Bibliographic Attribute Extraction

Ganesh D Gourshete[1] Prof. Sharvari S Govilkar[2]
[1]Department of Information Technology [2]Department of Computer Science & Engineering
[1,2]PIIT New Panvel, India

Abstract— Enormous amount of information is generated by the Proceedings. Online archives are repositories for technical reports. Here we review different bibliographic attribute extraction methods of PDF document to extract Title, Authors, Publication, Date, Pages, etc. Work done in the past can be classified into three major approaches: regular expression based heuristics, learning based algorithm and knowledge based systems.

Key words: Tokenization, Lexicons, Regular Expressions, HMM, CRF, SVM, DVHMM, Information extraction, Document processing

I. INTRODUCTION

A bibliography is a list of the sources used to get information for our report, dissertation or any other form of academic writing. It is essential that we acknowledge our debt to the sources of data, research and ideas on which we have drawn by including references to, and full details of, these sources in our work.

Entities and relationships are qualified by attributes representing their descriptive properties. Each entity has a set of attributes. Attributes can be viewed as data elements associated with bibliographic entities to describe and identify them clearly during the processes of creation, publication, production and cataloguing

To identify an entity uniquely, a combination of attributes such as the author's name, the title, the edition and the date of publication can be used. Even this combination might need further clarification, such as format.

Bibliographic entities are identified, related, located and accessed through their attributes. Thus, each attribute has a specific function or set of functions that are essential to the catalogue.

Some of the methods used are Tokenization, Lexicons, Regular Expressions, HMM, CRF, SVM, DVHMM, Information extraction, Document processing.

II. LITERATURE SURVEY

Many researchers have proposed multiple ways of extracting information from bibliographical references, some of the prominent ones are presented herein:

A. Bibliographic Attribute Extraction from Erroneous References Based On a Statistical Model:

Atsuhiro Takasu [1] has proposed a statistical model for attribute extraction that represents both the syntactical structure of references and OCR error patterns.

Bibliographic attribute extraction can be used in two ways: a) reference parsing in which attribute values are extracted from OCR-processed references for bibliographic matching. & b) reference alignment in which attribute values are aligned to the bibliographic record to enrich the vocabulary of the bibliographic database.

A statistical model for attribute extraction from erroneous references is a combination of OCR error models and syntactical structure represented by the DVHMM.

1) Observations:

DVHMMs have the preferred characteristic of being able to learn from non-aligned training data, which reduces the cost of preparing large training datasets.

2) Limitations:

As Japanese characters have very similar sizes in printed documents, substitution errors tend to occur in the OCR process whereas the proportion of framing errors may increase in English text.

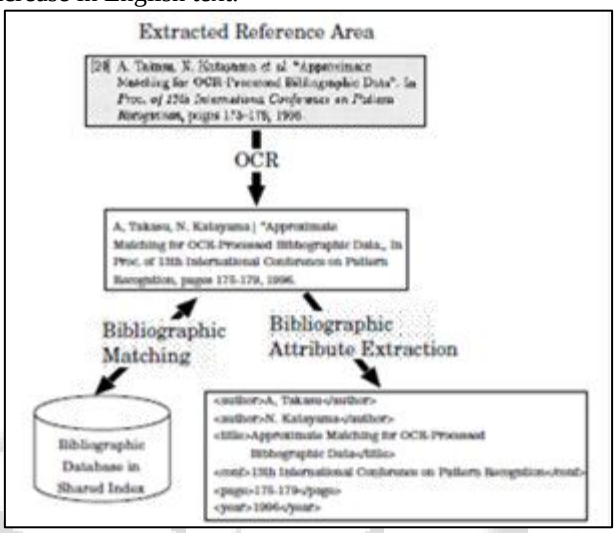

Fig. 1: Outline of bibliographic reference analysis [1]

Makes use of syntactical structures such as Component DVHMM [1] for Attributes and Delimiters. It is a combination of OCR error models and syntactical structure represented by the DVHMM, using the following steps:

1) construct a syntactical structure of bibliographic strings consisting of bibliographic attributes and delimiters,
2) replace each attribute component with a DVHMM that produces a pair comprising the attribute values and the recognized string, and
3) Replace each delimiter using a DVHMM in the same way as for a bibliographic attribute.

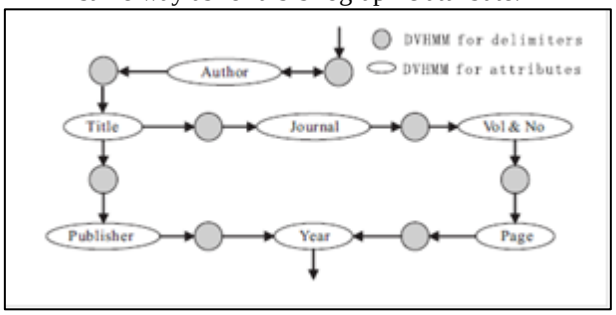

Fig. 2: Syntactical structure of bibliographic references [1]

B. Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields [2]:

Manabu Ohta , Takayuki Yakushi , Atsuhiro Takasu [2] proposed an automatic bibliographic element extraction

method for academic articles scanned with OCR markup which first labels text blocks as predetermined bibliographic elements and then further labels the characters in each labeled text block. Makes use of CRF.

This system extracts bibliographic data from scanned images by using the following steps:

1) Layout analysis and character recognition by using the OCR,
2) CRF-based text block labeling, and
3) CRF-based character labeling.

1) Observations:

The experiments showed that more than 99% of the author name strings were correctly extracted when the textual and layout features of the labeled and its previous and next characters were used.

CRF a statistical sequence modeling framework was proposed by Lafferty et al. [6] for part-of speech tagging and syntactical analysis. CRF outperforms 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.

Text block labeling uses layout information as the features for CRF, while the character labeling uses textual information in addition to the layout information as its features.

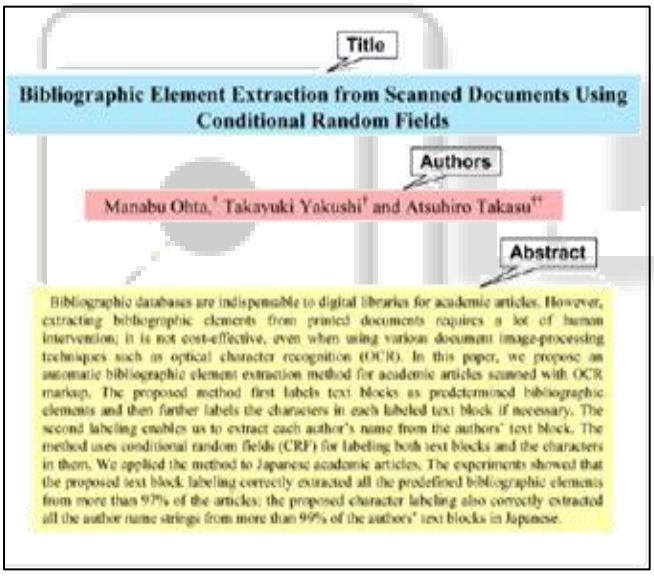

Fig. 3: Authors' block example in a title page [2].

The set of adopted feature templates that automatically generate a set of feature functions for the text block labeling. As shown in the table, all the features belong to the layout information of the title page of academic articles. We take into account not only the block location and size, but also the gap between the blocks and the size and number of the characters constituting each block.

C. A Simple Extraction Procedure for Bibliographical Author Field [3];

Pere Constans introduced a procedure based on the identification of simple, yet general templates or regular expressions [5, 7]. Author segments are recognized solely upon capitalization patterns and line break delimiters.

This process is grouped into two categories

1) Named knowledge representation and template mining.
2) General machine learning techniques such as Hidden Markov models, Support Vector Machines, and Conditional Random Fields, to infer segmentations.

1) Observations:

Approximately one hundred thousand PubMed citations [15] have been processed to analyze title words and author names. The procedure for the author field extraction has been implemented in the cb2Bib [9, 14] program, version 1.1.1, and it is part of its set of recognition algorithms.

D. A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching [4]:

Gupta D, Morris B, Catapano T & Sautter G describes automatic recognition, parsing and normalization of bibliographic references to enable easy search and retrieval of related information content.

The procedure for extracting is divided into 4 stages:

1) Obtaining Input Files in the TaxonX Schema
2) Obtaining a reference block in the document and identifying references.
3) Parsing the references into Author Name, Year, Title, Publication and other metadata.
4) Identification and matching of the corresponding citations with the references in the document.

1) Observations:

This paper studies the limitations of existing open-source solutions such as Paracite in bibliographic reference parsing along with the description and implementation of a new approach. This approach combines multiple techniques to enhance the accuracy and provide better recognition rates.

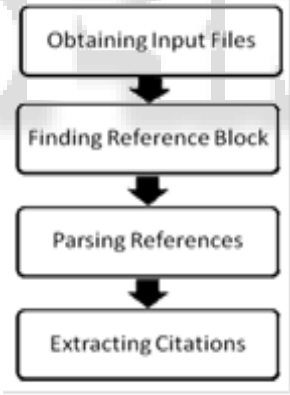

Fig. 4: Process outline

The text produced by using OCR may consists erros in test corpus. Also capitalization of the text cannot be relied upon as useful indicator for parsing. GoldenGate editor is used to add annotations and then export to an internal format which is converted to TaxonX by a purpose built XSLT transformation which is being used internally in Plazi. The TaxonX Schema based xml files are then used by the Reference Block Identifier as an input. Reference Parsing approach consists of:

1) Template matching i.e. use of regular expressions to Classify various portions of the text as particular fields.
2) Use of domain based information like a list of publications to classify a portion as a Publication in case of failure by the first approach.

III. CONCLUSIONS

Results of different papers reviewed are as follows for those papers discussed above

A. Bibliographic Attribute Extraction from Erroneous References Based On a Statistical Model [1]:

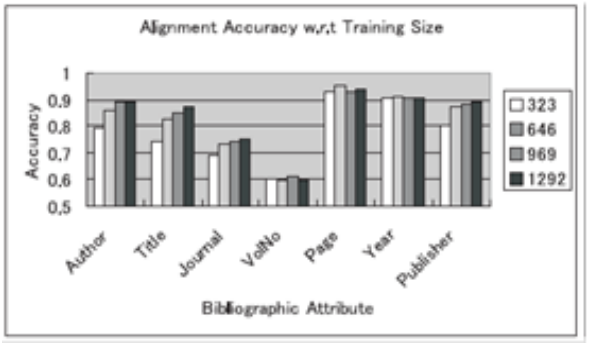

Fig. 5: Accuracy of alignment with respect to training data size

DVHMM can be trained using non-aligned pairs of training data, it has advantages in reducing the cost of preparing training data, a critical problem in rule-based systems. DVHMM trained with non-aligned training data has performance similar to one trained with aligned training data. The accuracies differ depending on the attributes and some of them are not high enough.

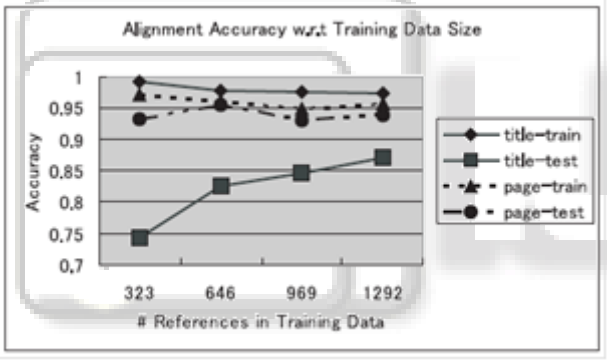

Fig. 6: Empirical and generalized accuracy with respect to training data size

B. Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields [2]:

Bibl. element	j-title	j-authors	j-abstract	e-title	e-authors	e-abstract	#Article
Accuracy (%)	99.27	98.53	99.27	99.27	99.02	99.51	97.56

Table 1: Text block labeling accuracy for test Data [2]

	Current		+Previous		+Next		+Both	
	2-tag	2+pos-tag	2-tag	2+pos-tag	2-tag	2+pos-tag	2-tag	2+pos-tag
#Author	99.18	99.18	99.45	99.73	99.27	99.45	99.36	99.82
#Article	97.52	97.52	98.76	99.07	97.83	98.14	98.45	99.07

Table 2: Japanese author's name labeling accuracy for test data [2]

The text block labeling, more than 98% of the bibliographic elements were correctly extracted: the accuracy based on the number of articles was 97.56%. After applying the CRF-based character labeling to the author/delimiter labeling for the authors' blocks in Japanese. The experiments showed that more than 99% of the author name strings were correctly extracted when the textual and layout features of the labeled and its previous and next characters were used. The experiments for each author labeling also showed that the tag set with the character position in the author/delimiter string, i.e., the 2+pos-tag set, was superior to the one without it, i.e., the 2-tag set.

C. A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching [4]:

Title	ParaCite Unmodified	Modified Code I	Modified Code II	Our Approach
Number of References(Hand Counted)	664	664	664	664
True Positives Identified	330	660	660	660
False Positives Identified	310	550	86	3
Percentage of False Positives	48.4%	45.5%	13.0%	0.5%
Percentage of False Negatives	50.4%	0.6%	0.6%	0.6%

Table 3: Reference Block Identification

Fields		Number Correctly Identified	Percentage Correctly Identified
Authors		660	99.5%
Year of Publication		660	99.5%
Title		504	76.0%
Publication	Combined Approach	490	73.9%
	Regular Expression Heuristic	445	91.8%
	Knowledge Based System	45	9.2%

Table 4: Reference Parsing

REFERENCES

[1] Atsuhiro Takasu, "Bibliographic Attribute Extraction from Erroneous References Based on a statistical Model," jcdl, pp.49, Third ACM/IEEE-CS Joint Conference on Digital Libraries(JCDL'03), 2003.

[2] Ohta, M., Yakushi, T, Takasu, A. "Bibliographic Element Extraction from Scanned DocumentsUsing Conditional Random Fields" In Proceedings of ICDIM, 2008, 99-104.

[3] Constans, Pere. "A Simple Extraction Procedure for Bibliographical Author Field," World Journal OF The International Linguistic Association, February, 2009, Available at http://arxiv.org/abs/0902.0755.

[4] Gupta, D.; Morris, B.; Catapano, T.; Sautter, G. "A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching," In Proceedings of IC3. 2009, 93-102.

[5] "Regular expression - Wikipedia, the free encyclopedia", July 19 2010. [Online].Available http://en.wikipedia.org/wiki/Regular_expression.[Accessed: July 22, 2010].

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In

Proc. of 18th International Conference on Machine Learning, pp. 282-289, 2001.
[7] RegExLab - Regular Expression Laboratory. [Online]. Available: http://www.regexlab.com/en/.Accessed: July 22, 2010].
[8] A. Takasu and K. Aihara. "DVHMM: Variable Length Text Recognition Error Model". In Proc. of International Conference on Pattern Recognition (ICPR02), Vol.III, pages 110–114, 2002.
[9] P. Constans. The cb2Bib: A tool for rapidly extracting unformatted biblographic references from email alerts, journal web pages, and PDF files, 2004 – 2009.
[10] Ray Smith (2007). "An Overview of the Tesseract OCR Engine" http://tesseract-ocr.googlecode.com/svn/trunk/doc/tesseracticdar2007.pdf
[11] CRFF Package http://crfpp.sourceforge.net/.
[12] Dan Moldovan and MihaiSurdeanu. On the Role of Information Retrieval and Information Extraction in Question Answering Systems. Information Extraction in the Web Era: NaturalLanguage Communication for Knowledge Acquisition and Intelligent Information Agents, 2003
[13] NLP tools for the context of developing IE systems: http://alias-i.com/lingpipe/web/competition.html.
[14] The cb2Bib http://www.molspaces.com/d_cb2bib-overview.php
[15] PubMed Central. http://www.ncbi.nlm.nih.gov/pubmed.
[16] NER http://nlp.stanford.edu:8080/ner/process
[17] Text mining: http://times.cs.uiuc.edu/czhai/pub/text-mining.ppt
[18] Tagging http://nlp.stanford.edu/software/tagger.shtml
[19] HMM http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html
[20] The Collection of CS Bibliographies. http://liinwww.ira.uka.de/bibliography
[21] Gavin Zhang http://ai.arizona.edu/mis510/slides/MIS510_2009Spring_SVM_and_CRF_short.ppt
[22] Machine Learning http://www.cs.waikato.ac.nz/ml/weka/
[23] SVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[24] CRF NER http://nlp.stanford.edu/software/CRF-NER.shtml
[25] CRFPP http://crfpp.sourceforge.net/

