
Atsuhiro Takasu and Manabu Ohta[2]
[1]National Institute of Informatics, Tokyo, Ja
[2] Okayama University, Okayama, Japan
takasu@nii.ac.jp, ohta@de.cs.okayama-u.ac.jp
Keywords: Digital Library, Document Understanding, Information Extraction, CRF
Abstract: This paper discusses the problem of managing rules for page layout analysis and information extr
have been developing a system to extract information from academic papers that exploits both pag
textual information. For this purpose, a conditional random field (CRF) analyzer is designed accord
layout of the object pages. Because various layouts are used in academic papers, we must prepa
rules for each type ofl ayout to achieve high extraction accuracy. As the number of papers in a syst
rule management becomes a big problem. For example, when should we make a new set of rule

Figure 1: Example of page layout.

Because bibliographic information is located in a two-dimensional space, some researchers have proposed rules that can analyze components of a page based on a page grammar (Krishnamoorthy et al., 1992) and a 2D CRF (Zhu et al., 2005; Nicolas et al., 2007). In another approach, a sequential analysis can be applied after serializing components of the page in the preceding step. For example, Peng et al. proposed a CRF-based method of extracting bibliographies from the title pages and reference sections of academic papers in PDF format (Peng and McCallum, 2004). Councill et al. developed a CRF-based toolkit for page analysis and information extraction (Councill et al., 2008).

We adopted the latter approach. We label each text line on the title page of an academic article as an appropriate bibliographic element. For this purpose, linear-chain CRFs (Lafferty et al., 2001) were used. One linear-chain CRF was constructed for each journal to achieve high information extraction accuracy. The layout of a journal's title pages is, however, sometimes redesigned, which causes serious deterioration in information extraction accuracy. Therefore, this paper addresses the following problems:

- to detect such changes in the title page layout of academic papers, and
- to obtain a new CRF for analyzing pages in the new layout.

3 RULE MANAGEMENT

3.1 System Overview

We are developing a digital library system that covers various journals published in our country. Because their bibliographic information is stored in multiple databases, the system creates linkages by finding the papers in the multiple databases and provides a testbed for scholarly information study such as citation analysis and paper recommendation.

This system takes both newly published papers and those published previously but not yet included in the system. As stated in the previous section, we use multiple CRFs to extract information from various journals. The system chooses a CRF according to the journal title and applies it to papers to extract bibliographic information.

When the layout of a paper changes or a new journal is incorporated, we must judge whether we can use an existing CRF in the system or whether we must build a new CRF. The system supports rule maintenance by:

- checking the fitness of a CRF for given papers and alerting the user if the CRF does not analyze them with high confidence, and
- supporting labeling of training data when a new CRF is made.

3.2 The CRF

As described above, we adopted a linear-chain CRF for extracting bibliographic information from title pages of academic papers. Suppose L denotes a set of labels. For a token sequence $x := x_1 x_2 \cdots x_l$, a linear-chain CRF derives a sequence $y := y_1 y_2 \cdots y_l$ of labels, i.e., $y \in L^l$. A CRF M defines a conditional probability by:

$$P(y \mid x, M) = \frac{1}{Z(x)} \exp\left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x) \right\}, \tag{1}$$

where $Z(x)$ is the partition constant. The feature function $f_k(y_{i-1}, y_i, x)$ is defined over consecutive labels y_{i-1} and y_i, and the input sequence x. Each feature function is associated with a parameter λ_k giving the weight of the feature.

In the learning phase, the parameter λ_k is estimated from labeled token sequences. In the prediction phase, CRF assigns the label sequence, y^* to the given-token sequence x that maximizes Eq. (1).

3.3 Metrics for Change Detection

To detect a layout change from a token sequence, we use metrics that show how unlikely the test token sequence is generated from the model. This problem is similar to the sampling problem in active sampling (Saar-Tsechansky and Provost, 2004a).

In Eq. (1), the CRF calculates the likelihood based on the order of the hidden label sequence and each feature vector x_i generated from the estimated hidden label y_i. A change of page layout may affect the order of hidden labels as well as layout features in x_i. This leads to a decrease of the likelihood $P(y^* \mid x, M)$ given by Eqs. (1) and (2). A natural way to measure the model fitness is to use the likelihood. The CRF calculates the hidden label sequence, y, that maximizes the conditional probability given by Eq. (1). Higher $P(y^* \mid x, M)$ means more confident assignment of labels, while lower $P(y^* \mid x, M)$ means that the token sequence makes it hard for the current CRF model to assign labels.

The conditional probability is affected by the length of the token sequence, x; therefore, we use the following normalized conditional probability as a model fitness measure:

$$C_l(x) := \frac{\log\left(P(y^* \mid x)\right)}{|x|}, \tag{2}$$

where $|x|$ denotes the length of the token sequence, x. We refer to the metric given by Eq. (2) as a *normalized likelihood*.

The normalized likelihood is a kind of confidence measure when the model assigns labels to all tokens in the sequence, x. The second measure is based on the confidence measure for assigning labels to a single token in the sequence. For sequence x, let Y_i denote a random variable for assigning a label to the ith token in x. For label l in a set L of labels, $P(Y_i = l)$ denotes the marginal probability that label l is assigned to the ith token. If the token has feature values clearly supporting a specific label, for example, $l \in L$, $P(Y_i = l)$ must be significantly high and $P(Y_i = l')$ $(l' \neq l)$ must be low. Hence, the following entropy quantifies this feature:

$$\sum_{l \in L} -P(Y_i = l) \log(P(Y_i = l)) \,. \tag{3}$$

Lower entropy signifies that the label of token x_i is more likely to be l. For the sequence x, we use the following *average entropy* as another model fitness measure:

$$C_e(x) := \frac{\sum_{x_i \in x} \sum_{l \in L} -P(Y_i = l) \log(P(Y_i = l))}{|x|} \,. \tag{4}$$

3.4 Change Detection

Suppose CRF M is used to label a token sequence obtained from a title page. There are a couple of ways to define the change detection problem. The most basic definition is as follows. Given a new token sequence x, determine whether the sequence is from the same information source as that from which the current CRF M is learned.

For the journal layout detection problem, one issue of a journal usually contains multiple papers, so the problem is defined as detecting the change when given a set $\{x_i\}_i$ of token sequences. The rest of this paper addresses this problem.

A token sequence x is judged to be a token sequence from the same information source if $C(x) > \sigma$ holds where σ is a threshold, where C is C_i or C_e defined in Section 3.3. Otherwise, the layout has changed.

3.5 Learning a CRF for a New Layout

Once we detect papers with a page layout that is different from those already known, we must derive a new CRF for the detected papers. We apply the following active sampling technique (Saar-Tsechansky and Provost, 2004b) to this task,

1. gather a significant number of papers T without labeling,
2. choose an initial small number of papers T_0 from T, label them, and learn an *initial CRF* M_0 using the labeled papers,
3. repeat until convergence:

(a) choose a small number of papers T_t from the pool $T - \cup_{i=0}^{t-1} T_i$ using the CRF M_{t-1} we obtained in the previous loop,

(b) label the papers T_t manually,

(c) learn CRF M_t using the labeled papers $\cup_{i=0}^{t} T_i$.

The purpose of active sampling is to reduce the cost of labeling required for learning the CRF. One drawback is that we must delay learning the new CRF until we have gathered enough papers in the new layout in Step 1.

In active sampling, the sampling strategy for the initial CRF in Step 2 and for updating the CRF in Step 3-(c) is important. For the initial CRF, we choose the k papers in T with the lowest values of the metrics C introduced in Section 3.3, where C is calculated using the CRFs that we have at that time. This strategy means that we choose training papers for the initial CRF having the most different layout from those that we have so far.

In the tth update phase, we choose the k training papers from $T - \cup_{i=0}^{t-1} T_i$ with the lowest values of the metrics C, where C is calculated using the CRF M_{t-1} that we obtained in the previous step, instead of the CRF for the original layout. This strategy means that we choose training papers with a different layout from those in $\cup_{i=0}^{t-1} T_i$.

4 EXPERIMENTAL RESULTS

4.1 Dataset

For this experiment, we used the dataset prepared for our previous study (Ohta et al., 2010). It is taken from the following three journals:

- Journal of Information Processing by the Information Processing Society of Japan (IPSJ): We used papers published in 2003 in this experiment. This dataset contains 479 papers, most of them written in Japanese.
- IEICE Transactions by The Institute of Electronics, Information and Communication Engineering in Japan (IEICE-E): We used papers published in 2003. This dataset contains 473 papers, all written in English.
- IEICE Transactions by The Institute of Electronics, Information and Communication Engineering in Japan (IEICE-J): We used papers published in 2003 and 2004. This dataset consisted of 174 papers, most of them written in Japanese.

We used the following labels for the bibliographic elements as in (Ohta et al., 2010).

- Title: We used separate labels for Japanese and English titles because Japanese articles contained titles in both languages.
- Authors: We used separate labels for Japanese and English authors as in the title.
- Abstract: As for title and author, we used labels for English and Japanese abstracts.
- Other: Title pages usually contain paragraphs of articles such as those for introductory paragraphs for the article. We assigned the label “other” to lines in these paragraphs.

Note that different journals have different bibliographic components on their title pages.

Because we used the chain-model CRF, the tokens must be serialized. In this experiment, we regard each line as a token. We used lines extracted by OCR and serialized the lines according to the order generated by the OCR system. We manually labeled each line for training and evaluation.

4.2 Features of the CRF

Fifteen features were adapted as in (Ohta et al., 2010). Among them, 14 are unigram features, and the remaining one is a bigram feature. They are also classified into two other kinds of features: layout features, such as location, size, and gaps between lines; and linguistic features, such as the proportions of several kinds of characters in the tokens and the appearance of characteristic keywords that often appear in a specific bibliographic component such as “institute” for affiliations. Table 1 summarizes the set of feature templates. Their instances are automatically generated from training token sequences.

For example, an instance of the bigram feature template $< y(-1), y(0) >$ is:

$$f_k(y_{i-1}, y_i, x) = \begin{cases} 1 & \text{if } y_{i-1} = \text{title}, y_i = \text{authors} \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

This bigram feature indicates that the author follows the title in a token sequence, and the corresponding parameter λ_k shows how likely it is that this token sequence occurs. CRF++ 5.8 [2] (Kudo et al., 2004) was used to learn and label the token sequence of each title page.

4.3 Evaluation Metrics

For our experiments, we used two evaluation metrics. One was for evaluating the performance of the

[2] http://code.google.com/p/crfpp/

Table 1: Feature templates for bibliography labeling (Ohta et al., 2010).

Type	Feature	Description
unigram	$< i(0) >$	Current line ID
	$< x(0) >$	Current line abscissa
	$< y(0) >$	Current line ordinate
	$< w(0) >$	Current line width
	$< h(0) >$	Current line height
	$< g(0) >$	Gap between current and preceding lines
	$< cw(0) >$	Median of characters' width in the current line
	$< ch(0) >$	Median of characters' height in the current line
	$< \#c(0) >$	# of characters in the current line
	$< ec(0) >$	Proportion of alphanumerics in the current line
	$< kc(0) >$	Proportion of kanji in the current line
	$< jc(0) >$	Proportion of hiragana and katakana in the current line
	$< s(0) >$	Proportion of symbols in the current line
	$< kw(0) >$	Presence of predefined keywords in the current line
bigram	$< y(-1), y(0) >$	Previous and current labels

sequence analysis; i.e., its accuracy. It was defined as

$$\frac{\text{\# successfully labeled sequences}}{\text{\# test sequences}}. \qquad (6)$$

Note that a CRF was only regarded as having succeeded in labeling when it assigned correct labels to all tokens in the token sequence. In other words, if a CRF assigned an incorrect label to one token but correctly labeled all other tokens in a sequence, *x*, it was regarded as having failed.

The other metric was the accuracy of change detection. For the change detection, we first learned a CRF by using training data for each journal. In the test phase, we mixed token sequences from two journals and let the CRF judge whether a token sequence came from the journal used for learning or from the other one. If a test token sequence was judged to come from the same journal as that used for learning, we regarded the sequence as *positive*. Otherwise it was regarded as *negative*.

The receiver operating characteristic (ROC) curve was used for evaluation. That is, the mixed sets of test token sequences were ranked according to the metrics explained in Section 3.3. By regarding the top *k* token sequences in the list as *positive*, we obtained the true positive and false positive rates for each *k*. We plotted the ROC curve by changing *k*.

4.4 Sequence Analysis Accuracy

We first examined the accuracy of CRFs learned separately for each journal. Although their accuracies were not the main concern of this paper, we measured them as one of the basic statistics of the CRFs for this experiment; they are also helpful for the later analysis of the change detection.

We applied fivefold cross-validation to evaluate the sequence analysis accuracy. We first learned CRFs for each journal by using four out of five evenly split datasets. Then, we evaluated the learned CRFs using the remaining dataset as a test set. Table 2 shows the average accuracies defined by Eq. (6). As shown in the table, we obtained CRFs with various levels of accuracy.

Table 2: Average Accuracy of CRFs.

IPSJ	IEICE-E	IEICE-J
0.947	0.891	0.752

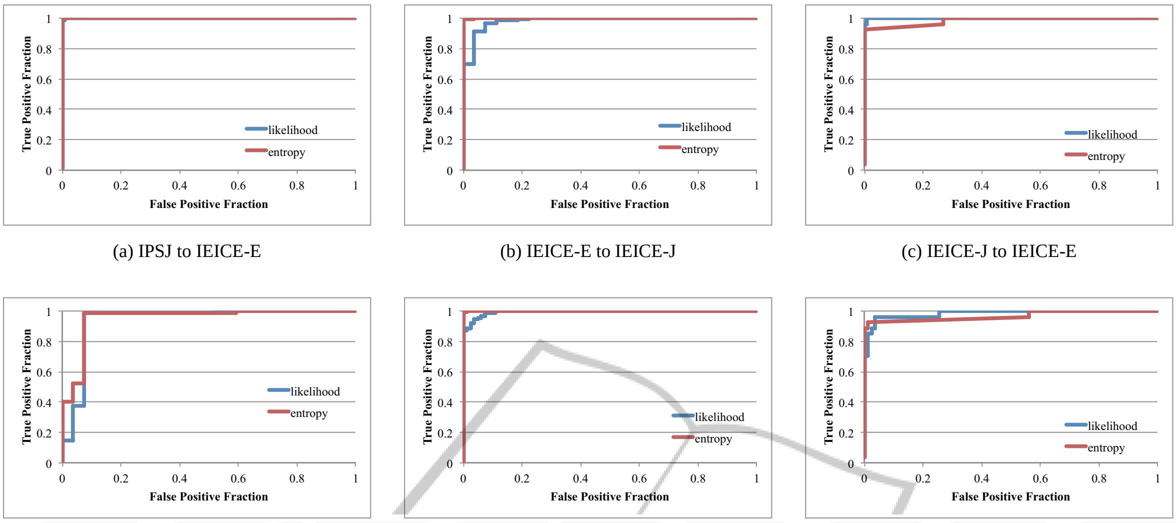
4.5 Change Detection Performance

To measure the performance of the proposed metrics, we made test data by mixing two test datasets. More precisely, we applied each learned CRF described in Section 4.4 to the set of sequences consisting of

- the test set of the journal used for learning the CRF, and
- one test set from another journal.

For each pair of journals, Figure 2 depicts ROC curves. Each panel contains the ROC curve by the normalized likelihood ("likelihood") and the average entropy ("entropy"). For example, the ROC curve in panel (a) in Figure 2 is the result of detecting token sequences of IEICE-E from those of IPSJ using the CRF learned from labeled IPSJ token sequences. Similarly, the ROC curve in panel (b) is the result of detecting token sequences of IEICE-J from those of IEICE-E using the CRF learned by labeled IEICE-E token sequences.

First, the ROC curves show that both the normal-



(a) IPSJ to IEICE-E

(b) IEICE-E to IEICE-J

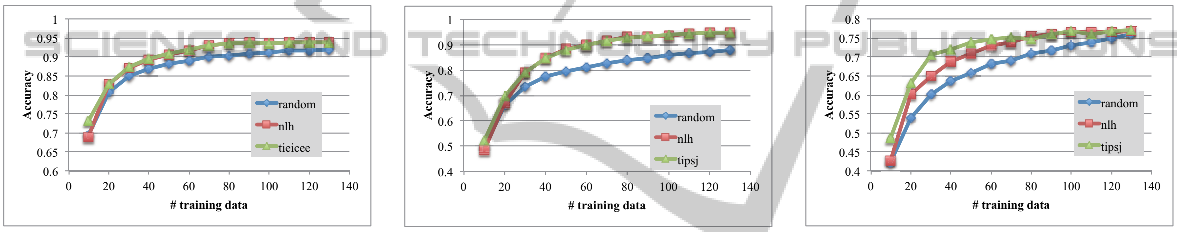
(c) IEICE-J to IEICE-E

(d) IPSJ to IEICE-J

(e) IEICE-E to IPSJ

(f) IEICE-J to IPSJ

Figure 2: Change detection performance.



(a) IPSJ

(b) IEICE-E

(c) IEICE-J

Figure 3: CRF learning by active sampling.

ized likelihood and the average entropy are very effective for detecting token sequences from a different journal from the one used for learning the CRF. The ranked test token sequence lists according to these metrics are clearly separated. Panels (a), (b), and (e) in Figure 2 show that the ranked lists according to the average entropy are perfectly separated into two journals: IPSJ and IEICE-E in (a), IEICE-E and IEICE-J in (b), and IEICE-E and IPSJ in (e).

Second, the training journal used affects the change detection. For example, compare panels (d) and (f) in Figure 2. In both panels, the test data were the same, i.e., the mixture of token sequences of IPSJ and IEICE-J, but the CRF was learned from IPSJ in (d) and from IEICE-J in (f). From the table 2, the CRF of (d) is more accurate than that of (f), whereas the detection accuracy of CRF (f) is better than that of (d). This is an interesting phenomenon, and we plan to analyze it further.

Third, both the normalized likelihood and the average entropy work very well for change detection; we observed no significant difference in detection accuracy between them.

4.6 CRF Learning

To evaluate the method for learning CRFs, we observed the accuracy of CRFs for three journals. More precisely, for each journal

1. obtain a CRF M for another journal,
2. choose an initial training sample using M and obtain an initial CRF M_0,
3. repeat updating CRF to M_t by choosing a training sample using M_{t-1}.

In this experiment, we fixed the sample size to 10 in both the initial and update phases. The accuracy of the CRF is measured by Eq. (6). For comparison, we measured the accuracy of the following sampling strategies:

- *random*: 10 training samples are randomly chosen in both the initial and update phases,
- *nlh*: 10 training samples are randomly chosen in the initial phase, and 10 training samples are chosen according to the normalized likelihood in each update phase,

as well as the proposed method:

- *journal*: 10 training samples are chosen according to the normalized likelihood of the CRF for *journal* in the initial phase, and 10 training samples are chosen according to the normalized likelihood in each update phase.

Because we obtained similar results for the metric *average entropy*, we show only the results for *normalized likelihood* in this section.

Figures 3 (a), (b), and (c) respectively show the accuracy of CRFs for journals IPSJ, IEICE-E, and IEICE-J. Each graph in the figure plots the accuracy of the CRF with respect to the size of training samples by three sampling strategies.

First, we observed that both the proposed strategy and *nlh* obtained accurate CRFs with fewer samples than with *random*. This indicates that the sampling strategy for the update phase is effective.

Second, when we compare the proposed strategy and *nlh*, the proposed strategy obtains a slightly better initial CRF; its accuracy is plotted at the training data size of 10. This indicates that the sampling strategy using a CRF designed for another journal can improve the active learning process.

5 CONCLUSIONS

We have examined two statistical measures obtained using a linear-chain CRF for detecting layout changes of title pages of academic papers and obtaining new CRFs for extracting information from academic title pages. The experiments revealed that both statistical measures are very effective at detecting layout changes. We also showed that the measures can be used for active sampling to reduce the labeling cost of training data.

We plan to extend this study in several directions. First, it is unknown how the CRF's sequence labeling accuracy affects the change detection accuracy. To study this problem, we plan two kinds of experiments: (1) controlling the labeling accuracy by the size of training data, obtaining CRFs with various labeling accuracy, and comparing them for change detection, and (2) applying our approach to more complex sequence labeling problems.

In this paper, we used datasets that we prepared. To make comparison easier, we plan to evaluate the method using other open datasets such as the ICDAR2009 layout dataset (Antonacopoulos et al., 2009).

REFERENCES

Antonacopoulos, A., Bridson, D., Papadopoulos, C., and Pletschacher, S. (2009). A realistic dataset for performance evaluation of document layout analysis. In *ICDAR2009*, pages 296 – 300.

Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). Parscit: An open-source crf reference string parsing package. In *LREC*, page 8.

Krishnamoorthy, M., Nagy, G., and Seth, S. (1992). Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Computer*, 25(7):10–22.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *EMNLP 2004*.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th ICML*, pages 282–289.

Nicolas, S., Dardenne, J., Paquet, T., and Heutte, L. (2007). Document image segmentation using a 2d conditional random field model. In *ICDAR 2007*, pages 407 – 411.

Ohta, M., Inoue, R., and Takasu, A. (2010). Empirical evaluation of active sampling for crf-based analysis of pages. In *IEEE IRI 2010*, pages 13–18.

Ohta, M. and Takasu, A. (2008). CRF-based authors' name tagging for scanned documents. In *JCDL'08*, pages 272–275.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):1345 – 1359.

Peng, F. and McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL*, pages 329–336.

Saar-Tsechansky, M. and Provost, F. (2004a). Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153–178.

Saar-Tsechansky, M. and Provost, F. (2004b). Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153–178.

Takasu, A. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. In *JCDL '03*, pages 49–60.

Wang, Y., Phillips, I. T., R.M.Robert, and Haralick, M. (2004). Table structure understanding and its performance evaluation. *Pattern Recognition*, 37(7):1479–1497.

Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y. (2005). 2D conditional random fields for web information extraction. In *ICML 2005*.