

Automated Labeling of Zones from Scanned Documents

Daniel X. Le, Jongwoo Kim, Glenn Pearson, and George R. Thoma

National Library of Medicine
Bethesda, Maryland 20894

Abstract

The Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine (NLM), is developing an automated system, the Medical Article Record System (MARS), to identify and convert bibliographic information from printed biomedical journals to electronic format for inclusion in the MEDLINE database. This paper describes one aspect of this ongoing effort: the automated labeling of zones from scanned images with labels such as titles, authors, affiliations, and abstracts. This labeling is based on features calculated from optical character recognition (OCR) output, neural network models, machine learning methods, and a set of rules that is derived from an analysis of the page layout for each journal and from generic typesetting knowledge for English text.

Several learning systems are considered including back-propagation neural networks, decision trees, and rule-based systems. Experiments are carried out on a variety of medical journals, and the performance of these techniques are analyzed and compared in terms of development times, training times, and classification accuracy.

1 Introduction and Background

Automated document conversion systems are being developed for a variety of document related applications to convert paper-based document information to electronic format. Paper documents usually consist of text zones, or a mixture of text and non-text zones, and each text zone has its own label such as titles, authors, affiliations, abstracts, etc. In order to support automated document searching, automated document delivery, and electronic publishing (converting papers from one format to another or modifying manuals and references, etc.), document labeling techniques are then required to extract the meanings of text zone contents.

Most document labeling techniques proposed so far in the literature [1, 2, 3, and 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 '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. Other techniques [5, 6] used the outputs of OCR to further improve labeling accuracy. In this paper, we propose an automated technique to label text zones with labels such as titles, authors, affiliations, and abstracts using integrated image and OCR processing, rule-based technology and back-propagation neural network. Preliminary evaluation results show that the system is capable of labeling text zones at a classification accuracy of 99.6% for the rule-based system and of 97.0% for the back-propagation neural network.

The rest of this paper is divided into six sections. Section 2 provides a system overview. Section 3 presents zone features. Sections 4, 5 and 6 describe in detail the labeling techniques and experimental results. Section 7 contains conclusions and future work.

2 System Overview

The automated labeling technique described here is one prototype component of our second-generation MARS system under development. The process consists of three steps as follows:

- Scan journal images.
- Perform optical character recognition (OCR). This includes detecting zones around paragraphs.
- Apply automated labeling. This associates a label, such as "Title", with each zone of interest.

Additionally, since verification and correction steps are needed to collect ground truth data for training, a Zone Checker system was also implemented to serve this purpose and the interested reader might refer to reference [7] for more information about this system.

Using a commercial 5-engine OCR system developed by Prime Recognition Inc. (PR) [8], scanned binary document images are first segmented into rectangular text zones. Each zone is then processed to deliver an OCR output (including zone coordinates, text line information, characters and their bounding boxes,

confidence levels, font sizes, and certain style attributes). From this output, features for each zone are calculated and input to several learning systems for label classification. Finally, it is planned that the label classification outcomes from these learning systems for each zone are combined by voting to reach the final decision on the zone's label.

The calculated features include geometric ones, such as the zone's height/width ratio, the zone area or its position in a page, as well as those based on character statistics or substring recognition against word lists. These features are extracted from the output of the PR OCR system that provides information at the page, zone, line, and character levels, as given below:

Zone Level

- Zone boundaries
- Number of text lines

Line Level

- Number of characters
- Baseline
- Average character height
- Average font size

Character Level

- Recognized 8-bit character
- Confidence level ($1 = \text{lowest}$, $9 = \text{highest}$)
- Bounding box
- Font size
- Font attributes (*normal*, *bold*, *underlined*, *italics*, *superscript*, *subscript*, and *fixed pitch*)

3 Zone Features for Document Labeling

Most features and rules derived for labeling techniques in this paper are based on an analysis of the page layout for each journal, and generic typesetting knowledge for English text [9]. Both geometric and non-geometric features are considered here.

The geometric layout features are calculated based on the zone location, zone order, and zone dimensions. Title zone is usually located in the top half of the first page of an article with the biggest font size. As reported in a title page study [10], roughly 96% of titles have the largest font compared to other zones in the top half of the first page. Normally, title is followed by author, affiliation, and other publisher information. The font size of the author zone is usually smaller than that of the title zone. The non-geometric features derive from the zone contents, and can involve aggregate statistics, font characteristics such as total characters, total capital letters, total punctuation marks, etc. Table 1 shows a list of features used both in the rule-based system and in the back-propagation neural network.

For the rules-based system, approximately 50 features are extracted from the PR OCR output. In addition to the features shown in column 2 of Table 1,

extensive word matching based on cue words is used as shown in Table 2. Word matching relies upon lists of cue words commonly associated with particular label types.

Word matching is very important for the system since a zone has a higher probability of being labeled as "Affiliation" zone when a zone has many country, city, and school names. Seven database tables with word lists have been assembled and the Ternary Search Tree algorithm [11] is used as a search engine for the word matching shown in Table 2.

4 Labeling using Rule-Based System

NLM's MEDLINE database contains bibliographic records from about 3800 journals. Their physical layouts can be categorized into several hundred types. Figures 1(a), 1(b), and 1(c) show some examples of types consisting of one column, a combination of one and two columns, and two columns, respectively. We define Figure 1(a) as Type 1, Figure 1(b) as Type 12, and Figure 1(c) as Type 2, respectively. It is very difficult to design a single automatic labeling (AL) module that can handle all types of journals. Therefore, we classify journals as belonging to specific types, and design an AL module for each particular type. Since Type 12 occurs most frequently in our journal collection, we will make an AL module for it first and will handle other types in the future.

For our purpose, we are interested in five zone labels in an article: title, author, affiliation in 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, called rules 1, 2, 3, and 4 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 proposed AL technique consists of four steps as shown in Figure 2 and described in the following paragraphs.

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 ($\text{PID} \geq 100$), the zone is labeled as title.

In the second step, previous labeling results are checked again by rule 4. For example, when two separate zones are both labeled as author but they are not close to each other, one zone is then removed from the author category.

In the third step, in addition to rule 2, rules 1 and 4 also 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 labeled as author does not have any information about author ($\text{Number_Middlename} = 0$ and $\text{Number_Degree} = 0$), geometric features are then used

to do the labeling. That is, if a zone does not have any information about title and upper affiliation and it is located between title and upper affiliation, the zone is labeled as author.

In the fourth step, the PR 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 final step. The detailed rules for each label type are shown in the following:

Let Max_Font_Size represent the biggest font size in a page and Height_Article be the difference between the bottom and top coordinates of the bottom-most and top-most zones, respectively.

4.1 Rules for Title

Rule 1

1. Sentence_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, Avg_Line_Height, and Avg_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 upper affiliation
 < Coordinate_Upper of abstract
 < Coordinate_Upper of lower affiliation

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. Sentence_Headtitle == Sentence_Abstract == 0
Sentence_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, Avg_Line_Height, and Avg_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. Sentence_Headtitle == Sentence_Abstract == 0
Sentence_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 Sentence_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, Avg_Line_Height, and Avg_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. Sentence_Headtitle == Sentence_Abstract == 0
Sentence_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 (Sentence_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, Avg_Line_Height, and Avg_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 Sentence_Abstract == 1)
 - PID = 100
 - If (Previous Zone has Sentence_Received == 1)
 - PID = 100
 - If (Next Zone has Sentence_Introduction == 1)
 - PID = 100
 - If (Next Zone has Sentence_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, Avg_Line_Height, Avg_Line_Length, and Avg_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 Labeling using Neural Network System

Before a neural network model can be used as a pattern classifier, its structure has to be designed and trained. We discuss in this section the selection of training and testing data sets, a method to train and test the neural network, and the neural network structure design.

5.1 Training and Testing Data Sets

Since each journal has its own page layout and style setting, our preliminary approach is to create a neural

network for each journal type. A neural network for each particular journal type is designed, trained, and tested with its own data. For each journal type, a group of at least four journal issues is selected to create the training and data sets. The training data set is used to design the neural network while the testing data set is used to estimate the classification accuracy.

Sixteen different journal types consisting of 66 issues were selected for the experiment for a total of 2176 binary images. These images are 8.5 x 11 inches and scanned at 300 dpi resolution.

5.2 Cross-Validation Method

For purposes of generalization, the cross-validation (CV) technique [12] is used by randomly dividing the training data set into five data groups of which four data groups create a *CV-train set* and one remaining data group is considered as a *CV-test set*. As a result, there are five pairs of a CV-train set and a CV-test set that are used to train and test the back-propagation neural network. The modified weights corresponding to the winning pair of a CV-train set and a CV-test set, the one yielding the highest classification accuracy, are chosen to be the final weights for the neural network.

5.3 Back-Propagation Neural Network

Back-propagation (BP) [12, 13, and 14] is a multi-layer neural network using sigmoidal activation functions. The network is made up of an input layer, hidden layers, and an output layer and nodes in each layer are fully connected to those in the layers above and below. Each connection is associated with a synaptic weight. The BP network is trained by *supervised learning*, using a gradient descent method, which is based on the least squared error between the desired and the actual response of the network.

In this project, a two-layer BP network is implemented with an input layer – thirty-eight text zone features shown in Table 1, a five output layer (title, author, affiliation, abstract, and others), and one single hidden layer of which the number of nodes is 16. Therefore, the two-layer BP network architecture is 38-16-5. Each input vector of the training data set is presented to the network many times and the weights are adjusted on each presentation to improve the network's performance until the network stops improving. Two learning factors that significantly affect convergence speed, as well as accomplish avoiding local minima, are the learning rate and the momentum. The learning rate determines the portion of weight needed to be adjusted. Even though a small learning rate guarantees a true gradient descent [14], it slows down the network convergence process. The momentum determines the fraction of the previous

weight adjustment that will be added to the current weight adjustment. It accelerates the network convergence process. During the training process, the learning rate was adjusted to bring the network out of either its local minima (where the network has converged but its output error is still large) or its no-gain mode (the network mode in which its output error does not change or changes very little over many cycles). The learning rate ranges from 0.001 to 0.1, and the momentum is 0.6.

6. Experimental Results

6.1 The Rule-Based System

There were 90 rules generated for the Type 12 and 38 journals consisting of 1407 articles were selected for experiment. Experimental results showed that 1402 articles were labeled correctly with 99.6% of correct recognition rate. We had five errors in labeling affiliation zones due to the incorrect font attributes and poor contents obtained from the output of the PR OCR system.

6.2 The Neural Network System

The BP neural network was trained with all five pairs of a CV-train set and a CV-test set. The average training time spent for the each pair was about 4 hours. The BP neural network configuration associated with the winning pair was evaluated on the testing data set. The result showed that the average classification accuracy on the testing data set was about 97.0 %. Most errors were due to the segmentation problem generated from the PR OCR output that split zone of interest (such as title zone) into multiple zones, as well as merged several different zones (such as author and affiliation zones) into a single zone.

7 Summary and Conclusion

Two automated labeling techniques, a rule-based system and a back-propagation neural network, have been presented in this paper. Both techniques yielded very good performance and showed the possibility of extension to other journals as prototypes.

The rule-based labeling technique uses 90 rules for the journal layouts designated “Type 12” journals. More rules are expected to be added to handle other types of journals. This labeling technique employed both geometric and non-geometric zone features as well as geometric relations between zones as the basis for the proposed set of rules. Other rules can be changed or added easily since there is no training procedure. However, label classification time is proportional to the number of rules. In addition, much time and effort are

needed to devise rules that can be used for more than one type of journal.

In the case of the back-propagation neural network technique, label classification time is very fast and the results are stable regardless of the journal types. However, it is hard to use the geometric relations between labels as features and it is time consuming to train the module and tune its learning parameters. It is also hard to analyze wrong labeling results. The most serious drawback is that the whole neural network must be trained again when we have new types of journals to label.

Since each system has advantages and disadvantages, our future approach is to combine these systems together with a voting procedure to improve the labeling results. Another AL module using a decision tree algorithm shall be implemented and the results of these three AL modules will be voted on to improve the accuracy of label classification.

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] 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
- [6] 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.
- [7] G. Pearson and G. Thoma, Manual Verification and Correction of Automatically Labeled Zones: User Interface Considerations, *Proc. SDUIT '99*, Annapolis, MD, April 1999.
- [8] Prime Recognition Inc., Prime OCR Access Kit Guide, version 2.70, San Carlos, CA, 1997.
- [9] G. Nagy, At the Frontiers of OCR, *Proc. IEEE*, Vol. 80, No. 7, pp. 1093-1100, 1992.
- [10] 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.
- [11] J. Bentley and B. Sedgewick, Ternary Search Trees, *Dr. Dobb's Journal*, pp. 20-25, April 1998.
- [12] D. R. Hush and B. G. Horne, Progress in Supervised Neural Networks - What's New Since Lippmann? *IEEE Signal Processing Magazine*: pp. 8-39, 1993.
- [13] R. P. Lippmann, An Introduction to Computing with Neural Nets, *IEEE Acoustics, Speech and Signal Processing Magazine* 4(2), pp. 4-22, 1987.
- [14] J. M. Zurada, Introduction to Artificial Neural Systems, *West Publishing Company*, St. Paul, Minnesota, 1992.

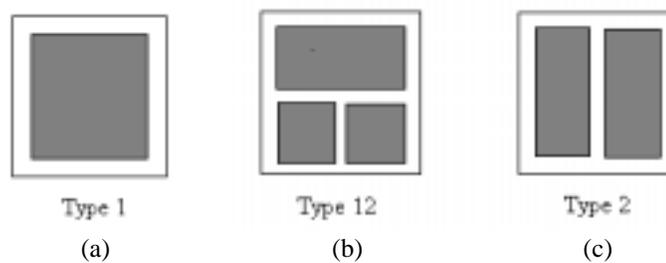


Figure 1. Examples of journal types.

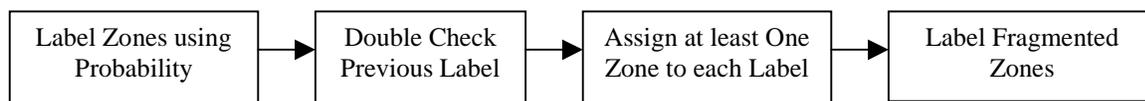


Figure 2. The procedure for automatic labeling.

Table 1. Features Associated with Each Zone

The rule-based system variables in this table are unnormalized while the neural network-based system variables are normalized either to page dimensions (NTPD) or to zone contents (NTZC).

Zone Features	Rule-based System	Neural Network System
<i>Geometric Features:</i>		
Zone coordinates, pixels	Coordinate_Left, _Right, _Upper, _Lower	NTPD
Zone height and width, pixels	Height_Zone, Length_Zone	NTPD
Zone centroid (X and Y)		NTPD
Zone shape: 100 log(height/width)		NTPD
Zone area		NTPD
Median height and length of lines*	Avg_Line_Height, _Length	
Median vertical spacing between lines*	Avg_Line_Space	
Zone order, in sequence by top edge	R****	
<i>Non-Geometric Features:</i>		
Lines	Number_Line	R
Total characters	Number_Character	R
7-bit capital characters [A-Z]	Number_Capitalcharacter	NTZC
7-bit lower case characters		NTZC
Numerals		NTZC
Punctuation group !"#\$%&'()*+,-.:;?@[\\]^_`{ }~		NTZC
Math symbol group (w/o minus) + / < = >		NTZC
3 groups of symbol pairs: [] () { }	(special cases: parentheses)	NTZC
Commas	Number_Comma	NTZC
Other separately-totaled punctuation characters { . - ; : ' " * }	(special cases: . - ;)**	NTZC
Other separately-totaled characters	(special cases, e.g., ©)	
Highest-confidence characters (= 9)		NTZC
Less-than-highest confidence character (< 9)		NTZC
Characters with particular font attributes (6 non-disjoint totals, & total of attribute-frees)		NTZC
Number of words	Number_Words	
Number of initials (e.g., "A.")	Number_Middlename***	
Average font size		R
Maximum font size	R	
Dominant {Font Attribute, Font Size} Pair	Font_Attribute, Font_Size	
Zone avg. font size:Page avg. font size		Ratio

* The median line height, length, and vertical spacing are derived by first calculating the median upper and lower character boundaries for each line.

** While the rule-based system doesn't total up individual punctuation marks throughout the zone, it does calculate totals within particular places, e.g., number of dashes at the ends of lines during word count generation.

*** includes "jr.", "sr.", "II", etc.

**** R = numerical raw counts.

Table 2. Some features used in the Rule-Based System.

Zone Features	Description
Number_Degree	Number of “M.D., Ph.D., M.S., R.N, ...”
Number_Affiliation	Number of “names of city, country, hospital, department, ...”
Percent_Capitalcharacter	Percentage of Number_Capitalcharacter per character
Percent_Degree	Percentage of Number_Degree per word
Percent_Affiliation	Percentage of Number_Affiliation per word
Percent_Middlename	Percentage of Number_Middlename per word
Sentence_Headtitle	Check the existence of a word such as “review, letter, note, ...”
Sentence_Abstract	Check the existence of a word such as “abstract, summary, ...”
Sentence_Subabstract	Check the existence of a word such as “aim, background, design, result, ...”
Sentence_Keyword	Check the existence of a word such as “keyword, index word, ...”
Sentence_Introduction	Check the existence of a word such as “introduction, ...”
Sentence_Received	Check the existence of a word such as “received, revised, ...”
Sentence_Affiliation	Check the existence of a word such as “correspondence, to whom, mailing, ...”