Automated Labeling in Document Images

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

ABSTRACT

The National Library of Medicine (NLM) is developing an automated system to produce bibliographic records for its MEDLINE® database. This system, named Medical Article Record System (MARS), employs document image analysis and understanding techniques and optical character recognition (OCR). This paper describes a key module in MARS called the Automated Labeling (AL) module, which labels all zones of interest (title, author, affiliation, and abstract) automatically. The AL algorithm is based on 120 rules that are derived from an analysis of journal page layouts and features extracted from OCR output. Experiments carried out on more than 11,000 articles in over 1,000 biomedical journals show the accuracy of this rule-based algorithm to exceed 96%.

Keyword: OCR, automated data entry, automated zoning, automated labeling, rule-based algorithm, MARS, NLM

1. INTRODUCTION

Journal articles usually consist of text zones, non-text zones such as graphics, or a mixture of both. Text zones of interest in a journal article contain bibliographic information such as the title of the article, author names, the institutional 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.). Aside from automatically extracting bibliographic data from biomedical journals, the application of interest here, document labeling techniques are required in a variety of other applications such as automated document searching, automated document delivery, and electronic publishing (converting papers from one format to another or modifying manuals and references, etc.).

Most proposed 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. Tsujimoto et al.[3] presented a technique based on the transformation from a geometric structure to a logical structure. Tateisi et al.[4] proposed a method based on stochastic syntactic analysis to extract the logical structure of a printed document. Tang et. al.[5] proposed conceptual and concrete structures to describe document structures. Other techniques[6,7,8] 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. SYSTEM OVERVIEW

The automated labeling module described here is part of our second-generation MARS system used in the automated extraction of bibliographic fields from scanned journal articles. The steps leading to automatically labeling the bibliographic fields of interest are:

- Scan journal articles.
- Perform optical character recognition (OCR).
- Apply automated zoning (AZ) using rules derived from OCR output.
- Apply automated labeling (AL). Every zone is associated with a label such as title, author, etc.

*Correspondence: Email: kimjw@ceb.nlm.nih.gov

Scanned binary document images are segmented into rectangular text zones using a commercial 5-engine OCR system[9], and in addition to the recognized characters, each zone contains zone coordinates, text line information, character bounding boxes, confidence levels, font sizes, and attributes. AZ follows OCR, and from the outputs of OCR and AZ, features for each zone are calculated and input to an AL system for label classification. Since AZ is another research area, we will focus on our AL algorithm in this paper.

The features calculated for AL include geometric ones, such as height and width of zone, zone area, and position of a zone in a page, as well as non-geometric ones such as character statistics and word recognition against word lists. These features are extracted from the output of the OCR system that provides information at the zone, line, and character levels, as given below:

Zone Level	Zone boundaries, number of text lines
Line Level	Line boundaries, number of characters, average character height
Character Level	8-bit code for each recognized character, confidence level (*1 = lowest, 9 = highest*), bounding box, font size, font attributes (*normal, bold, underlined, italics, superscript, subscript, and fixed pitch*)

3. FEATURES USED IN THE AUTOMATED LABELING MODULE

Most features and rules derived for labeling algorithms are based on an analysis of the page layout for each journal, and generic typesetting knowledge for English text[10]. Both geometric and non-geometric features are considered here.

Geometric features are based on a zone's location, order of appearance, and dimensions. For example, title zone is usually located in the top half of the first page of an article, and usually has the largest font size[11]. In most cases, the title is followed by author, affiliation and abstract, in that order. Font sizes of author and affiliation zones are usually smaller than those in the title zone. Non-geometric features are derived from contents of zone, aggregate statistics, and font characteristics.

Since a zone is often characterized by the words in the zone, word matching is an important function in the AL module. 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 positioned between the words "abstract" and "keywords" has a higher probability of being labeled as "abstract" than other labels. Fifteen database tables with word lists have been assembled as shown in Table 1 and the Ternary Search Tree algorithm [12] is used as a search engine for the word matching. Some of the features extracted from the OCR output for the AL module are shown in Table 2.

4. STRUCTURE AND RULES USED IN THE AUTOMATED LABELING MODULE

NLM's MEDLINE database contains bibliographic records from over 4,300 journals. The physical layout of the first page of articles in these journals can be categorized into several types, and the order of occurrence of the four zones of interest (title, author, affiliation, and abstract) may be used to further categorize the layout into subtypes. A single rule-based algorithm that can handle all journals is unlikely; the rules have to be tailored for each layout type and have to accommodate layouts that may be geometrically similar but have different orders of occurrence of the zones.

Figure 1 shows examples of common layout types consisting of a single column or a combination of multiple columns. Figures 1(a), 1(b), and 1(c) show Type 1, Type 12, and Type 122, respectively. Our current work focuses on layout types with a "regular" zone order in which the title is followed by author, affiliation in the upper portion of a page, and abstract. We define journals of Type 1, Type 12, and Type 122 with this regular zone order as Type 10000, Type 12000, and Type 12200 and make rules for these three types first. Journals with other layout types will be handled in the future.

In Type 10000, every zone is located in a single column. In Type 12000, the abstract is located in the first column and/or in the second column of the two-column zones; other labels are located in the one-column zone on the top of the page. In Type12200, the abstract is located in both columns of the two-column zones in the middle of the page and other labels are located in the one-column zone on the top of the page. Since the major difference between the three layout types is the location of the abstract, we create three special rules for labeling abstract zones and common rules for labeling the other zones (title, author and affiliation).

Figure 2 shows the structure of the AL module. A database of journal information is used to save relevant information on every journal. It contains journal name, ISSN, layout type, physical size, and some identifying feature particular to a journal. When the zoned page information is input to the AL module, the information on the journal in the database is also sent to the AL module so that the algorithm specific to the particular journal type is activated. An identifying feature in the database may be used to handle exceptional case journals. For example, most journals have the word "keywords" after the abstract, and so is an important feature to label abstract zones. However, in some journals "keywords" appears before the abstract, and this fact would be in the database of journal information and used in such cases to label the abstract zone correctly.

In the current MARS system, we are interested in five zone labels in an article: title, author, affiliation printed 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 type. Rules 1, 2 and 3 are different for each label classification, while rule 4 is the same for all. The rule-based algorithm consists of four steps as shown in Table 3.

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 specific to each of the five fields of interest, and are defined in Sections 4.1 to 4.5.

In the second step, previous labeling results are rechecked by rule 4. For example, when two different zones are both labeled as author but one zone is located between title and upper affiliation and the other is located between upper affiliation and abstract, a zone between upper affiliation and abstract is then removed from the author category.

In the third step, in addition to rule 2, rules 1 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 initially labeled as author does not have any information about author (Number_Middlename = 0 and Number_Degree = 0), its location is then used to do the labeling. That is, its 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 a labeled title and an upper affiliation.

In the fourth step, problems caused by errors in zoning such as splitting a zone into multiple zones are handled by all rules, and any remaining unlabeled zones are labeled. The detailed rules for each label type are shown below, and variables are defined in Table 2.

4.1 RULES FOR TITLE

RULE 1

1. Number_Headtitle == 0
2. Font_Size == Max_Font_Size
3. Number_Degree < 3 or Percent_Degree < 10
4. Number_Middlename < 3 or Percent_Middlename < 10
5. Coordinate_Upper < Height_Article /3
6. Coordinate_Lower < Height_Article /2
7. If all of above conditions are satisfied {

```
    If ( Font_Size == Max_Font_Size )        PID = 100
    Else If ( | Font_Size - Max_Font_Size | < 3 )      PID = 99
    Else         PID = ( Font_Size – Min_Font_Size ) ×100/( Max_Font_Size – Min_Font_Size )
}
Else {
    PID = 0
}
```

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, Font_Attribute, 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

4.2 RULES FOR AUTHOR

RULE 1

1. Coordinate_Upper < Height_Article /2
2. Font_Size <= Font_Size of Title
3. Number_Word >= 2
4. Number_Affiliation <= 3 or Percent_Affiliation <= 30
5. Number_Headtitle == Number_Abstract == Number_Introduction == 0
6. If all of above conditions are satisfied {

```
    If ( Percent_Degree+Percent_Middlename > 28 )
       PID = 100;
    Else
       PID= ( Percent_Degree+Percent_Middlename ) × 100/28
    If ( Percent_Capitalcharacter > 50 ) {
       If ( PID > 50 )
          PID = 100;
       Else
          PID = PID + PID /2
    }
}
Else {
    PID = 0
}
```

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, Font_Attribute, 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 described in section 4.1.

4.3 RULES FOR UPPER AFFILIATION

RULE 1

1. Upper_Coordinate < Height_Article /2
2. Lower_Coordinate < Height_Article×3/4
3. Number_Word >= 2
4. Number_Degree < 3 or Percent_Degree < 30
5. Number_Middlename < 3 or Percent_Middlename < 30
6. Percent_Capitalcharacter < 50
7. Number_Headtitle == Number_Abstract == Number_Introduction==0
8. If all of above conditions are satisfied {

```
    If ( Number_Affiliation >= 2 ) {
       If ( Percent_Affiliation >= 30 )
          PID =100;
       Else
          PID = Percent_Affiliation×100/30
    }
    Else {
       If ( Percent_Affiliation >= 30 )
          PID =50;
       Else
```

```
            PID = Percent_Affiliation×50/30
        }
    }
    Else {
        PID = 0
    }
```

RULE 2
If (PID < 100), pick a zone having the highest PID for upper affiliation.

RULE 3
1. If (PID > 25 and the next zone has Number_Received ==1) PID = 100.
2. Distance from a zone to upper affiliation zone is smaller than any other label zones.
3. Font_Size, Font_Attribute, Med_Line_Height, and Med_Line_Space of a zone must be similar to upper affiliation zone.

RULE 4
Same as rule 4 for title described in section 4.1.

4.4 RULES FOR LOWER AFFILIATION

RULE 1
1. Upper_Coordinate > Height_Article /2
2. Lower_Coordinate > Height_Article ×3/4
3. Number_Words >= 2
4. Number_Degree < 3 or Percent_Degree <= 25
5. Number_Middlename < 3 or Percent_Middlename <= 25
6. Percent_Capitalcharacter < 50
7. Number_Headtitle == Number_Abstract == Number_Introduction == 0
8. If all of above conditions are satisfied {

```
        If ( Number_Affiliation >= 2 ) {
            If( Percent_Affiliation >= 30 )
                PID =100
            Else
                PID = Percent_Affiliation×100/30
        }
        Else {
            If( Percent_Affiliation >= 30 )
                PID =50
            Else
                PID = Percent_Affiliation×50/30
        }
        If ( Number_Affiliation > 0 ) PID=PID+50
    }
    Else {
        PID = 0
    }
```

RULE 2
If(PID < 100), pick a zone which has the highest PID for lower affiliation.

RULE 3
1. Distance from a zone to lower affiliation zone is smaller than any other label zones.
2. Font_Size, Font_Attribute, Med_Line_Height, and Med_Line_Space of a zone must be similar to lower affiliation zone.

RULE 4
Same as rule 4 for title described in section 4.1.

4.5 RULES FOR ABSTRACT

RULE 1
1. Zone is bigger than title, author, upper affiliation, and lower affiliation zones**.**
2. If all of above conditions are satisfied {
    If (Previous Zone has Number_Abstract == 1)
      PID = 100
    If (Current Zone has Number_Abstract == 1)
      PID = 100
    If (Previous Zone has Number_Received == 1)
      PID = 100
    If (Next Zone has Number_Introduction == 1)
      PID = 100
    If (Next Zone has Number_Keyword == 1)
      PID = 100
  }
  Else {
    PID = 0
  }

RULE 2
None

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 abstract zone.

RULE 4
Same as rule 4 for title described in section 4.1.

5. EXPERIMENTAL RESULTS

120 rules were generated for the three layout types (Type10000, Type12000, and Type12200) and 1,054 journals consisting of 11,651 articles were selected for experiments. Figure 3 shows an example of the automated labeling procedure. Figures 3(a)-(c) show an input article of Type 10000, the result of AZ, and the result of AL, respectively. Errors are of three types: "bad" OCR (incorrect conversion), automated zoning, and automated labeling.

Figure 4 shows a consequence of incorrect OCR. Figure 4(a) is an input image and Figure 4(b) is the OCR output, magnified for easy viewing. Each character should have its own bounding box as the result of OCR. However, many characters in the author name area do not have bounding boxes because the OCR failed to correctly perform symbol isolation. Figure 5 shows another example of incorrect OCR performance. Figure 5(a) is an input image and Figure 5(b) is the magnified OCR output. Some characters (a, r, t, and m) are not recognized by OCR system, thereby creating a zoning problem as shown in Figure 5(c). AL module only labeled the left portion of the zone as author.

Figure 6 shows an example of AZ error. Figures 6(a)-(d) show a part of the input image, the magnified OCR output, AZ result, and AL result. As shown in these figures, the affiliation area was zoned incorrectly and only one of the affiliation zones is correctly labeled.

Figure 7 shows examples of errors caused by the AL algorithms not tailored to the specific journals being processed. In Figure 7(a), an "introduction" zone was labeled as abstract since the size of the title zone is larger than that of the abstract zone (i.e., number of lines in title zone is greater than that in abstract) and the word "abstract" above the abstract zone was not recognized by the OCR system. In Figure 7(b), two introduction zones are labeled as abstract since the word "background" is located before the introduction zones and many journals use "background" as "abstract".

Table 4 gives experimental results. 0.3% of the errors is due to incorrect OCR output and 2.0% is due to wrong AZ. Since incorrect OCR outputs usually create automated zoning problems, many of the AZ errors are basically caused by incorrect OCR. The error rate attributed to the AL algorithm itself is 1.0% when OCR and automated zoning are correct. The reason for the high error rate in the affiliation field is that many journals use italics in this field, and current OCR systems do not recognize italics well. In overall performance, our AL module delivers an accuracy of 96.7%.

6. CONCLUSION

This paper describes a rule-based algorithm to automatically label zoned bibliographic fields in scanned pages from medical journals. This automated labeling is a key stage in the automated production of bibliographic citation records for MEDLINE, the flagship database of the National Library of Medicine. This algorithm employs both geometric and non-geometric zone features as well as geometric relations between zones as the basis for the set of rules, which numbers about 120. The accuracy of correct labeling is 96.7% for 1,054 journals tested to date. Research continues toward tailoring the algorithm to expand the number of journals covered.

REFERENCES

1. F. Hones and J. Lichter, "Layout Extraction of Mixed Mode Documents," *Machine Vision and Applications* 7, pp. 237-246, 1994.
2. S. Taylor, R. Fritzson, and J. Pastor, "Extraction of Data from Preprinted Forms," *Machine Vision and Applications* 5, pp. 211-222, 1992.
3. S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," *Proc. IEEE*, Vol. 80, No. 7, pp. 1133-1149, 1992.
4. 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, pp. 391-394, 1994.
5. Y. Tang and C. Suen, "Document Structures: A Survey," *Document Image Analysis*, World Scientific, pp.1081-1111, 1994.
6. T. Hu et. al., "A Prototype for Extracting Logical Elements from Tables of Contents of Journals," *Int. Assoc. Patt. Recog. Workshop on Doc. Analysis System*, Malvern, PA, 1996
7. J. Liang et. al., "The Prototype of a Complete Document Image Understanding System," *Int. Assoc. Patt. Recog. Workshop on Document Analysis System*, Malvern, PA, 1996.
8. D. Le, J. Kim, G. Pearson, and G. Thoma, "Automated Labeling of Zones from Scanned Documents," *Proceedings 1999 Symposium on Document Image Understanding Technology*, pp.219-226, 1999.
9. Prime Recognition Inc., Prime OCR Access Kit Guide, version 2.70, San Carlos, CA, 1997.
10. G. Nagy, "At the Frontiers of OCR," *Proc. IEEE*, Vol. 80, No. 7, pp. 1093-1100, 1992.
11. J. H. Ling, "The Title Page as The Source of Information for Bibliographic Description: An Analysis of its Visual and Linguistic Characteristics," University Texas at Austin, 1987.
12. J. Bentley and B. Sedgewick, "Ternary Search Trees,"*Dr. Dobb's Journal*, pp. 20-25, April 1998.

Table 1. Word List Tables.

Table Name	Words in the Table
HeadTitle	Review, Orginal Article, etc.
KeyOfTitle	Study, case, method, etc.
Author	Smith, John, Kim, etc.
Degree	Ph.D., MD, RN, etc.
Affiliation	University, Department, Institute, etc.
Abstract	Abstract, Summary, Background, etc.
SubAbstract	Aim, Result, Conclusion, etc.
Keyword	Keyword, Index word, etc.
Received	Received, Revised, Accepted, etc.
Introduction	Introduction, Introduzione, etc.
KeyOfAffiliation	Corresponding, Address, To whom, etc.
KeyOfLowAffiliation	Mail, fax, tel, etc.
Date	January, February, 2000, etc.
Publisher	Elsevier, John Wiley, etc.
JournalName	Diabetes, endocrinology, etc.

Table 2. Features used in the Automated Labeling Module.

Zone Features	Variable Names
Geometric Features:	
Zone coordinates	Coordinate_Left, _Right, _Upper, _Lower
Zone height and width	Height_Zone, Length_Zone
Median value of height, length and space of lines	Med_Line_Height, _Length, _Space
Median value of space between lines	Med_Line_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 lines	Number_Line
Number of characters and words	Number_Character, Number_Words
Number of Capital characters	Number_Capitalcharacter
Dominant Font Attribute and Font Size	Font_Attribute, Font_Size
Confidence of characters	(A number between 1 and 9)
Number of “M.D.”, “Ph.D.”, “RN”, etc.	Number_Degree
Number of Middle Name, “Jr”, “Sr”, “II”, etc.	Number_Middlename
Number of city, state, country, school, etc.	Number_Affiliation
Number of “abstract”, “summary”, etc.	Number_Abstract
Number of “keywords”, “index words”, etc.	Number_Keyword
Number of “review”, “article”, etc.	Number_Headtitle
Number of “received”, “accepted”, etc.	Number_Received
Number of “Introduction”, “Introduzione”, etc.	Number_Introduction
Percentage of Number_Degree per word	Percent_Degree
Percentage of Number_Middlename per word	Percent_Middlename
Percentage of Number_Affiliation per word	Percent_Affiliation
Percentage of Number_Capitalcharacter per zone	Percent_Capitalcharacter

Table 3. Sequential Process for Applying Rules in the AL 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 category.
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 (Number_Middlename = 0 and Number_Degree = 0), geometric relations and non-geometric features are then 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.

Table 4. Results of the proposed AL Module.

Error Type	Label Field					
	Title	**Author**	**Affiliation**	**Abstract**	**Totals**	**% of Error**
Bad OCR	4	3	26	1	34	0.3
Automated Zoning	47	35	84	65	231	2.0
Automated Labeling	23	24	24	47	118	1.0
Totals	74	62	134	113	384	3.3

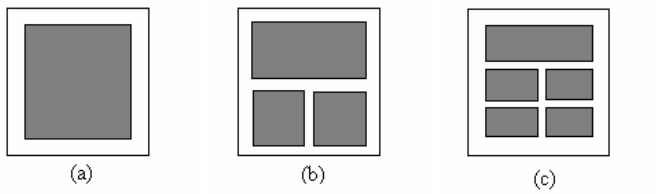


Figure 1. Examples of journal layout types. (a) Type 1, (b) Type 12, and (c) Type 122.

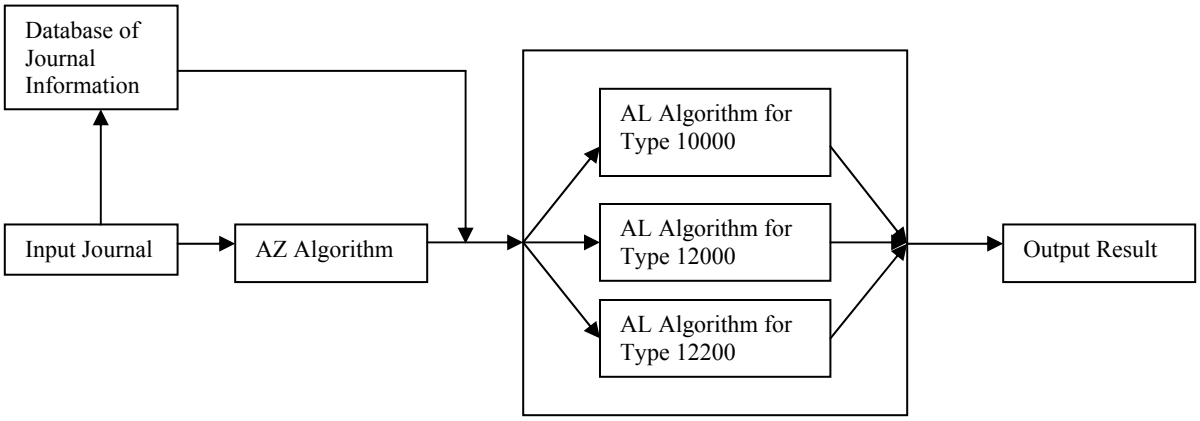
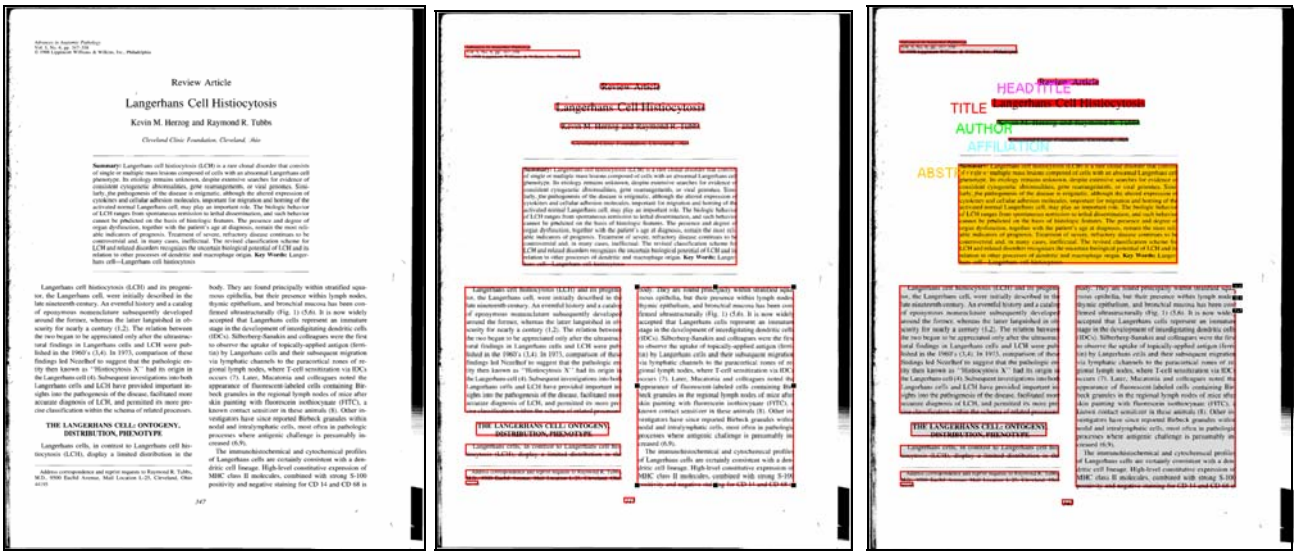
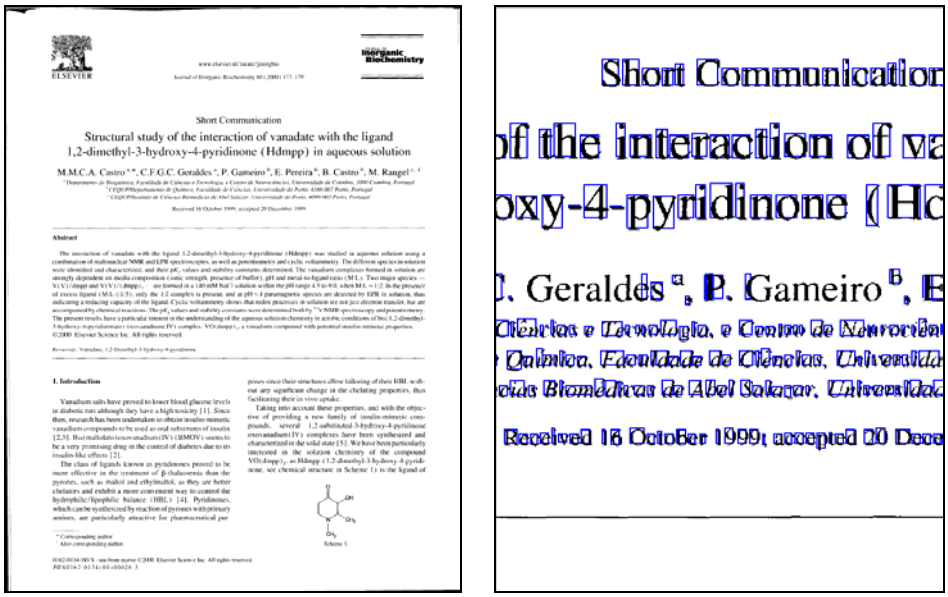


Figure 2. Structure of the AL module



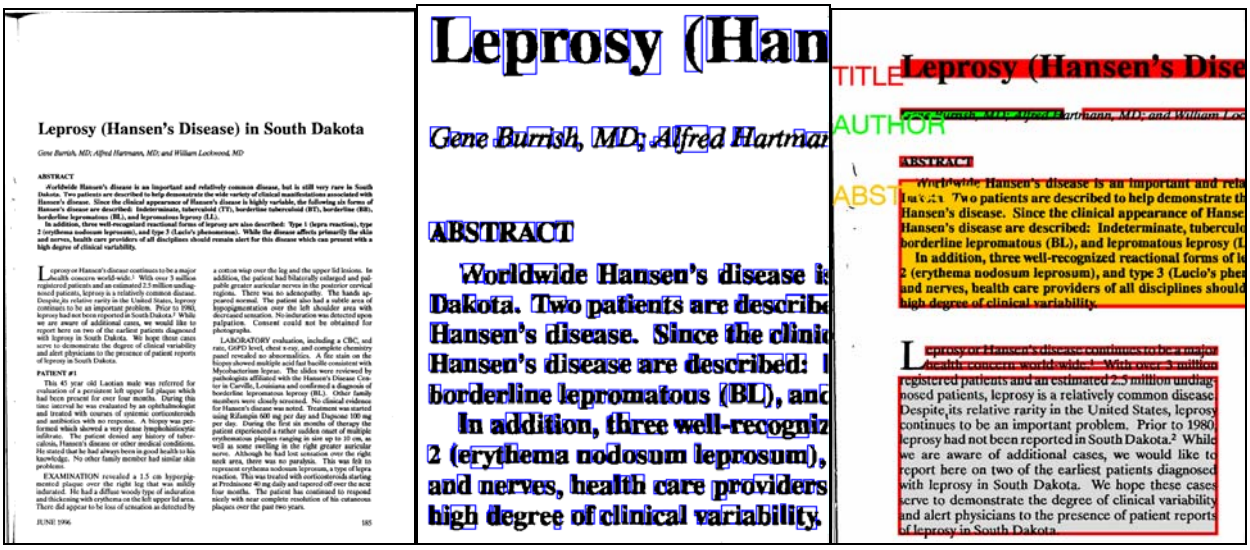
(a) (b) (c)

Figure 3. Example of the Automated Labeling algorithm. (a) Input image, (b) AZ result, and (c) AL result.



(a) (b)

Figure 4. Consequence of incorrect OCR. (a) Original article bitmap (b) OCR result (No bounding box in several author name characters)



(a) (b) (c)

Figure 5. Consequence of incorrect OCR and AZ. (a) Original article bitmap, (b) OCR result. (No bounding box in several author name characters), (c) AL result.

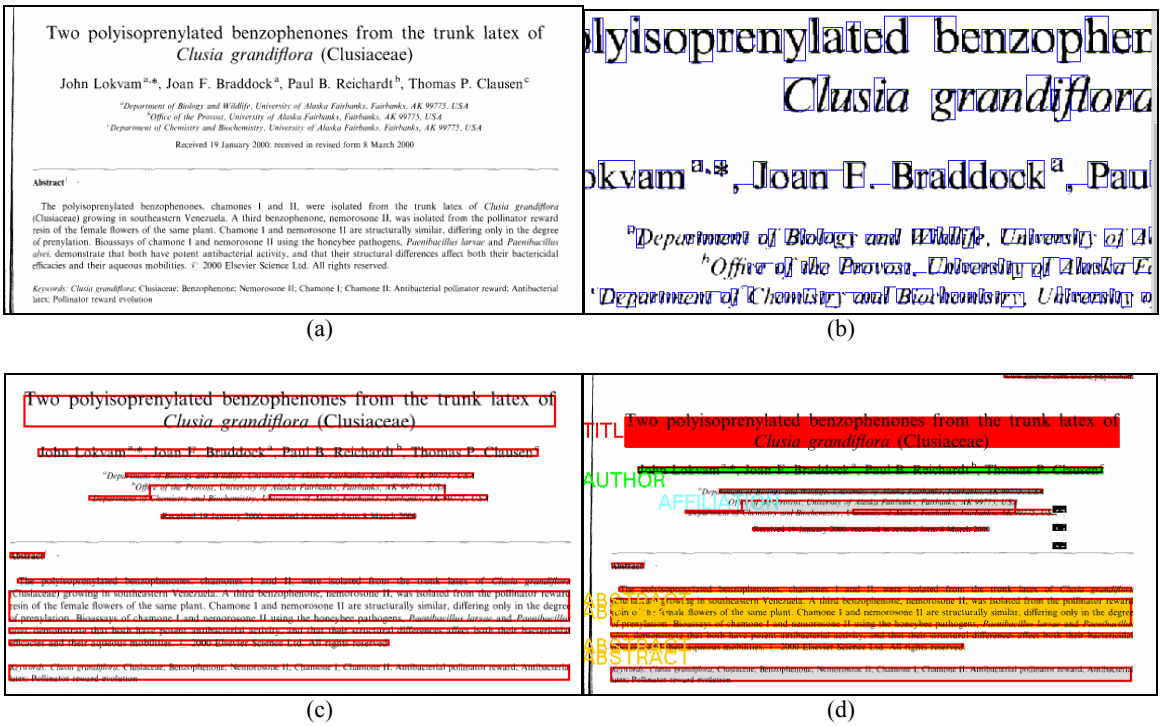


Figure 6. Example of incorrect OCR and AZ results. (a) Original article bitmap, (b) OCR result (No bounding boxes in several author name characters), (c) AZ result (d) AL result.

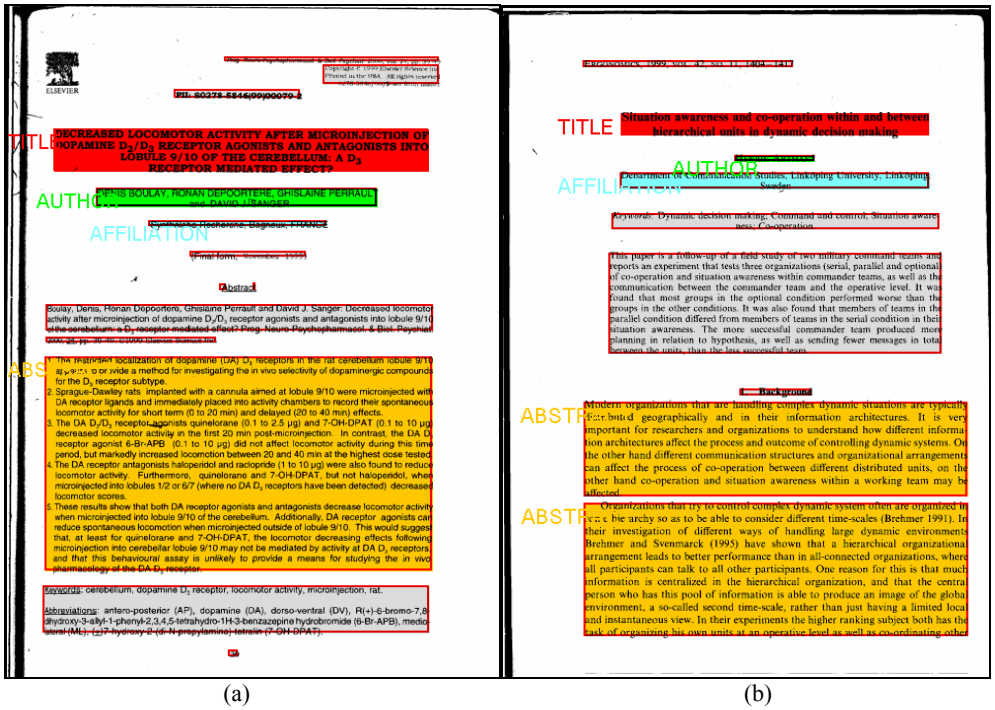


Figure 7. Example of AL algorithm errors. “Introduction” zones are labeled as abstract because of (a) large title zone and incorrect OCR result, and (b) the word “Background” located before introduction.