Part-of-Speech Tagging for Table of Contents Recognition

A. Belaïd[1], L. Pierron[2] and N. Valverde[3]
[1]*LORIA-CNRS,* [2]*LORIA-INRIA,* [3]*LORIA-ITESOFT*
Campus Scientifique B.P. 239
54506 Vandoeuvre-Lès-Nancy France
{abelaid,pierron,valverde}@loria.fr

Abstract

A labeling approach to automatic recognition of tables of contents (TOC)s is described. A prototype is used for consulting electronically scientific papers in a digital library system named Calliope. This method operates on an a roughly structured ASCII file, produced with OCR.. Labeling is based on a part of speech (POS) tagging. Tagging is initiated by a primary labeling of text component using some specific dictionaries. Significant tags are then grouped in title and author strings and reduced in canonical forms according to contextual rules. Non labeled tokens are integrated in one or another field per either applying contextual correction rules or using a structure model generated from well detected articles. The designed prototype operates with a great satisfaction on different TOC layouts and character recognition qualities. Without manual intervention, 95.41% rate of correct segmentation was obtained on 38 journals including 2703 articles *and 81.74% rate of correct field extraction.*

1. Introduction

Document analysis is often based on the detection of some regularities on a document structure and content. There are two methodologies for regularity detection: 1) by using an *a priori* model summarizing different regularity schemes. This imposes to assure a certain generality in the model writing in order to cover the maximum of possible cases; 2) by discovering progressively the regularity from some indices directly extracted from the content, related to linguistic, typographical or contextual aspects. This avoids the use of an a priori model and leads to discover an adapted model for each circumstance.

POS tagging [1] takes a key place for this modeling approach essentially when document is produced by OCR, containing some corrupted characters and structure defaults. The interest of POS tagging is to restore the content syntax and also the document structure. It can help to identify keywords and to associate them to specific fields.

The application related in this paper concerns the recognition of TOCs by identifying their articles. Articles have a relatively simple structure composed of three essential fields: *title*, *authors* and *page number*. In spite of the simplicity of this structure, two factors may handicap the straightforward extraction of fields: order unstable and bad separation. These two factors may be accentuated by OCR errors such as the suppression of the separation between two fields and the introduction of errors in field words which can corrupt their identification. The POS tagging may help marking out a field from the identification of some keywords as proper nouns for author fields and nominal groups for the title.

A few TOC recognizers appear in the literature. On one hand, Takasu and al. [4,5] propose a system named CyberMagazine, based on image segmentation into blocks and syntactic analysis of their contents. The article recognition combines the use of decision tree classification and a syntactic analysis using a matrix grammar. On the other hand, Story and O'Gorman [2,3] propose a method combining OCR techniques and image processing. Blocks are first located by the image processing ``docstrum''. Then, the TOC layout and relationships between the different article references are found according to an *a priori* model given manually for each kind of journal. These relationships are used, for example, to determine automatically the page number when the user clicks on one article title, or to give specific information concerning one article.

Contrary to these systems, ours works directly on a text file produced by OCR without any text preprocessing. The article structure is discovered progressively by a recursive and adapted labeling of the text components and its extension to the nearest context according to the article reference properties. Then the model generated from some reliable articles is used to complete the structure of the unachieved articles.

2. TOC Analysis

The TOC analysis follows three major steps. First, a primary labeling of lines and text components is performed. Second, based on these labels more syntactic forms are constructed to represent each article reference. At last, the final structure of the TOC articles is built and used to improve the structure of the bad structured articles.

2.1. Primary Labeling

The text file is examined line by line and each line, space or token within a line is labeled according to labels given in Table 1. A tabulation is a regular long space, occurring at the same position in several lines. A dotted line is a consecutive list of dots. Common nouns correspond to words belonging to a common dictionary, revealing the possible presence of a title in the area examined. Proper nouns are extracted from an author dictionary indicating the possible presence of authors. Initials correspond to first name abbreviations (i.e. capital letter followed by a dot) , revealing the possible presence of an author. The connectors "and" and "by" are important indices revealing the possible presence of authors in an area containing initials and proper nouns. NL is reserved for unknown tokens because they are not found in the dictionaries or because they are bad recognized by OCR.

SL	Space Line
SP	Long Space
TB	Tabulation
DL	Dotted Line
CN	Common Noun
PN	Proper Noun
IT	Initial for first name
NS	Numerical String
PT	Punctuation
PU	PT starting an article
CR	Connector like AND
NL	Not Labeled

Table 1 : Token Labels

2.2. Article Location

A TOC is rarely written in a distinctive manner in the page. Textual zones accompanies the TOC as headers in the top, footnotes in the bottom, or sometimes as editorial zones at different places. So, in order to make the method location independent, we have based the search for the TOC location, on the detection of page numbers.

2.2.1. Page Number Extraction. In our method, only references accompanied by their page numbers are considered. So, as the numerical strings are easy to extract, the first step for the reference location is the NS location. Knowing that these NSs are regular, all the NSs are first extracted and then only those presenting some location regularities are considered. The regularities correspond to 1) position at the beginning, at the end or within the reference, 2) vertical alignment, 3) after a specific punctuation like dotted line, 4) after or before a space line. A weight between 0 and 100 is assigned to each regularity. Then, for each NS, a total weight is performed by adding regularity weights. Those presenting a high amount greater than a given threshold are retained.

2.1.3. Article Delimitation. Some logical rules are used to mark up the obvious beginning or ending lines of TOC references. Let CL, PL, NL, BL, EL be respectively the current line, the previous line, the next line, the beginning line and the ending line. Rules used for reference location are :

CL = BL iff:
- ✓ CL begins by PU or by NS
- ✓ CL contains NS and PL=SL
- ✓ CL contains NS and PL = EL

CL = EL iff:
- ✓ CL contains NS and NL=SL
- ✓ CL contains NS and NL = BL

Then, knowing the beginning or the end of an article, the systems tries to delimit it by deducing the missing label. This is done either at this level if the context is rich enough or later by using more information from labeling and article modeling.

3. Authors and Title Extraction

Reduction rules are applied recursively on the initial tags, grouping progressively in different steps the authors (Aut) and title (Tit) components.

3.1. Gathering Rules

The first step deals with author and title field initialization by gathering some obvious consecutive labels. Table 2 outlines these gathering rules. "+" means a succession.

AUTHOR (Aut)	TITLE (Tit)
IT + PN ⇒ Aut	CN + CN ⇒ Tit
PN + IT ⇒ Aut	CN + CR ⇒ Tit
IT + IT + PN ⇒ Aut	
PN + IT + IT ⇒ Aut	
PN + PN ⇒ Aut	
PN + PN + IT ⇒ Aut	
PN + PN + PN ⇒ Aut	

Table 2 : Gathering Rules.

3.2. Reduction Rules

In the second step, the sub-fields are assembled by grouping either similar elements or by assimilating embedded terms representing punctuation or connectors. Table 3 shows some of these different rules.

AUTHOR	TITLE
IT + Aut ⇒ Aut	Tit + Tit ⇒ Tit
Aut + IT ⇒ Aut	CN + Tit ⇒ Tit
Aut + CR + Aut ⇒ Aut	Tit + CN ⇒ Tit
BY + Aut ⇒ Aut	Tit + CR + Tit ⇒ Tit
	Tit + CR + Tit ⇒ Tit

Table 3 : Grouping Rules

3.3 Contextual Cleaning Rules

At the issue of the sub-field grouping some typographic elements remain non classified. This is always the case for the punctuation, spaces or tabulations situated at the end lines (EL) or at the beginning of lines when an article is written on several lines. With the help of the context, it is possible to know if these indices can be considered as field delimiters or simply if they belong to one of the surrounding fields. Table 4 gives some of these cleaning rules.

AUTHOR	TITLE
Aut + SP/PT+Aut ⇒ Aut	Tit+SP/PT+Tit ⇒ Tit
Aut + EL+TB+Aut ⇒ Aut	Tit+EL+TB+Tit ⇒ Tit

Tableau 4 : Contextual Cleaning Rules

3.4. NL Contextual Assimilation

Due to current OCR errors and to the existence of unknown proper nouns in the author fields, a great number of tokens remain non labeled. The use of part of speech tagging will allow us to find contextual situations where it is possible to rectify these bad marks. This is made in three different phases.

3.4.1. Contextual Correction. Some contextual situations can promote the identification of missing fields. Table 5 lists some of that situations. In fact, it is easy to observe that for author, for example, the presence of the first or last name accompanied with an NL before an ending mark or before a title, can be interpreted as an author. Similar situations can be examined for the title.

AUTHOR
IT/PN+NL+EL+ TB ⇒ Aut+EL+TB
IT/PN + NL + DL ⇒ Aut + DL
IT/PN + NL + IT EL+ Tit ⇒ Aut + EL + Tit
Tit + EL+ TB + NL/CN/PN/ ⇒ Tit+EL+TB+Aut
NL/CN/PN/IT + EL + TB + Aut ⇒ Tit+EL+TB+Aut
TITLE
CN+NL + EL+ TB ⇒ Tit
CN+NL + DL ⇒ Tit + DL
CN+NL + EL+ Aut ⇒ Tit + EL + Aut
Aut+EL+TB + NL/CN/PN/IT ⇒ Aut+EL+TB+Tit
NL/CN/PN/IT + EL+TB+Tit ⇒ Aut+EL+TB+Tit

Table 5: Contextual Assimilation Rules

3.4.2. POS Tagging. A contextual assimilation is tried when the NL tags are surrounded by two similar sub-fields. The POS examines the favorable cases for this assimilation by studying the "meaning" of tokens placed at the end and at the beginning of respectively the left and the right sub-fields. For example, for authors, the intermediate non labeled term can be a connector if the two surroundings authors are complete, i.e. composed each one by two first names and one last name. It can be an initial if the author at the right contains only one first name, etc. For the title, The POS examines the grammatical categories of the surrounding tokens. For example, if the left token is an article or a subject, there is a chance to assimilate the non labeled term as a verb or a noun. Table 6 summarizes these different contextual situations.

AUTHOR	TITLE
Aut+NL+Aut ⇒ Aut	Tit+NL+Tit ⇒ Tit
Aut+NL+EL+Aut ⇒ Aut	Tit+NL+EL+Tit ⇒ Tit
Aut+EL+NL+Aut ⇒ Aut	Tit+EL+NL+Tit ⇒ Tit
Aut+NL+EL+NL+Aut ⇒ Aut	Tit+NL+EL+NL+Tit ⇒ Tit

Table 6 : POS Rules.

3.4.3. Reference Model Generation. Once the labeling reduction phase is finished, we try to determine the most repeated form articles which we regard as model to go to rectify the non well labeled articles. Thus, for each article, we determine some structure indices related to : the number of columns, the position of the page numbers, the apparition order of the articles references, etc. Then, we determine for all the articles extracted the most regular structure.

Two cases can occur:

- Authors and title are located in two different columns. It is the easiest case because one can decide to assign directly the non labeled tokens to more pertaining field in each column. Let x and y be tokens representing respectively a title and an author in article lines. Let $\delta_i = x_j$ TAB y_k the lines representing an article (assuming that the model

have given title and author in this order). Than, the tokens are modified as follows:

$\forall i$ If $x_i \cong Tit \Rightarrow x_i = Tit$, $\forall j$ If $y_j \cong Aut \Rightarrow y_j = Aut$

- Authors and title are located in the same column, separated or not by a tabulation or a punctuation. It is the most frequent case (> 80%) for periodic. In this case we consider the article as a sequence: 1) $\{ = \{x_i\}$ TAB|PN$\{y_j\}$ or :2) $\{ = \{x_i\}$. If the model proposes a title in the beginning, then:

  - in 1) : $\forall i$ If $x_i \cong Tit \Rightarrow x_i = Tit$, $\forall j$ If $y_j \cong Aut \Rightarrow y_j = Aut$
  - in 2) : $\forall i<k$ If $x_i \cong Tit \Rightarrow x_i = Tit$, ,$\forall j>k$ If $y_j \cong Aut \Rightarrow y_j = Aut$

  where k is the indice of the last title in $\{$.

Table 7 gives a complete labeling example. In the final labeling, articles are surrounded by "<" and ">" and ignored lines are preceded by "-".

4. Experiments and results

We tested this prototype on 38 reviews, including 2703 articles and 1486 fields author. The rate of localization of articles is 95.43% and the rate of recognition of fields is 95.43% for the numbers of page and 81.74% for the separation of the titles and the authors. For the delimitation of articles, the difficulties come, on one hand, from the suppression by the OCR of the blank lines due to the presence bordering on particular fonts, and on the other hand from the proximity between the titles of headings and articles. This introduces ambiguities on the article limits (starting and ending points), due to the merger between the heading and the article line. The consequence might that we don't get enough of well recognized articles to validate a meaningful model. Another concern is the ambiguity within the identification of all the items making of an article line. This is due to either a weak field separation or not significant token identification. The identification of fields works well if separation is honest between authors and titles of articles: fields in different columns, authors on a line with share.

5. Conclusion and prospects

We presented in this paper a system of TOC analysis of a homogeneous type (textual). The TOCs are digitized and converted into text with OCR. The recognition method uses a syntactical approach. It is based on linguistic labeling of words and syntactic reduction. The method works by field sweep line by line, by separating the articles between them initially, then article by article by separating the fields inside articles. The system operates without any a priori model. It adapts the extraction process on each new TOC by only taking into account some general and logical knowledge on the article structure. In this version, only one OCR (TextBridge) was used without any parameterization. In the future, we will extend this prototype by 1) combining many OCRs in order to improve the data quality, 2) reinforcing the POS rules for improving the incorporation of more linguistic rules, 3) enlarging the use of more complicated TOCs within magazines having complex layout.

6. References

[1] E. Brill, A simple Rule-Bases Part of Speech Tagger, In Proceedings of ANLP, 1992

[2] L. O'GORMAN, " Image and Document Processing Techniques for the Right Pages Electronic Library System ", ICPR, Vol. 2, pp. 260-263, 1992.

[3] G. A. STORY, L. O'GORMAN, D. FOX, L. LEVY SCHAPER, H. V. JAGADISH, " The Right Pages Image-Based Electronic Library for Alerting and Browsing ", Computer, September 1992.

[4] A. TAKASU, S. SATOH, E. ATSURA, " A Document Understanding Method for Database Construction of an Electronic Library ", ICPR, pp. 463-466, 1994.

[5] A. TAKASU, S. SATOH, E. KATSURA, " A Rule Learning Method for Academic Document Image Processing ", ICDAR'95, Vol. I, pp. 239-242.

a) TOC Text			b) Primary Labeling	b) Final labeling
Contents			CN	-
			SL	-
			SL	-
Articles			CN	-
			SL	-
Spatial Representation for Navigation in Animats	Tony J. Prescon	85	CN CN CN TB PN IT PN TB NS	<Tit TB Aut TB NS
			CN CN NL	>Tit
			SL	-
Monitoring Strategies for Embedded Agents: Experiments and Analysis	Marc S. Ackin and Paul R. Cohen	125	CN CN CN TB PN IT NL CR NL IT PN TB NS	<Tit TB Aut TB NS
			CN CN PT CN	= Tit
			CR CN	> Tit
			SL	-
Discovring the competitors	Luc Steels	173	CN CN CN TB PN PN TB NS	#Tit TB Aut TB NS
			SL	-
Rovlows			NL	-
			SL	-
Biological Adaptations and Evolutionary Epistemology	James H. Fetzer	201	CN CN CN TB PN IT NL TB NS	< Tit TB Aut TB NS
			CN CN	> Tit

Table 7 : Example of TOC Labeling