Automated Labeling from Biomedical Journals published in Foreign Languages

Jongwoo Kim, Daniel X. Le, George R. Thoma
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA

ABSTRACT

An automated labeling (AL) module is developed to produce bibliographic records such as English title, vernacular title, author, affiliation, and English abstract from biomedical articles published in foreign language journals. Optical character recognition (OCR) output from scanned biomedical journals is used in this labeling process. Since frequently occurring words in a zone are important features, word lists are used as key features in the AL module. The AL module uses geometric and contextual features, and geometric relations between zones, as the basis for the rule-based labeling algorithms in the module. The algorithms uses 131 rules derived for foreign language journals. Experiments conducted with several medical journal articles show about 95% accuracy.

Keywords: Labeling Module, Rule-based Algorithm, Foreign Language Journals, OCR, MARS.

1. INTRODUCTION

Journal articles written in foreign languages contain bibliographic information such as vernacular title and/or English title of the article, author names, affiliations of authors, vernacular abstract and/or English abstract, and other information such as pagination, email address, etc. a few more than in typical English-language articles.
The process of extracting such information begins with scanning the article, converting the bitmapped image to text by optical character recognition (OCR), zoning the contiguous text, and then identifying the zones by labels (title, author, affiliation, abstract, etc.).

Many document labeling techniques [1-4] are based on the layout (geometric) structure and/or the logical structure of a document. Hones et al. [1] described an algorithm for layout extraction of mixed-mode documents. Taylor et al. [2] described a prototype system using a 'feature extraction and model-based' approach. Tateisi et al. [3] proposed a method based on stochastic syntactic analysis to extract the logical structure of a printed document. Kanungo et al [4] used Hidden Markov models for the layout analysis of a document image. Kim et al. used rule-based techniques [5-6] and fuzzy theory [7] to improve labeling accuracy. In this paper, we propose automated labeling algorithms to extract bibliographic information such as vernacular title, English title, author, affiliation, and English abstract from hard-copy journals articles using image and OCR processing, and rule-based technique.

Section 2 provides an overview of an existing system for automated data extraction, Section 3 describes the definition of layout and label types, Section 4 shows the structure of the proposed labeling module, Section 5 presents features used in the labeling module, and Section 6 describes the rules used in the labeling module in detail. Experimental results and a summary are in Sections 7 and 8.

2. MARS: AUTOMATED DATA EXTRACTION

The National Library of Medicine (NLM) has developed an automated system, Medical Article Records System (MARS), to extract bibliographic data from biomedical journal articles for its MEDLINE® database. MARS consists of several modules using OCR and document image analysis/understanding technologies. Scanned articles are processed by an OCR module which segments the articles into rectangular text zones using a commercial 5-engine OCR system [8]. This system also recognizes coordinates of zones, text lines, characters, bounding boxes of the characters, confidence levels, font sizes, and font attributes. A Zoning module (ZM) [9] corrects zoning errors produced by the OCR module, and a labeling module (LM) labels the zones as title, author, affiliation, or abstract. The results of the LM are processed by other modules, e.g., syntax reformat and reconcile modules, and the final results are uploaded to the MEDLINE database.

About 10% of the bibliographic records in the MEDLINE database are from journals published in foreign languages. These articles are usually composed of vernacular title and/or English title, author, affiliation, vernacular abstract, and/or English abstract. Some articles have both English title and English abstract, some have one, and some have neither. The vernacular title, English title, author, affiliation, and English abstract are important elements in the bibliographic records of articles written in foreign languages.

3. PAGE LAYOUT TYPES

There are around 470 journals published in foreign languages indexes at NLM for the MEDLINE database. The physical layout of the first page of articles from the journals can be categorized into several layout types, and the order in which the five important zones (vernacular/English titles, author, affiliation, and English abstracts) appear may be used to further categorize the layout types into subtypes. It is inefficient to create a single labeling algorithm in the LM that can handle all label types. Therefore, several journals are analyzed to classify several common label types and a labeling algorithm is developed for each common label type.

Figure 1 shows examples of common layout types consisting of a single column or a combination of multiple columns. Figures 1(a)-(e) show layout types 1, 11, 12, 121, and 122, respectively. Every gray block is composed of single column and the

numbers in the blocks indicate block numbers. Our current work focuses on layout types with "first regular" and "second regular" zone orders. "First regular" zone order has vernacular (or English) title followed by author, affiliation, and vernacular (or English) abstract. "Second regular" zone order has vernacular (or English) title followed by author, vernacular (or English) abstract, and affiliation.

Layout type 1, 11, 12, 121, and 122 journals are defined as label type A when every important label zone is in block 1 with "first regular" zone order. Layout type 11 journals are defined as label type B when lower affiliation zone is in block 2 and other important label zones in block 1 with "second regular" zone order. Six common label types, from type A to F, are defined and they are named as general label types. Other types, which are not included in general label types, are defined as arbitrary label types.

4. STRUCTURE OF THE LABELING MODULE (LM)

The LM is composed of several labeling algorithms, as shown in Figure 2. The MARS database has tables to save information of each journal as shown in Table 1. In Figure 2, an input journal is processed by the Scan and OCR modules and the results of the OCR module are processed by the Zoning module (ZM). When a zoning result of a journal article is input to the LM, ISSN of the journal is sent to the database and information of the journal corresponding to the ISSN in the JournalName table (Table 1) is sent to the LM so that a labeling algorithm related to the journal is activated. Label type in the JournalName table is used for the activation. When the input is one of label type journals (e.g., the journals that have Label Type A, B, and D in the first to third rows in Table 1), the LM activates one of the related labeling algorithms. When an input journal does not belong to one of the label types (e.g., the journal that has Label Type AB (AB means arbitrary type)) in the fourth row in Table 1), the LM activates the labeling algorithm for Type AB.

JournalName table shown in Table 1 contains journal information such as label type and published language of the journal. The first row in Table 1 means that journal name is "Comptes Rendus Biologies", label type is A, and the language is french. Since the label type of this journal is A, the LM selects "Label Algorithm for Type A" in the LM for the operation.

5. FEATURES USED IN THE LABELING MODULE

The features used in this experiment are divided into two categories: geometric and non-geometric features. Geometric features are based on location, order of appearance, and dimensions of a zone. Contextual features are derived from contents of zone and font characteristics. For example, vernacular title zone is usually located in the top half of the first page of an article (geometric feature), and usually has the largest font size (non-geometric feature). Font sizes of author and affiliation zones are usually smaller than those in the vernacular title zone (non-geometric feature).

Since a zone is often characterized by the words in the zone, word matching is an important function in the LM. For example, a zone has a higher probability of being labeled as "affiliation" when it has words representing country, city, and school names. Also, a zone located between the words "abstract" and "keywords" has a higher probability of being labeled as "abstract" than other labels. A word list with most frequent foreign words in vernacular title zones (Table 2) and a word list with most frequent English words in title zones (Table 3) are used to distinguish vernacular/English titles and abstracts.

Fifteen tables with word lists have been collected and some of them are shown in Table 4. The Ternary Search Tree algorithm (TST) [10] is used as a search engine for the word matching.

Table 5 shows some of the features extracted from the OCR output for the LM. Some features are extracted using the word lists (Table 5) and the TST algorithm, and others are extracted directly from the OCR output.

6. RULES USED IN THE LABELING MODULE

Rule-based algorithms in our system have 131 rules for the LM. The LM in MARS is designed for five label zones in an article: English title, vernacular title, author, and English abstract. The remaining zones are labeled as "others". Four kinds of rules are developed for each label. Rules 1, 2 and 3 are different for each label, while rule 4 is the same for all. In the first step, a zone is labeled by rule 1. For example, when a zone has a higher Probability of Correct Identification (PID) for title (PID >= 100), the zone is labeled as title. The PIDs are derived from features related to each of the five labels.

In the second step, previous labeling results are rechecked by rule 4. For example, when two different zones are both labeled as author, (i.e., One zone is located between title and upper affiliation. The other is located between upper affiliation and abstract.), a zone between upper affiliation and abstract is removed from the author zones.

In the third step, rules 1, 2, and 4 are applied again to make sure that at least one zone is labeled as title, author, abstract, and upper affiliation or lower affiliation. For example, when a zone, which is initially labeled as author, does not have any information about author (Nbr_Middlename=0 and Nbr_Degree =0), its location is then used to do the labeling. That is, the label as author is inferred by the facts that (a) it does not contain information suggestive of a title or upper affiliation, and (b) it is located between title and upper affiliation zones.

In the fourth step, problems caused by zoning errors such as splitting a zone into multiple zones are handled by all rules. Any remaining unlabeled zones are labeled.

Rules for Vernacular Title

Rule 1:
1. Font_Size == Max_Font_Size
2. Nbr_Degree < T_{gvt1} (=3) or Pct_Degree < T_{gvt2} (=10)
3. Nbr_Middlename<T_{gvt3} (=3) or Pct_Middlename < T_{gvt4} (=10)
4. Nbr_Author < T_{gt5} (=3) or Pct_Author < T_{gvt6} (=10)
5. Coordinate_Upper < Height_Article/3 and Coordinate_Lower < Height_Article/2
6. NbrTitle < Nbr_ForeignWord
7. Pct_Title < T_{gvt7} (=30)
8. If all of above conditions are satisfied {
   If (Font_Size == Max_Font_Size) PID = 100
   Else If (|Font_Size-Max_Font_Size|<T_{gvt8} (=3)) PID=99
   Else PID=(Font_Size–Min_Font_Size)×10/ (Max_Font_Size – Min_Font_Size)
   }

Rule 2:
If (PID < 100) pick a zone having the highest PID for title.

Rule 3:
1.Distance from a zone to title is smaller than that of any other labels.

2.Font_Size, Med_Line_Height, and Med_ Line_Space of a zone must be similar to those of title zone.

Rule 4:

Coordinate_Upper of vernacular title < Coordinate_Upper of author <Coordinate_Upper of affiliation < Coordinate_Upper of abstract

Rules for Author

Rule 1:

1. Coordinate_Upper < Height_Article/2
2. Font_Size <= Font_Size of Title
3. Nbr_Word >= T_{ga1} (=3)
4. Nbr_Affiliation<=T_{ga2} (=3) or Pct_Affiliation<=T_{ga3} (=30)
5. If all of above conditions are satisfied {
   If (Pct_Degree+Pct_Middlename+Pct_Author > T_{ga4} (=28)) PID = 100;
   Else PID=(Pct_Degree+Pct_Middlename+ Pct_Author) × 100/28
   If (Pct_Capitalcharacter > T_{ga5} (=50)) {
      If (PID > 50) PID = 100;
      Else PID = PID + PID/2
   }
}

Rule 2:

If (PID < 100) pick a zone having the highest PID for author.

Rule 3:

1. Distance from a zone to Author zone is smaller than any other label zones.
2. Font_Size, Med_Line_Height, and Med_Line_Space of a zone must be similar to those of author zone.

Rule 4:

Same as rule 4 for vernacular and English titles.

Rules for English Abstract

Rule 1:

1. Zone is bigger than titles, author, affiliation zones
2. Font_Size <= Font_Sizes of Vernacular and English Titles
3. Pct_Lexicon >= T_{gab1} (=30)
4. If all of above conditions are satisfied {
   If (Previous Zone has a word "Abstract") PID=100
   If (Current Zone has a word "Abstract") PID=100
   If (Next Zone has a word "Introduction") PID=100
   If (Next Zone has a word "Keyword") PID=100
}

Rule 2:

If (PID < 100) pick a zone having the highest PID for English abstract.

RULE 3

1. Distance from a zone to abstract zone is smaller than any other label zones
2. Font_Size, Font_Attribute, Med_Line_Height, Med_Line_Length, and Med_Line_Space of a zone must be similar to those of English zone.

Rule 4:

Same as rule 4 for vernacular and English titles

.

7. EXPERIMENTAL RESULTS

Figures 3, 4, and 5 show examples of the labeling process in LM. Figure 3(a) is an input journal article with label type equal to A. Figure 3(b) is the zoning result. The results are shown with red bounding boxes. Figure 3(c) shows the labeling result. Figure 4(a) is an input journal article. Figure 4(b) is the zoning result. Figure 4(c) shows the labeling result. Figure 5(a) is an input journal article. Figure 5(b) is the zoning result. Figure 5(c) shows the labeling result.

260 journal articles from 22 journals are used for the experiment of the LM. The results are shown in Table 5. The result shows 96.2% labeling accuracy in vernacular title, 98.8% in English title, 96.5% in author, 92.3% in affiliation, and 91.9% in English abstract zones. The English abstract has the highest error rate. However, most of them are caused by zoning error. The affiliation has the highest error rate caused by labeling module. It is because the affiliation word list table is based on addresses written in English. This can be easily solved if we collect more words from affiliations written in foreign language. In total, incorrect OCR output generates 0.1% of the errors and incorrect zoning generates 1.4% of the errors. The error related to the LM is 3.7%. In overall performance, the proposed labeling module shows 94.5% labeling accuracy.

8. SUMMARY

This paper describes a rule-based module to label the first pages of scanned medical journals published in foreign languages for the automated production of bibliographic citation records for MEDLINE at the National Library of Medicine. The module is composed of several labeling algorithms to process the different label types. The labeling algorithms in the modules employ both geometric and non-geometric zone features. The algorithms also use geometric relations among zones. The proposed modules show relatively high accuracy in labeling.

9. REFERENCES

[1] F. Hones and J. Lichter, "Layout Extraction of Mixed Mode Documents," **Machine Vision and Applications** 7, 1994, pp. 237-246.

[2] S. Taylor, R. Fritzson, and J. Pastor, "Extraction of Data from Preprinted Forms," **Machine Vision and Applications** 5, 1992, pp. 211-222.

[3] Y. Tateisi and N. Itoh, "Using Stochastic Syntactic Analysis for Extracting a Logical Structure from a Document Image," **Proc. IEEE Int. Conf. Neural Networks**, Vol. 2, 1994, pp. 391-394.

[4] T. Kanungo, S. Mao, "Stochastic Language Model for Style-Directed Physical Layout Analysis of Documents," **IEEE Transactions on Image Processing**, vol. 12, no. 5, May 2003, pp. 583-596.

[5] J. Kim, D. Le, and G. Thoma, "Automated Labeling Document Images," **Proc. of SPIE**, Vol. 4307, Document Recognition and Retrieval VIII, San Jose, CA January 2001, pp.111-122.

[6] J. Kim, D. Le, and G. Thoma, "Automated labeling algorithms for biomedical document images**," Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics**, Vol. V, Orlando FL, July 2003, pp.352-57.

[7] J. Kim, D. Le, and G. Thoma, **"**Automated labeling of bibliographic data extracted from biomedical online journals," **Proc. SPIE Electronic Imaging**, January 2003. SPIE Vol. 5010, pp. 47-56

[8] Prime Recognition Inc., **Prime OCR Access Kit Guide**, version 2.70, San Carlos, CA, 1997.

[9] S. Hauser, D. Le and G. Thoma,. "Automated zone correction in bitmapped document images," **Proc. SPIE**: Document Recognition and Retrieval VII, San Jose, CA, January 2000, SPIE Vol. 3976, 248-58

[10] J. Bentley and B. Sedgewick, "Ternary Search Trees,"**Dr. Dobb's Journal**, April 1998, pp. 20-25.

Table 1. JournalName Table.

ISSN	Journal Title	Label Type	Language
1631-0691	Comptes Rendus Biologies	A	French
0009-9074	La Clinica terapeutica	B	Italian
0036-3634	Salud Publica de Mexico	D	Spanish
0004-0614	Archivos Espanoles de Urologia	AB	Spanish

Table 2. Word List (Foreign Words)

Foreign Word	Country
aber	German
alla	Italian
als	German
bis	German
con	Italian
dann	German
en	French
fur	German
La	French
les	French
nei	Italian
um	German
vor	German
zu	German

Table 3. Word List (Key of Title)

Key of Title
after
an
as
at
but
by
for
from
in
of
or
the
to
with

Table 4. Word List Tables.

Table Name	Words in the Table
Rubric	Review, Orginal Article, etc.
KeyOfTitle	Study, Case, Method, etc.
AcademicDegree	Ph.D., MD, RN, etc.
Affiliation	University, Department, Lab, etc.
Abstract	Abstract, Summary, etc.
StructuredAbstract	Aim, Result, Conclusion, etc.
Keyword	Keyword, Index word, etc.
KeyOfAffiliation	Corresponding, To whom, etc.
ForeignWord	Abr, alla ,fur, etc.

Table 5. Features used in the Labeling Module.

Zone Features	Variable Names
Geometric Features:	
Zone coordinates	Coordinate_Left, _Right, _Upper, _Lower
Median value of height, length and space of lines	Med_Line_Height, _Length, _Space
Biggest and smallest font sizes in an article	Max_Font_Size, Min_Font_Size
Difference between the bottom and top coordinates of the bottom-most and top-most zone	Height_Article
Zone order in sequence of top left edge	(A number)
Contextual Features:	
Number of characters and words	Nbr_Character, Nbr_Words
Number of Capital characters	Nbr_Capitalcharacter
Dominant Font Attribute and Font Size	Font_Attribute, Font_Size
Number of Title, "study", "method", etc.	Nbr_Title
Number of "M.D.", "Ph.D.", "RN", etc.	Nbr_Degree
Number of Middle Name, "Jr", "Sr", "II", etc.	Nbr_Middlename
Number of Author Name, "Kim", "Le", etc.	Nbr_Author
Number of city, state, country, school, etc.	Nbr_Affiliation
Number of "abstract", "summary", etc.	Nbr_Abstract
Number of "review", "article", etc.	Nbr_Rubric
Number of ForeignWord, "fur", "alla", "de", etc.	Nbr_ForeignWord
Percentage of Nbr_Title per word	Pct_Title
Percentage of Nbr_Degree per word	Pct_Degree
Percentage of Nbr-ForeignWord	Pct_ForeignWord
Percentage of Nbr_Middlename per word	Pct_Middlename
Percentage of Nbr_Author per word	Pct_Author
Percentage of Nbr_Affiliation per word	Pct_Affiliation
Percentage of Nbr_Capitalcharacter per zone	Pct_Capitalcharacter

Table 6. Test Results of the Labeling Module.

Label	Number of Zone	Error				Error (%) of each label
		OCR	Zoning	Labeling	Total	
Vernacular Title	180			7	7	3.8
English Title	85			1	1	1.2
Author	198		2	5	7	3.5
Affiliation	221			17	17	7.7
English Abstract	148	1	10	1	12	8.1
Total Label Zones	**832**					
Total Error		**1**	**12**	**31**	**44**	
Total Error (%)		**0.1**	**1.4**	**3.7**	**5.3**	

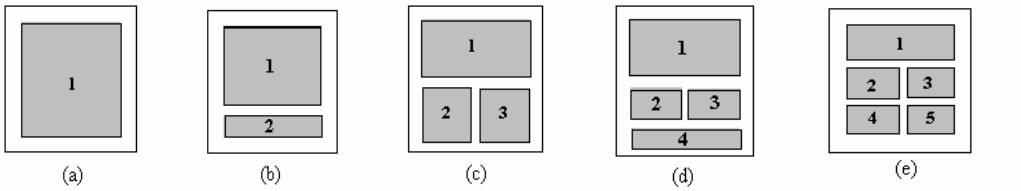


Figure 1. Examples of journal layout types. The numbers in the gray block show block numbers. (a) Layout Type 1, (b) Layout Type 11, (c) Layout Type 12, (d) Layout Type 121, (e) Layout Type 122

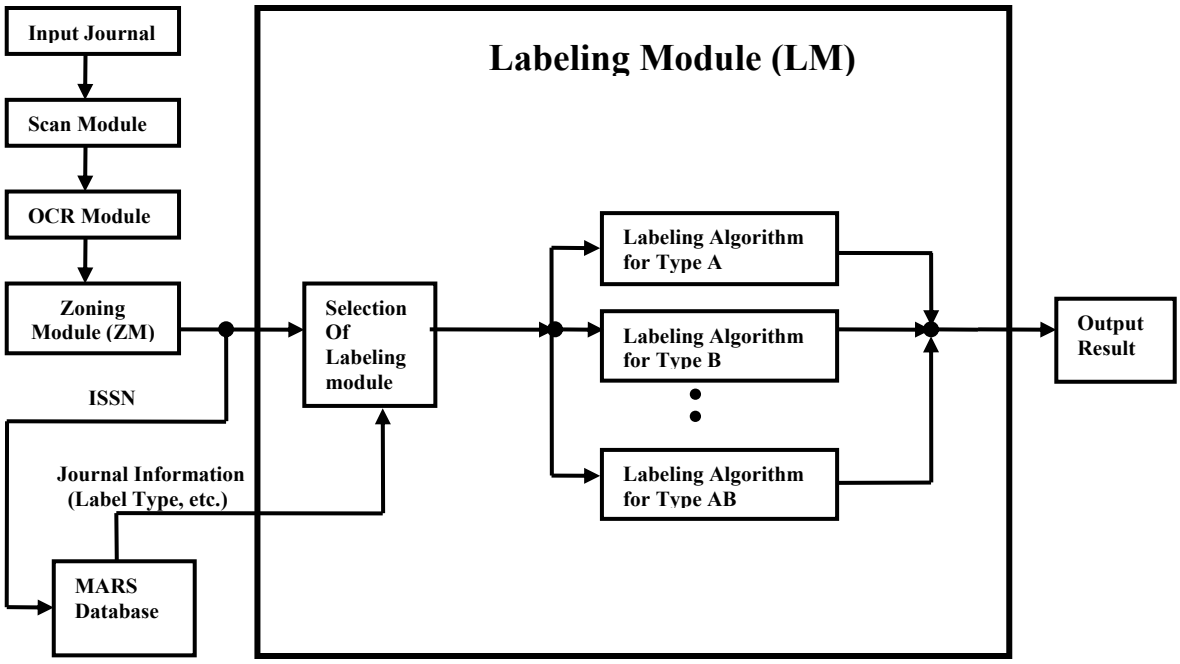
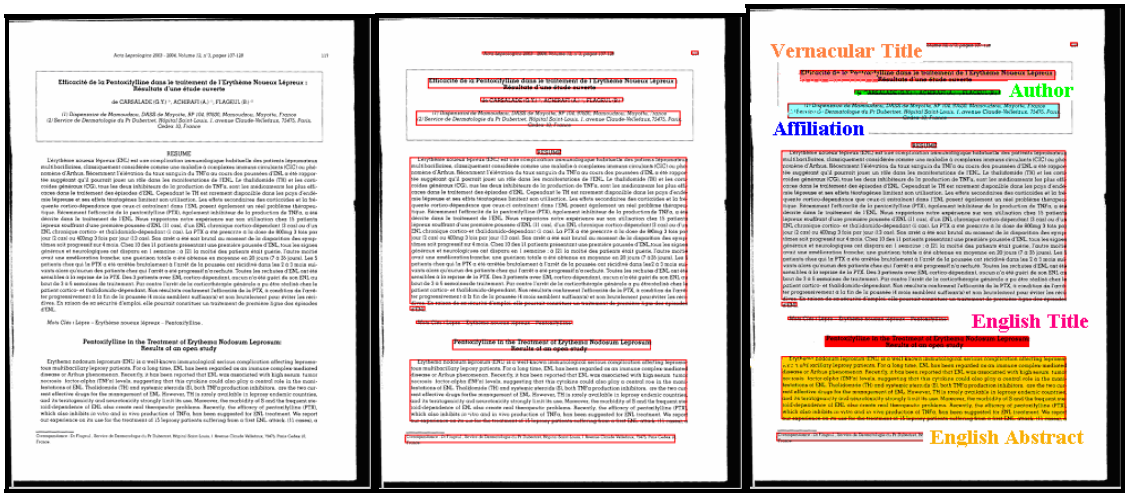


Figure 2. Structure of the Labeling module (LM).



(a) (b) (c)

Figure 3. Example 1. (a) Input image, (b) Zoning result, and (c) Labeling result.

(a) (b) (c)

Figure 4. Example 2. (a) Input image, (b) Zoning result, and (c) Labeling result.

(a) (b) (c)

Figure 5. Example 3. (a) Input image, (b) Zoning result, and (c) Labeling result.