IADIS International Journal on Computer Science and Information Systems
Vol. 7, No.2, pp. 18-31
ISSN: 1646-3692

EMPIRICAL EVALUATION OF CRF-BASED BIBLIOGRAPHY EXTRACTION FROM RESEARCH PAPERS

Manabu Ohta. *Okayama University, Japan*

Ryohei Inoue. *Shikoku Hitachi Systems, Ltd., Japan*

Atsuhiro Takasu. *National Institute of Informatics, Japan*

ABSTRACT

We proposed an automatic bibliography extraction method for research papers scanned with OCR markup. The method uses conditional random fields (CRFs) to label serially OCRed text lines in the article title page as appropriate bibliographic element names. Although we achieved good extraction accuracies for some Japanese academic journals, extraction errors are inevitable. Therefore, this paper proposes three confidence measures for bibliography labeling to detect such extraction errors. This paper also reports an empirical evaluation of CRF-based page analysis for research papers on the basis not only of labeling accuracy but also of labeling error detection. We applied the three confidence measures to detecting errors of labeling articles selected from three academic journals published in Japan. The experiments showed that the proposed confidence measures reasonably indicated the labeling accuracies and could be used for error detection. This paper also discusses the tradeoff between the quality of bibliographic data assured by human post-editing of detected errors and its cost.

KEYWORDS

Bibliography extraction, conditional random field (CRF), error detection, OCR, digital library.

1. INTRODUCTION

Nowadays many publishers and academic societies provide articles in digital formats. Owing to these services we can quickly obtain articles. Early digital library systems stored articles independently from each other. Hence, we needed to make another search to obtain cited papers. Recently, they begin to make networked documents where cited papers are linked to each other and authors are also connected to the articles they wrote.

The linkage between papers and authors enhances the function of digital library systems in various ways. From the viewpoint of information retrieval, it enables us to easily access cited papers by just clicking links in a reference list. By following the linkage between authors and articles, we can gather articles written by a specific group of authors. From the viewpoint of bibliometrics, we can count the number of citations of each paper which is a fundamental metric for measuring the quality of articles and journals. Similarly, the number of publications of researchers obtained through linkage analysis is also an important metric to measure their productivity.

Since articles are usually published without explicit linkage to their related contents, we need to find them from the text of articles. For this purpose, we first need to extract bibliographic entities to be linked such as authors and titles appearing in the title pages and references. This is an information extraction problem extensively studied in natural language processing (NLP) and machine learning (ML) communities. Some researchers applied sequence labeling techniques to extract entities (Xin et al., 2008). Entity extraction is also studied as a problem of document layout analysis in pattern recognition community (Nagy et al., 1992). After extracting entities, various machine learning techniques were also applied to entity matching problem.

Most of the studies on entity extraction and linkage analysis focus on improving the extraction and linkage accuracies as much as possible. This approach leads to so-called best effort systems. For information retrieval, best effort systems are reasonable, however, for analysis of articles or researchers as in bibliometrics, the quality of extracted linkage should be assured. In early studies of entity linkage, Fellegi and Sunter proposed an entity linkage model that assures linkage accuracy (Fellegi and Sunter, 1969). Most entity linkage systems judge whether a given pair of entities is identical or not, i.e., the systems are regarded as a binary classifier. In Fellegi-Sunter model, systems classify a pair of entities into three categories, i.e., identical, unknown, and not identical. The pairs judged as unknown are manually classified. By introducing human judgment, the model assures the quality of linkage.

In this study, we aim to develop an entity extraction model that assures the quality as in Fellegi-Sunter model. The first step for the model construction is to develop a method that can detect unknown results of entity extraction. In this paper, we define the problem as bibliography extraction from a title page of research papers. We first describe our CRF-based bibliography extraction briefly and then empirically discuss the effectiveness of several measures proposed for error detection of CRF-based bibliography labeling.

As for bibliography extraction from PDFs, Okada et al. proposed a method to extract bibliographies from reference strings (Okada et al., 2004). They combined a support vector machine (SVM) and a hidden Markov model (HMM) where the SVM is used for handling features of each token in reference strings, whereas the HMM is used for handling features of label transition. Peng et al. proposed a CRF-based method of extracting bibliographies from the title pages and reference strings in research papers in PDF format (Peng and McCallum, 2004). They correctly labeled entire entities in title pages of research papers with 73.3% accuracy using 13 bibliographic labels defined for title pages. They compared CRFs with HMMs and SVMs and experimentally showed that the CRF outperformed the other methods. None of these studies, however, discussed how to detect errors and pass them to human judgment to assure the quality.

For extracting bibliographies from legacy articles that are digitized via scanning and OCR processing, we need methods that are robust against noises caused by OCRs. Takasu et al. proposed a robust method of extracting references from scanned research papers and applied it to articles in various journals (Takasu, 2003). Their method was based on HMM and could handle OCR errors. We also developed a method of extracting bibliographies from a title page of OCRed academic articles. The method uses a CRF to assign labels to text lines in title pages. The input of the CRF is the text lines serialized by the OCR. Since OCRed documents involve physical layout features such as height of lines and distance between lines, we exploited the layout features as well as textual features obtained from the text in lines to improve the extraction accuracy (Ohta et al., 2008).

2. CRF-BASED BIBLIOGRAPHY EXTRACTION

2.1 CRF

A CRF is a statistical sequence labeling framework proposed by Lafferty et al. (Lafferty et al., 2001) for part-of-speech tagging and syntactical analysis. CRFs outperform other popular models, such as HMMs and maximum entropy models, when the true data distribution has higher order dependencies than the models, which is often the case under practical circumstances. Moreover, CRFs have performed well in many studies in fields ranging from bioinformatics to natural language processing (Kudo et al., 2004).

We adopt a common linear-chain CRF for text line labeling. That is, we define a conditional probability of a label sequence $\mathbf{y} = y_1,...,y_n$ given an input token sequence $\mathbf{x} = x_1,...,x_n$ as follows:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1}, y_i, \mathbf{x})), \tag{1}$$

where $Z(\mathbf{x})$ is the normalization constant that makes the probability of all candidate label sequences sum to one, $f_k(y_{i-1}, y_i, \mathbf{x})$ is an arbitrary feature function over the *i*th label y_i, its previous label y_{i-1}, and the input sequence **x**, and λ_k is a learned weight associated with the feature function f_k.

The CRF assigns the label sequence $\mathbf{y}^*$ to the given token sequence **x** that maximizes Eq. (1), i.e.,

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}). \tag{2}$$

Note that the input token sequence **x** is the sequence of text line IDs, while the label sequence **y** is the sequence of bibliographic element names such as a title, authors, and an abstract.

Our CRF-based labeling uses the CRF++ package (http://crfpp.sourceforge.net/) which is an open source implementation of CRFs for labeling sequential data. When training the CRF, we set the learning parameters such as balancing the degree of fitting to default values given by the CRF++.

2.2 CRF-Based Bibliographic Labeling

We label each text line in the title page of an academic article as an appropriate bibliographic element. Bibliographic elements include paper titles, authors, abstracts, and whatever other components found in title pages of the target journal papers. It should be noted that a bibliographic element includes at least a text line produced by the OCR and is often comprised of several lines.

For page layout analysis and character recognition, we have developed an OCR system in collaboration with an OCR vendor. For each scanned page, the OCR system produces not only recognized text, but also XML markup indicating the bounding rectangles for the characters, words, lines, and blocks. The labeling target is the text lines composed of one or more words. Moreover, these XML elements have the layout attributes of x, y, width, and height, and therefore, we know where the text blocks, lines, words, or characters are located in the page and how large they are.

We prepared nine kinds of bibliographic element labels listed in Table 1 for extracting them from three target academic journals published in Japan, i.e., IPSJ Journal (IPSJ), IEICE Trans. Commun. in English (IEICE-E), and IEICE Trans. Inf. & Syst. in Japanese (IEICE-J). In Table 1, prefixes of "j-" and "e-" respectively stand for Japanese and English, and "type" is the article type specifically defined for IEICE-E. Note that different journals have different bibliographic elements in their title pages.

Table 2 summarizes the set of adopted feature templates that automatically generate a set of feature functions for the text line labeling. As for visual features reflecting layout information of title pages, we take into account not only a line's location and size, i.e., <x(0)>, <y(0)>, <w(0)>, and <h(0)>, but also the gap between lines, <g(0)>, and the size and number of characters constituting each line, i.e., <cw(0)>, <ch(0)>, and <#c(0)>. As for linguistic features reflecting textual information of OCRed text, we adopt proportions of several kinds of characters in the text line, i.e., <ec(0)>, <kc(0)>, <jc(0)>, and <s(0)>, and appearances of characteristic keywords, <kw(0)>, which seem correlated with a specific bibliographic element, e.g., "university" often found at authors' affiliations.

An example of the feature functions generated by the bigram feature template <y(−1),y(0)> is as follows:

$$f_k(y_{i-1}, y_i, \mathbf{x}) = \begin{cases} 1 & \text{if } y_{i-1} = \text{j-title},\ y_i = \text{j-authors} \\ 0 & \text{otherwise} \end{cases}. \qquad (3)$$

This label bigram reflects the syntactic constraints of bibliographic elements, i.e., that the authors' area typically follows the title area and is followed by the abstract area, and so on.

The number of generated feature functions depends on that of kinds of bibliographic labels. As for IPSJ papers, for example, the number of feature functions generated by the unigram feature template, e.g., <i(0)>, is $7 \times N$, where 7 is the number of bibliographic labels used for IPSJ shown in Table 1 and *N* is the number of different unigrams, i.e., line IDs. The number of bigram feature functions generated by <y(−1),y(0)> amounts to 7×7.

Table 1. Bibliographic element labels

Bibl. element label	type	j-title	j-authors	j-abstract	j-keywords	e-title	e-authors	e-abstract	other
IPSJ	-	yes	yes	yes	-	yes	yes	yes	yes
IEICE-E	yes	-	-	-	-	yes	yes	yes	yes
IEICE-J	-	yes	yes	yes	yes	yes	yes	-	yes

Table 2. Feature templates for labeling text lines

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

3. DETECTION OF BIBLIOGRAPHY EXTRACTION ERRORS

3.1 Confidence Measure

We propose three confidence measures for evaluating the difficulty of CRF-based labeling in order to detect labeling errors. These measures should highly correlate with the accuracy of labeling. Therefore, we need to know how much the measures correlate with the accuracy and how the correlation is affected by the accuracy. We first explain two measures which we originally proposed for active sampling (Ohta et al., 2010). We then propose the other confidence measure on the basis of the entropy of label assignment to each token (Settles and Craven, 2008).

3.1.1 Normalized Likelihood

As we described in section 2, a CRF calculates the hidden label sequence, **y**, which maximizes the conditional probability given by Eq. (1). Higher $p(\mathbf{y}^* \mid \mathbf{x})$ means more confident assignment of labels. In contrast, lower $p(\mathbf{y}^* \mid \mathbf{x})$ means that it is hard for the CRF to assign labels to the token sequence.

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

$$c_{NLH}(\mathbf{x}) := \frac{\log(p(\mathbf{y}^* \mid \mathbf{x}))}{|\mathbf{x}|}, \tag{4}$$

where $|\mathbf{x}|$ denotes the length of the token sequence, **x**, i.e., the number of text lines in a title page. We denote the normalized likelihood as *NLH*.

3.1.2 Minimum Probability of Token Assignment

NLH is a confidence measure on the basis of label assignment to all tokens in a sequence, **x**. The second measure is based on the confidence in assigning a label to a single token in the sequence. For sequence **x**, let Y_i denote a random variable for assigning a label to the *i*th token in **x**, i.e., x_i. Let *L* be a set of labels. For label *l* in *L*, $p(Y_i = l)$ denotes the marginal probability that label *l* is assigned to x_i. We can then regard the maximum probability, $\max_{l \in L} p(Y_i = l)$, indicates confidence in labeling x_i. Using the confidence, we define the second confidence measure as follows:

$$c_{MP}(\mathbf{x}) := \min_{i \le |\mathbf{x}|} \max_{l \in L} p(Y_i = l). \tag{5}$$

We denote the probability as *MP*.

3.1.3 Maximum Token Entropy

The NLH focuses only on the most likely label sequence, $\mathbf{y}^*$, and the MP focuses only on the largest marginal probability of an assigned label, $\max_{l \in L} p(Y_i = l)$. However, we consider that the distribution of label assignment probabilities over all the possible label sequences also reflects the difficulty of labeling. Therefore, we propose using entropy of labeling as follows.

We take into consideration not only the most probable label assigned to each token but also the other labels to determine the third confidence measure. If there are many other label sequences with almost the same probability as the most probable one has, the CRF is considered less confident in labeling and so is in the assigned label sequence. While the CRF is considered confident in its labeling when it assigns to a token one label with a probability of nearly one and other labels have a probability of nearly zero. Therefore, we propose the following maximum token entropy as the third confidence measure:

$$c_{MTE}(\mathbf{x}) := -\max_{i \le |\mathbf{x}|} \sum_{l \in L} - p(Y_i = l) \log p(Y_i = l). \tag{6}$$

The minus sign in front is simply to ensure that $c_{MTE}(\mathbf{x})$ acts as a confidence measure just like $c_{NLH}(\mathbf{x})$ and $c_{MP}(\mathbf{x})$. We denote the maximum token entropy as *MTE*.

3.2 Error Detection Strategy

We need to detect labeling errors among a set of CRF-labeled token sequences. We consider that less-confident sequences are more likely to be erroneous than more-confident ones. Therefore, we detect such less-confident sequences as errors as follows:

1. Calculate the confidence measures, *c*.(**x**), for each token sequence **x** in the set of CRF-labeled data,
2. Order the sequences in ascending order w.r.t. *c*.(**x**), which can be regarded as difficulty order, and
3. Choose top-ranked token sequences as errors.

After detecting errors, we can manually check the detected token sequences to assign correct labels, which is expected considerably easier than manually checking all the sequences.

4. EMPIRICAL EVALUATION

4.1 Experimental Setup

The CRF-based bibliography extractor extracts bibliographic components from scanned and OCRed title pages of research papers by labeling text line sequences. We first describe experiments on extraction accuracies and then those on detection of extraction errors by using the confidence measures. We evaluated the performance of our CRF-based bibliography extractor and the effectiveness of the confidence measures for detecting errors by using three kinds of academic papers as follows.

1. Japanese papers issued by the Information Processing Society of Japan (IPSJ): We used those issued in 2003. This dataset consisted of 479 papers.
2. English papers issued by the Institute of Electronics, Information and Communication Engineers in Japan (IEICE-E): We used those issued in 2003 and this dataset consisted of 473 papers.
3. Japanese papers issued by the Institute of Electronics, Information and Communication Engineers in Japan (IEICE-J): We used those issued in 2003 and 2004 and this dataset consisted of 174 papers.

We applied five-fold cross validation to each dataset. We used real data since our OCR outputs were difficult to simulate. This is because the OCR outputs included errors caused by layout analysis as well as those by character recognition. The accuracy of the abstract was 99%, but that of the references was 97%. The misrecognitions were mainly caused by the mixture of Japanese and English characters, as well as the various fonts and punctuation symbols appearing in the references.

4.2 Bibliography Extraction Accuracies

We used the accuracy with which a CRF assigned labels to each token in the test token sequences as the evaluation metric. A CRF was only regarded as having succeeded in labeling a token sequence when it assigned correct labels to all tokens in the sequence. In other words, if a CRF assigned an incorrect label to a token and correctly labeled all other tokens in a sequence, **x**, the CRF was regarded as having failed in assigning labels to the sequence, **x**. Therefore, the labeling accuracy was

$$\frac{\#\text{ of correctly labeled sequences}}{\text{total }\#\text{ of test sequences}}.$$

We repeated the experiment with 30 random sampling of training data. That is, we randomly chose 20, 100, and 300 samples from the training dataset 30 times for each number of samples. The resultant extraction accuracies are the average for these 30 trials and shown in Table 3. Note here that the numbers of test sequences were 95.8 (IPSJ), 94.6 (IEICE-E), and 34.8 (IEICE-J) on average.

As seen in Table 3, the erroneously labeled test sequences decreased with the increase in the number of training samples. The result of 300 training samples for IEICE-J is not given because the total number of samples of this journal was 174 as described in section 4.1. We experimented with the three different numbers of training samples to evaluate the effectiveness of the proposed confidence measures for various accuracy levels.

Table 3. Extraction accuracies (%) and # of erroneously labeled test sequences (in parentheses)

# of training samples	20	100	300
IPSJ	83.4% (15.9)	91.9% (7.8)	93.8% (6.0)
IEICE-E	69.5% (28.8)	89.7% (9.8)	95.9% (3.9)
IEICE-J	65.7% (12.0)	79.8% (7.0)	-

4.3 Extraction Robustness against Text Line Permutation

We found not a few text line permutations in our experimental dataset. That is, the order of OCRed text lines of some articles was different from that in which human readers read them because of erroneous layout analysis. Therefore, we conducted the following experiments to evaluate the robustness of the CRF-based labeling against such text line permutation.

We first determined the correct order of bibliographies' appearance in a title page based on that of human readers for each journal as follows:

1. IPSJ: (other) → j-title → j-authors → j-abstract → e-title → e-authors → e-abstract → other
2. IEICE-E: type → e-title → e-authors → e-abstract → other
3. IEICE-J: (other) → j-title → j-authors → e-title → e-authors → (other) → j-abstract →
   j-keywords → other

Here "(other)" matches an "other" line zero or more times while "other" matches an "other" line one or more times. We then separated the experimental dataset into two: one was the samples which conformed to the above bibliography order and the other was those which did not. Table 4 summarizes the resultant classification. As seen in the table, IPSJ and IEICE-J had a relatively small number of articles including text line permutations while IEICE-E had many such articles. Note here that the articles which conformed to the above bibliography order were not necessarily completely permutation-free because we did not check the permutation of text lines in the same kind of bibliography.

For evaluating the robustness of our labeling, we conducted the experiment by using the permutation-free samples obtained through the classification. The resultant accuracies are shown in Table 5. In this table, the results of 100 and 300 training samples for IEICE-E are not given because the total number of permutation-free samples of this journal was 73 as shown in Table 4. Comparing Table 5 to Table 3, we can see that extraction from the permutation-free data was easier than from the original data irrespective of journal. Especially, the accuracy of IEICE-E with 20 training samples increased remarkably when we used only the permutation-free data, which indicates that eliminating permutation could lead to better accuracy.

Table 4. The number of classified samples

	Total	Permutation	Ratio (%)
IPSJ	479	24	5.01
IEICE-E	473	400	84.56
IEICE-J	174	17	9.77

Table 5. Extraction accuracies (%) and # of erroneously labeled test sequences (in parentheses) for permutation-free data

# of training samples	20	100	300
IPSJ	87.4% (11.5)	96.0% (3.7)	97.4% (2.4)
IEICE-E	91.0% (1.3)	-	-
IEICE-J	69.8% (9.5)	82.3% (5.6)	-

4.4 Extraction Error Detection

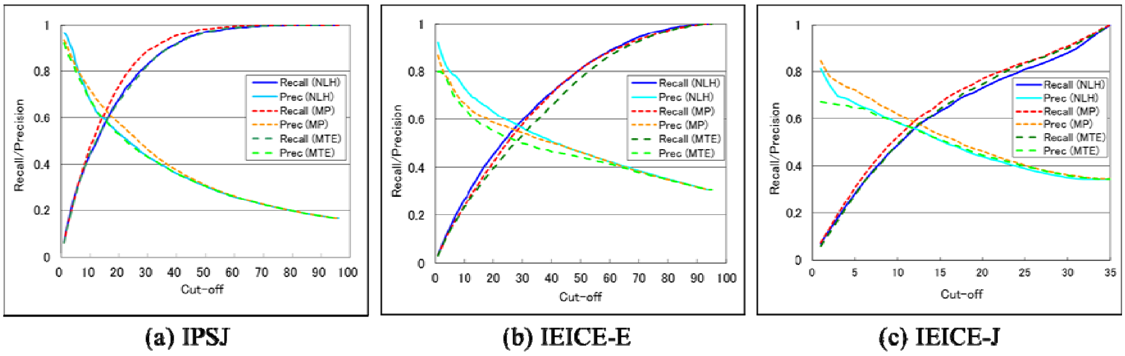
The task in the error detection experiment was to find erroneous label sequences among the sequences labeled by the CRF by using the three confidence measures. For this purpose, we first randomly chose 20, 100, and 300 samples from the training dataset and learned the CRF using them. Next, we made the CRF label the test sequences and then detected erroneously labeled sequences in accordance with each calculated confidence measure. Since all labeled test sequences were ranked by each confidence measure in ascending order, we detected top-*n*-ranked sequences as errors. Therefore, we calculated recall and precision of erroneous labeling detection as follows:

$$\text{Recall} = \frac{\text{\# of detected seqs actually including errors}}{\text{\# of erroneously labeled seqs}}, \quad \text{Precision} = \frac{\text{\# of detected seqs actually including error}}{\text{total \# of detected seqs}}.$$

Note here that “total # of detected seqs” equals the rank cut-off, *n*.

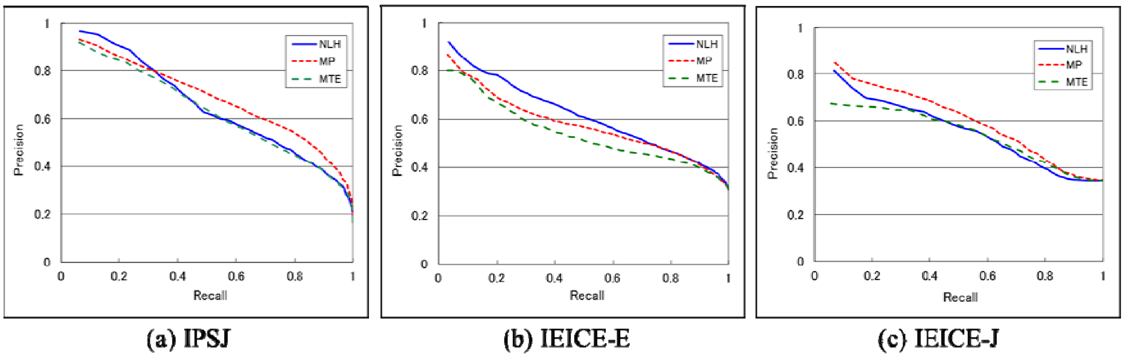
Figure 1 plots the recall and precision of error detection when we used 20 training samples for learning the CRF and applied each confidence measure to rank labeled test sequences with varying the rank cut-off *n*. Graphs (a), (b), and (c) respectively correspond to recalls and precisions for the IPSJ, IEICE-E, and IEICE-J datasets. In addition, Figure 2 shows the recall-precision curves for the three datasets when the three confidence measures were applied. As we can see in Figures 1 and 2, the retrieval effectiveness of erroneously labeled sequence search was better in IPSJ dataset than in IEICE-E and IEICE-J datasets. Comparing the three confidence measures, NLH and MP were better than MTE. For example, NLH showed the best performance among the three measures in Figure 2 (b) while MP showed the best performance in Figure 2 (c). However, NLH was best at low recall level and MP was best at high recall level in Figure 2 (a). It should also be noted that the recall did not saturate irrespective of the kinds of confidence measure until we detected all the test sequences in IEICE-J dataset as shown in Figure 1 (c).

Figure 3 also shows the recall and precision of error detection for the three datasets when we used 100 training samples for learning the CRF. In addition, Figure 4 shows the recall-precision curves for the experiments. As seen in Figures 3 and 4, the retrieval effectiveness of erroneously labeled sequence search was better in IPSJ dataset than in IEICE-E and IEICE-J datasets. Comparing the results in IEICE-E and IEICE-J datasets, those in IEICE-J were slightly better because its precision remained at about 0.6 while its recall increased to about 0.7 as shown in Figure 4 (c). Comparing the confidence measures, it is difficult to determine which one was the best measure for detecting errors because their performances differed in different datasets and at different recall levels even in the same dataset. For example, NLH showed the best performance among the three measures when its recall remained under about 0.6 in Figure 4 (a); however, it became the worst when its recall exceeded this recall level.



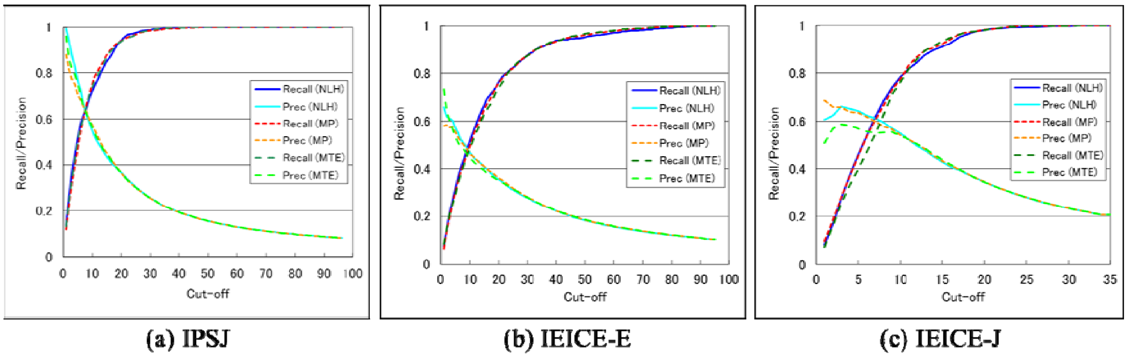
(a) IPSJ (b) IEICE-E (c) IEICE-J

Figure 1. Recall and precision w.r.t. rank cut-off *n* (# of training articles = 20)



(a) IPSJ (b) IEICE-E (c) IEICE-J

Figure 2. Recall-precision curves (# of training articles = 20)



(a) IPSJ (b) IEICE-E (c) IEICE-J

Figure 3. Recall and precision w.r.t. rank cut-off *n* (# of training articles = 100)

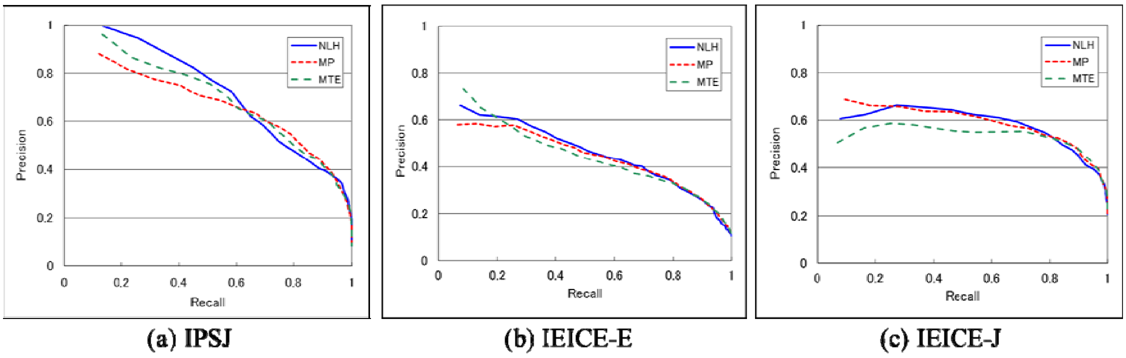


Figure 4. Recall-precision curves (# of training articles = 100)

Figure 5 shows the recall and precision of error detection with varying the rank cut-off *n* for the IPSJ and IEICE-E datasets when we used 300 training samples for learning the CRF. Figure 6 shows the recall-precision curves of the experiments. There are no graphs for IEICE-J dataset because the total number of articles in this dataset was 174. As seen in Figures 5 and 6, the retrieval effectiveness was much better in IPSJ dataset than in IEICE-E dataset. We can also say that NLH was the best performer irrespective of dataset throughout almost all recall levels. However, the performance in IEICE-E dataset shown in Figure 6 (b) was much poorer than those shown in Figures 2 (b) and 4 (b). This is considered partly because the extraction accuracy improved in accordance with the increase in the number of training samples as shown in Table 3. That is, we had to search for only 3.9 erroneously labeled sequences when using 300 training samples while 28.8 and 9.8 sequences when using 20 and 100 training samples, respectively. Figures 5 (a) and 6 (a) also show that the proposed measures such as NLH were good indicators of labeling confidence. Hence we could practically improve the labeling quality if we manually checked only a small fraction of CRF-labeled data with low confidence. We discuss the applicability of the confidence measures to controlling the tradeoff between assured bibliographic quality and necessary human intervention for achieving the quality in section 4.5.

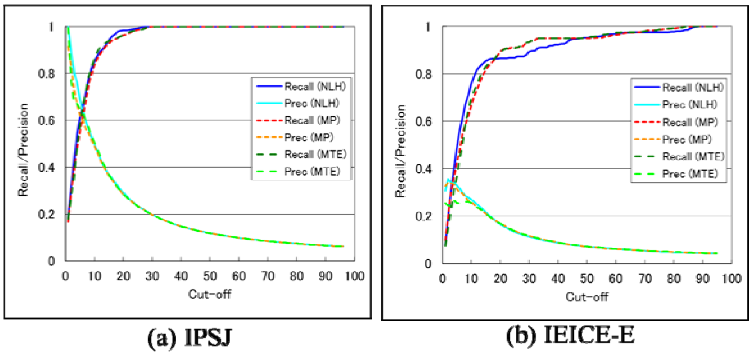
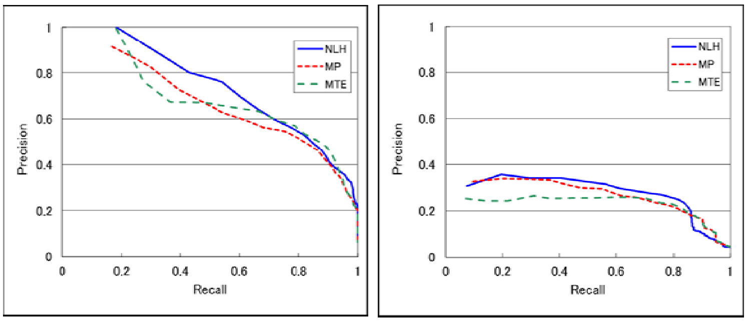


Figure 5. Recall and precision w.r.t. rank cut-off *n* (# of training articles = 300)



(a) IPSJ (b) IEICE-E

Figure 6. Recall-precision curves (# of training articles = 300)

4.5 Bibliographic Quality and Human Post-editing Cost

Finally, we examined the relationship between bibliographic quality assured by human post-editing and its cost. Let us suppose a task of realizing 97% bibliographic accuracy by the CRF-based bibliography extraction and the manual post-editing of detected articles as extraction errors by using the proposed confidence measures.

For example, Table 3 shows that the CRF-based extraction achieved 93.8 % accuracy and there were 6.0 articles with extraction errors when it was applied to IPSJ dataset with 300 training samples. We could achieve more than 97% accuracy if we detected 52% of the 6.0 articles with errors. As seen in Figure 5 (a), the recall exceeded 52% for the first time when the rank cut-off of NLH was four, and hence the four articles could be regarded as a manual checking cost to assure the 97% accuracy. By the same way, we estimated the deemed human cost, i.e., the number of articles that had to be manually checked after the CRF-based extraction for achieving 97% accuracy by using NLH for all the extraction results shown in Table 3. Table 6 summarizes the estimated number of articles that had to be manually checked and its ratio to the total.

Table 6 shows that we had to check many articles manually, i.e., 31% of the test articles for IPSJ dataset, 67% for IEICE-E dataset, and 92% for IEICE-J dataset when we used only 20 training samples. This is because the accuracies of the CRF-based extraction were poor as shown in Table 3. However, we could assure 97% accuracy by checking only 4% (IPSJ) and 3% (IEICE-E) of the test articles when we used 300 training samples.

Table 7 summarizes the estimated number of articles that had to be manually checked to assure 99% accuracy with human post-editing. As seen in the table, more than half of the test articles had to be checked, except in IPSJ dataset, when we used only 20 training samples, which is far from practical. However, we could assure 99% accuracy by checking only 10% (IPSJ) and 11% (IEICE-E) of the test articles when we used 300 training samples.

Table 6. # of articles that had to be manually checked and its ratio (%) to the total (in parentheses) for 97% accuracy

# of training samples	20	100	300
IPSJ	30 (31.3%)	8 (8.4%)	4 (4.2%)
IEICE-E	63 (66.6%)	18 (19.0%)	3 (3.2%)
IEICE-J	32 (92.0%)	13 (37.4%)	-

Table 7. # of articles that had to be manually checked and its ratio (%) to the total (in parentheses) for 99% accuracy

# of training samples	20	100	300
IPSJ	43 (44.9%)	17 (17.7%)	10 (10.4%)
IEICE-E	76 (80.3%)	49 (51.8%)	10 (10.6%)
IEICE-J	34 (97.7%)	18 (51.7%)	-

5. CONCLUSION

This paper reports an empirical evaluation of CRF-based bibliography extraction from scanned research papers. We specifically proposed three confidence measures for detecting bibliography labeling errors in order to assure bibliographic quality: i) normalized likelihood, ii) minimum probability of token assignment, and iii) maximum token entropy. Experiments showed that all the confidence measures reasonably indicated the labeling accuracies and could be used for labeling error detection for three academic journals used in the experiment. Moreover, this paper also discusses the tradeoff between the quality of bibliographic data assured by human post-editing of detected errors and its cost. The experiments showed that more than 99% accuracy could be assured for two of the journals if the post-editing was applied to about 10% of the articles detected as errors by using one of the proposed confidence measures. Note that the accuracies of the CRF-based bibliography extraction were about 94% for one journal and about 96% for the other by themselves.

We also observed the detection capabilities of the confidence measures were different in different journals, which suggests needs for a further investigation concerning this matter. Therefore, we plan to experiment on other journals for examining their applicability to various journals.

ACKNOWLEDGEMENT

This work was supported by two JSPS Grants-in-Aid for Scientific Research (B) (23300040 and 24300097), a JSPS Grant-in-Aid for Young Scientists (B) (23700119), and the Collaborative Research Program of National Institute of Informatics.

REFERENCES

Fellegi, I. P. and Sunter, A. B., 1969. A theory of record linkage. *Journal of American Statistical Association*, Vol. 64, No. 328, pp. 204-211.

Kudo, T. et al, 2004. Applying conditional random fields to Japanese morphological analysis. *Proc. of EMNLP 2004*, pp. 230-237.

Lafferty, J. et al, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of 18th International Conference on Machine Learning*, pp. 282-289.

Nagy, G. et al, 1992. A prototype document image analysis for technical journals. *IEEE Computer*, Vol. 25, No. 7, pp. 10-22.

Ohta, M. et al, 2010. Empirical evaluation of active sampling for crf-based analysis of pages. *Proc. of IEEE IRI 2010*, pp. 13-18.

Ohta, M. et al, 2008. Bibliographic element extraction from scanned documents using conditional random fields. *Proc. of ICDIM 2008*, pp. 99-104.

Okada, T. et al, 2004. Bibliographic component extraction using support vector machines and hidden markov models. *Proc. of ECDL 2004,* LNCS 3232, pp. 501-512.

Peng, F. and McCallum, A., 2004. Accurate information extraction from research papers using conditional random fields. *Proc. of HLT-NAACL 2004*, pp. 329-336.

Settles, B. and Craven, M., 2008. An analysis of active learning strategies for sequence labeling tasks. *Proc. of EMNLP 2008*, pp. 1070-1079.

Takasu, A., 2003. Bibliographic attribute extraction from erroneous references based on a statistical model. *Proc. of JCDL 2003*, pp. 49-60.

Xin, X. et al, 2008. Academic conference homepage understanding using constrained hierarchical conditional random fields. *Proc. of ACM Conf. on Information and Knowledge Management (CIKM '98)*, pp. 1301-1310.