Automated Labeling Algorithms for Biomedical Document Images

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

ABSTRACT

The National Library of Medicine (NLM) has developed an automated system, named Medical Article Records System (MARS), to process bibliographic data (title, authors, affiliation, abstract, etc.) in biomedical journal articles for its MEDLINE® database. This paper describes a labeling module in the MARS, which automatically extract the bibliographic data in biomedical journal articles. The labeling module is composed of two sub modules: General label type module (GLTM) and Arbitrary label type module (ALTM). Six label types, which are commonly used in the journals, are collected from several thousand journals. Journals are classified as general label types if label types of the journals belong to one of the six label types. Otherwise, journals are classified as arbitrary label types. The GLTM processes journals that belong to general label types and the ALTM processes journals that belong to arbitrary label types. Rule-based algorithms are used for both modules and the rules are derived from analysis of several journal articles and features extracted from the optical character recognition (OCR) results. There are 126 rules derived for the GLTM and 49 rules for the ALTM. Experiments conducted with several medical journal articles show relatively accurate labeling results.

Keywords: Labeling Module, Zoning Module, Rule-based Algorithm, OCR, MARS.

1. INTRODUCTION

Journal articles usually consist of text zones and non-text. Text zones of interest in a journal article contain bibliographic information such as the title of the article, author names, affiliations of authors, abstract and other descriptive information. The process of automatically extracting such information begins with scanning the article, converting the bitmapped image to text by optical character recognition (OCR), zoning the contiguous text to create the text zones, and then identifying the zones by labels (title, author, affiliation, abstract, etc.).

Most proposed document labeling techniques [1-3] 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. Other techniques [4,5] have used the outputs of OCR to further improve labeling accuracy. In this paper, we describe an automated labeling technique to label text zones as title, author, affiliation, and abstract using integrated image and OCR processing, and rule-based technology.

Section 2 provides a system overview, Section 3 presents features used in the automated labeling, and Section 4 describes the structure and rules used in the labeling module in detail. Experimental results and conclusion are in Sections 5 and 6.

2. OVERVIEW OF MARS

The MARS is composed of several modules and it adapts OCR and document image analysis/understanding technologies. Figure 1 shows some of the modules. Scan module scans journal articles, and saves the scanned images in the MARS database. The scanned articles are processed by an OCR module. The OCR module segments the articles into rectangular text zones using a commercial 5-engine OCR system [6], and recognizes coordinates of zones, text lines, characters, bounding boxes of the characters, confidence levels, font sizes, and font attributes. Zoning module (ZM) corrects zoning errors produced by the OCR module, and the labeling module (LM) labels the zones as article 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.

3. DEFINITION OF LAYOUT TYPES OF JOURNALS

NLM's MEDLINE® database contains bibliographic records from over 4,500 journals. The physical layout of the first page of articles in these journals can be categorized into several layout types, and the order in which the five important zones (title, author, upper affiliation, lower affiliation, and abstract) appear may be used to further categorize the layout types into subtypes. The subtypes are defined as label types in this paper. It is impossible and inefficient to make a single labeling algorithm in the LM that can handle all label types of journals. Therefore, several thousand journals are analyzed to classify several common label types and a labeling algorithm is developed for each common label type.

Figure 2 shows examples of common layout types consisting of a single column or a combination of multiple columns. Figures 2(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 title followed by author, affiliation in the upper portion of a page (upper affiliation), and abstract. "Second regular" zone order has title followed by author, abstract, and affiliation in the lower portion of a page (lower 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 divided into two sub modules, GLTM and ALTM, as shown in Figure 3. The MARS database has tables to save information of each journal as shown in Table 1 and 2. In Figure 3, 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 general 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 in the GLTM. When an input journal belongs to arbitrary label type (e.g., the journal that has Label Type AB (AB means arbitrary type)) in the fourth row in Table 1), the LM activates the ALTM and reads all information of the journal in the JournalSpecificInformation table in Table 2.

JournalName table shown in Table 1 contains journal information such as label type, width, and height of the physical journal. The first row in Table 1 means that journal name is “Adverse drug reactions and toxicological reviews”, label type is A, width is 6.125 inches and height is 9.250 inches. Since the label type of this journal is A, the LM selects “Label Algorithm for Type A” in the GLTM for the operation.

JournalSpecificInformation table shown in Table 2 contains more specific information of each journal such as character font sizes and bounding boxes of label zones. This information is used only for the ALTM. In Table 2, the first column is the ISSN of a journal. Issues and Pages columns represent numbers of journal issues and journal articles used to collect the data. LabelID column indicates the label of the feature. X and y columns are upper left coordinates of a bounding box of a label zone. Height and Width columns are height and width of a bounding box of a label zone. Font size column is character font size of a label zone. Ratio column is frequency ratio of a feature (from column four to nine) in a journal.

A journal “Psychological methods” is in the last row in Table 1 and journal specific information of the journal is in Table 2. In Table 2, the first row shows that 100% (Ratio=1.0) of the title zones has character font size of 12 point. The second row shows that upper left coordinate of a title zone-bounding box is (1.31, 1.82), and height and width of the box are 0.45 and 4.79 inches, respectively. The third and fourth rows show that there are two character font sizes in author zones. 36% (Ratio=0.36) of the author zones uses size 9 and 64% (Ratio=0.64) of author zones uses size 10. The fifth and sixth rows show that there are two bounding boxes for author zones. The symbol “x” means there is no related data in the table.

There are some noisy data in the JournalSpecificInformation table since OCR results are used to collect the information and the OCR module frequently generates errors. Therefore, the LM only considers information (row) in the table, which has Ratio greater than or equal to 0.1 in this experiment.

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. Non-geometric features are derived from contents of zone and font characteristics. For example, 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 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. Fifteen tables with word lists have been collected and some of them are shown in Table 3. The Ternary Search Tree algorithm (TST)[7] is used as a search engine for the word matching.

Table 4 shows some of the features extracted from the OCR output for the LM. Some features are extracted using the word lists (Table 3) and the TST algorithm, and others are extracted directly from the OCR output.

6. RULES USED IN THE LABELING MODULE

6.1 Rules for the General Label Type Module (GLTM)

Rule-based algorithms are used and 126 rules are generated for all labeling algorithms in the GLTM. The LM in the MARS are interested in five label zones in an article: title, author, affiliation in the upper portion of a page (upper affiliation), affiliation in lower portion (lower affiliation), and 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. The rule-based algorithm consists of four steps as shown in Table 5 and the thresholds in the rules can be changed in each step.

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 of 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. The detailed rules for some labels are shown below and some variables used in the rules are defined in Table 4.

Rules for Title (GLTM)

Rule 1:
1. Font_Size == Max_Font_Size
2. Nbr_Degree < T_{gt1} (=3) or Pct_Degree < T_{gt2} (=10)
3. Nbr_Middlename < T_{gt3} (=3) or Pct_Middlename < T_{gt4} (=10)
4. Nbr_Author < T_{gt5} (=3) or Pct_Author < T_{gt6} (=10)
5. Coordinate_Upper < Height_Article/3 and Coordinate_Lower < Height_Article/2
6. If all of above conditions are satisfied {
   If (Font_Size == Max_Font_Size) PID = 100
   Else If (|Font_Size-Max_Font_Size|<T_{gt7} (=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 title < Coordinate_Upper of author < Coordinate_Upper of affiliation < Coordinate_Upper of abstract

Rules for Author (GLTM)

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 title.

6.2 Rules for the Arbitrary Label Module (ALTM)

There are several journals that do not belong to the general label types. A single module, which is named arbitrary label type module (ALTM), is developed to process all existing arbitrary label type journals. The rules used in the ALTM module are similar to those used in the GLTM, but there are some differences. First, rules in the ALTM use journal specific information (JournalSpecificInformation table). Second, rules in the ALTM do not consider geometric relations between important labels. 49 rules are developed for the ALTM. The followings are rules for title and author. Similar rules are applied to other labels (affiliation and abstract). The labeling algorithm consists of two steps and the thresholds in the rules can be changed in each step. In the first step, rules 1, 2, and 3 are used to label zones that have a higher PID for each label. In the second step, rules 1, 2, 3, and 4 are used to label remaining unlabeled zones such as split zones caused by the ZM.

Rules for Title (ALTM)

Rule 1:
Select candidates of title zone using the information in the JournalSpecificInformation table.
1. A zone should be located inside of one of the bounding boxes of title zones.
2. Font size of a zone should be one of title font sizes.

Rule 2:
1. Nbr_Degree < T_{at1} (=3) or Pct_Degree < T_{at2} (=10).
2. Nbr_Middlename < T_{at3} (=3) or Pct_Middlename < T_{at4} (=10)
3. Nbr_Author < T_{at5} (=3) or Pct_Author < T_{at6} (=10)
4. If all of above conditions are satisfied {
   If (Font_Size == Max_Font_Size) PID = 100
   Else PID = (Font_Size – Min_Font_Size) ×100/ (Max_Font_Size – Min_Font_Size)
   }
5. If (PID = 100) label the zone as title.

Rule 3:
If (PID < 100) label a zone having the highest PID for title.

Rule 4:
1.Distance from a zone to title is next to each other.
2.Font_Size, Med_Line_Height, and Med_Line_Space of a zone must be similar to those of title zone.

Rules for Author (ALTM)

Rule 1:
Select candidates of author zone using the information in the JournalSpecificInformation table.
1. A zone should be located inside of one of the bounding boxes of author zones.
2. Font size of a zone should be one of author font sizes

Rule 2:
1. Nbr_Word >= T_{aa1} (=3)
2. Nbr_Affiliation <= T_{aa2} (=3) or Pct_Affiliation <= T_{aa3} (=30)
3. If all of above conditions are satisfied {
   If (Pct_Degree + Pct_Middlename + Pct_LastName > T_{aa4} (=28)) PID = 100;
   Else PID=(Pct_Degree+Pct_Middlename+Pct_LastName) × 100/28
   If (Pct_Capitalcharacter > T_{aa5} (=50)) {
   If (PID > 50) PID = 100;
   Else PID = PID + PID/2
   }
   }
5. If (PID = 100) label the zone as author.

Rule 3:
If (PID < 100) pick a zone having the highest PID for author.

Rule 4:
1.Distance from a zone to title is next to each other.
2.Font_Size, Med_Line_Height, and Med_Line_Space of a zone must be similar to those of title zone.

7. EXPERIMENTAL RESULTS

Figure 4 shows an example of the labeling process of the GLTM. Figure 4(a) is an input journal article with label type equal A. Figure 4(b) is the zoning result. The results are shown with red bounding boxes. Figure 4(c) shows the labeling result. Figure 5 shows an example of the labeling process of the ALTM. Figure 5(a) is an input journal article. Figure 5(b) is the zoning result. The results are shown with red bounding boxes. Figure 5(c) shows the bounding boxes of each label. Zones in the bounding box of each label can be candidates of each label zone. Figure 5(d) shows the labeling result. 11,651 journal articles from 1,054 journals are used for the experiment of the GLTM. Incorrect OCR output generates 0.3% of the errors and incorrect zoning generates 2.0% of the errors. The error related to the LM is 1.0%. In overall performance, the proposed GLTM module shows 96.7% labeling accuracy. Seven journal issues are collected to evaluate the performance of the ALTM. 161 articles in the journal issues are used to extract journal specific information of the journals and 76 articles are used to test the performance. The ALTM module shows 100, 95.64 and 95.85 % of labeling accuracy in title, author, and abstract, respectively. However, the module shows 63.13% labeling accuracy in affiliation. Since several arbitrary label type journals have more than one affiliation zone in an article and limited articles are used to estimate bounding boxes of the label, the estimated bounding boxes of affiliation do not cover all affiliation zones in test articles. This error can be easily solved by increasing the number of articles to estimate journal specific information. Overall, the ALTM shows promise in labeling arbitrary label type journals.

8. CONCLUSION

This paper describes a rule-based module to label the first pages of scanned medical journals for the automated production of bibliographic citation records for MEDLINE® in the National Library of Medicine. The module is composed of two sub modules to process the general label type and arbitrary label type journals. The labeling algorithms in the modules employ both geometric and non-geometric zone features. As the basis for the set of rules, the algorithms for the GLTM use geometric relations among zones while the algorithms for the ALTM use journal specific information. The proposed GLTM and ALTM modules show relatively accurate labeling results.

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] J. Liang et. al., "The Prototype of a Complete Document Image Understanding System," **Int. Assoc. Pattern Recognition Workshop on Document Analysis System**, Malvern, PA, 1996.

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

[7] 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	Width (Inches)	Height (Inches)
0964-198X	Adverse drug reactions and toxicological reviews	A	6.125	9.250
1076-898X	Journal of experimental psychology. Applied	B	8.500	11.000
0959-440X	Current Opinion in Structural Biology	D	8.500	11.000
1082-989X	Psychological methods	AB	8.500	11.000

Table 2. JournalSpecificInformation Table. (* Unit is inch.)

ISSN	Issues	Pages	LabelID	X *	Y *	Height *	Width *	Font Size	Ratio
1082-989X	2	15	Title	x	x	x	x	12	1.00
1082-989X	2	15	Title	1.31	1.82	0.45	4.79	x	1.00
1082-989X	2	15	Author	x	x	x	x	9	0.36
1082-989X	2	15	Author	x	x	x	x	10	0.64
1082-989X	2	15	Author	0.55	6.54	3.23	3.65	x	0.89
1082-989X	2	15	Author	1.35	2.31	0.11	1.23	x	0.11
1082-989X	2	15	Affiliation	x	x	x	x	17	1.00
1082-989X	2	15	Affiliation	0.58	1.35	0.78	6.83	x	1.00
1082-989X	2	15	Abstract	x	x	x	x	10	1.00
1082-989X	2	15	Abstract	1.55	2.31	2.34	4.88	x	1.00

Table 3. 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.

Table 4. 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)
Non-Geometric 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 “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
Percentage of Nbr_Degree per word	Pct_Degree
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 5. Sequential Process for Applying Rules in the Labeling Module.

Step	Rules used	Rule Description
1	Rule 1	Use Probability of Correct Identification (PID). Each label has its own PID equation. Example: When a zone has a higher PID for title (PID >= 100), the zone is labeled as title.
2	Rule 4	Use geometric relations between zones. Example: When two different zones are both labeled as author but they are not close to each other, one zone is then removed from the author zones.
3	Rules 1, 2, and 4	Label at least one zone as title, author, abstract, or affiliation. Example: When there is no zone labeled as author and a zone labeled as author does not have any information about author (Nbr_Middlename = 0 and Nbr_Degree = 0), geometric relations and non-geometric features are used to do the labeling. That is, when a zone between title and affiliation does not have any information about title and affiliation, the zone is labeled as author.
4	Rules 1, 2, 3, and 4	Label other remaining zones. The OCR segmentation problem of splitting a zone (such as title zone) into multiple zones (multiple title zones) is handled by all rules and any remaining unlabeled zones are labeled in this step.

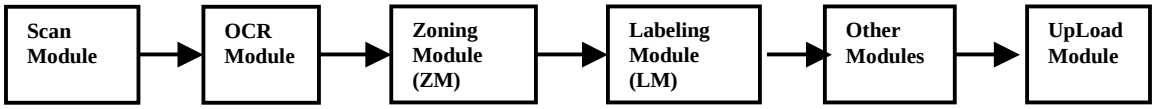


Figure 1. Overview of the MARS.

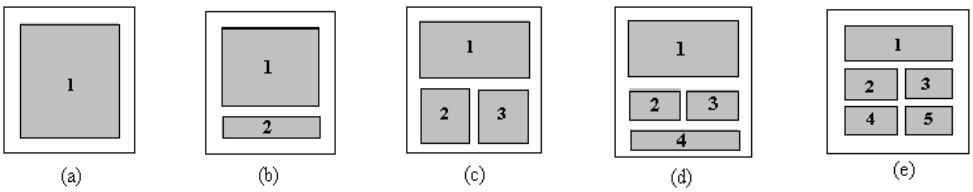


Figure 2. Examples of journal layout types. The numbers in the gray block show block numbers. (a) Layout Type 1, (b) Layout Type 11, (c) Layout Type 12, (d) Layout Type 121, (e) Layout Type 122

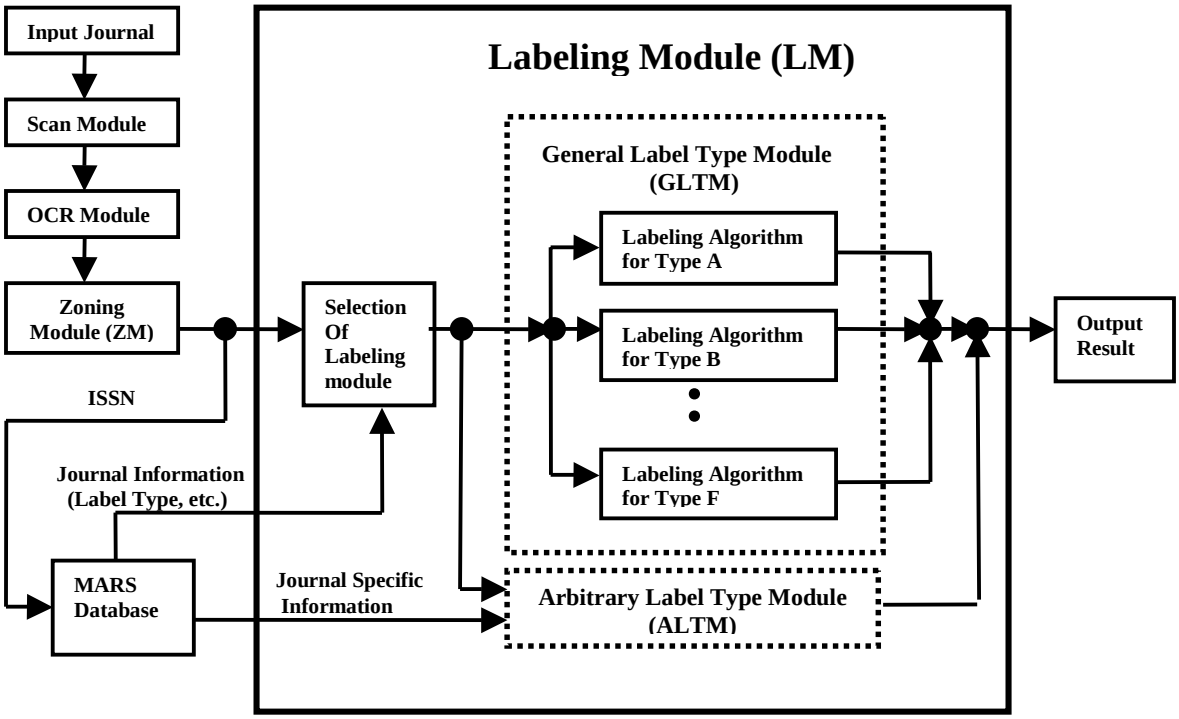
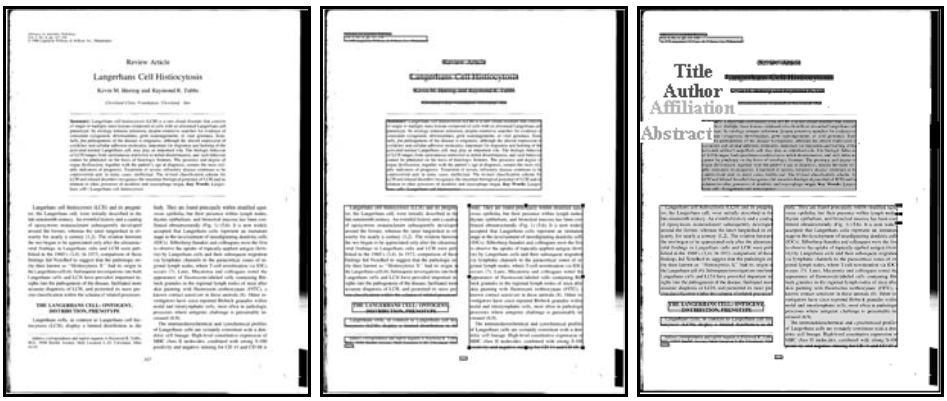
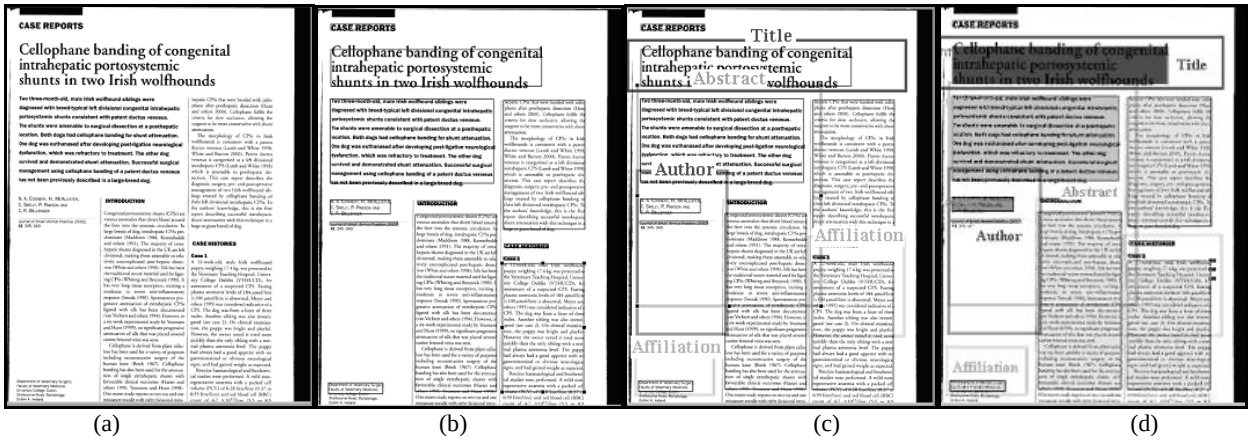


Figure 3. Structure of the Labeling module (LM).



(a) (b) (c)

Figure 4. Example of a labeling algorithm in the GLTM. (a) Input image, (b) Zoning result, and (c) Labeling result.



(a) (b) (c) (d)

Figure 5. Example of a labeling algorithm in the ALTM. (a) Input image, (b) Zoning result, (c) Bounding Boxes of important labels, and (d) Labeling result.