Une approche probabiliste pour la reconnaissance des sommaires

S. Souafi-Bensafi[1,2], H. Emptoz[1], F. Lebourgeois[1], M. Parizeau[2]

Laboratoire d'InfoRmatique en Images et Systèmes d'information, INSA de Lyon, France.

Laboratoire de Vision et Systèmes Numériques, Université Laval, Québec, Canada.

Résumé

L'analyse et la reconnaissance des documents écrits consistent à traduire leurs images numérisées sous une forme électronique réutilisable. L'analyse permet d'extraire à partir de l'image d'un document une structure dite physique, tandis que la reconnaissance associe aux composants de la structure physique leurs fonctions logiques dans le document. Le travail présenté dans cet article porte sur la phase de reconnaissance de documents dont la structuration logique est caractérisée par des marquages typographiques tels que les sommaires ou les tables des matières. Nous proposons une approche perceptuelle qui se base sur l'extraction de ces marquages typographiques directement à partir des images des documents. Ces documents présentent cependant une structuration variable et complexe. La complexité pose des difficultés au niveau de la phase d'analyse et peut conduire à des erreurs dans les données présentées à la phase de reconnaissance. Quant à la variabilité, elle impose d'entreprendre une modélisation générique de la structure logique et du processus de reconnaissance associé. Notre objectif est d'aborder ce problème de reconnaissance en présence de ces difficultés. Nous avons développé un système de reconnaissance automatique basé sur un modèle hybride combinant un classifieur bayésien et un automate probabiliste. Le rôle du classifieur est la correspondance entre les blocs de texte extraits dans les images des documents et les entités logiques à un niveau de structuration de base, alors que l'automate permet de regrouper ces entités logiques sur plusieurs niveaux hiérarchiques reconstruisant ainsi toute la structure logique. Ce modèle hybride est construit par apprentissage semi-supervisé, en s'appuyant d'une part sur la connaissance fournie de manière interactive par l'utilisateur, et d'autre part sur les propriétés typographiques des documents considérés. Nous avons expérimenté le système proposé pour l'indexation de sommaires de revues. La complexité et la variabilité de la structuration de ces documents nous ont permis de montrer l'efficacité de l'approche développée.

Mots clés : Analyse de documents, reconnaissance de documents, structure physique, structure logique, classifieur bayésien, automate probabiliste, typographie, apprentissage supervisé.

A probabilistic Approach for Table of Cotents Recognition

Abstract

Document Analysis and Recognition consist in translating their images into an electronic form that can be reusable. The analysis extracts the document layout structure from its image, and the recognition assigns to the layout structure components their logical functions in the document. In this article, we present our work on recognition of a category of documents in which the logical structure is based on typographical tagging such as table of contents. We propose a perceptual approach that extracts these typographical tagging directly from document images. However, the structures of such documents are complex and variable. Their complexity can cause errors in the analysis output, which influence directly the recognition task, while their variability requires defining a generic form of logical structures and the related recognition tasks. Our goal is to consider the document structure recognition problem even though these difficulties occur. We developed a automatic recognition system based on a hybrid model combining a bayesian classifier and a probabilistic automaton. The classifier is responsible of drawing a correspondence between text blocks extracted from document images and basic logical entities, while the automaton deals with grouping these entities into a hierarchical logical structure. This hybrid model is built by semi-supervised learning based on knowledge provided by the user on the one hand, and the typographical properties of our documents, on the other hand. This system has been experimented for automatic indexing of tables of contents in periodicals and journals. The complexity and the variability of these documents allow us to show the efficiency of the approach.

Keywords : Document analysis, document recognition, layout structure, logical structure, recognition, bayesian classifier, probabilistic automaton, typography, supervised learning.

1 Introduction

Avec l'expansion de l'Internet et des bibliothèques numériques, de plus en plus d'ouvrages comportent une édition électronique et sont consultables en ligne [10]. Mais, le challenge est de faire en sorte que les documents existant sous une forme non électronique ne soient pas exclus de la circulation. À présent, la numérisation est la solution adoptée, mais elle ne fournit que des images de documents, ce qui n'est pas toujours suffisant. En effet, il est souvent nécessaire d'accéder aux contenus des documents numérisés et de les modifier éventuellement. C'est l'objet de l'Analyse et la Reconnaissance des Documents (ARD) [17, 2, 8, 11]. Il s'agit de transformer des images de documents en une forme électronique symbolique adéquate pour le traitement par ordinateur. Dans certains cas, il suffit de convertir le contenu d'un document en mode texte. La localisation des zones textuelles dans l'image du document n'est cependant pas une tâche facile à réaliser, sa complexité étant fonction du type du document numérisé.

Outre la localisation et la reconnaissance du texte dans les images de documents, il est parfois nécessaire d'extraire des éléments de structuration. Le niveau de structuration à atteindre dépend, pour l'essentiel, de l'application considérée mais il reste conditionné par la nature du document qui regroupe le fond et la forme. L'information structurelle peut être de nature linguistique et/ou visuelle, les propriétés visuelles étant exprimée à l'aide de la typographie et de la mise en page. Notre travail se situe dans ce cadre de traitement automatique des documents visant à reconnaître la structure d'un document en adoptant une approche perceptuelle qui se base sur des propriétés de type visuel qu'il faut extraire à partir de son image. Nous nous intéressons plus particulièrement aux sommaires des revues qui appartiennent à une famille de documents que l'on qualifie de "documents à typographie riche et récurrente" [**?**, 16]. Ils sont caractérisés par la présence d'une typographie permettant de distinguer les éléments de structuration de leur contenu, en affichant plus ou moins de régularité basée sur la répétitivité. Le but est d'extraire la structure physique de ces documents et d'affecter à chacune de ses composantes sa fonction logique et de reconstituer ainsi la structure logique correspondante. Ces deux étapes de traitement correspondent à l'ARD.

L'ARD est confrontée à de nombreuses difficultés aux niveaux physique et logique. L'ensemble des méthodes et des techniques conçues à ce propos traitent souvent des documents ayant une mise

en page relativement simple et une structuration assez rigoureuse avec une tendance à développer des systèmes spécifiques. Diverses approches ont été adoptées basées principalement sur des règles de mises en formes [19, 9, 22, 5, 18, 4, 21, 20, 1]. Les sommaires de revues en particulier, et les documents à typographie riche et récurrente en général, constituent une famille étendue et trop diversifiée pour pouvoir rentrer dans une catégorie de documents déjà traités. De plus, ils regroupent un certain nombre de problèmes qui ne sont pas abordés dans l'ensemble et qui les font parfois classer comme cas difficiles à traiter, aussi bien au niveau physique qu'au niveau logique.

D'une manière générale, la modélisation et la reconnaissance des structures des documents restent conditionnées par les limites de la phase d'analyse, notamment dans les documents dits composites, qui présentent une mise en page complexe à cause de l'hétérogénéité de l'information qu'ils contiennent (couleur, texture, images, graphiques, variation de typographie, etc.). De ce fait, il est difficile de réaliser une phase d'analyse sans erreurs et de fournir à la phase de reconnaissance des conditions initiales parfaites. Les travaux réalisés par Belaïd sur les catalogues des bibliothèques [**?**] et les tables des matières dans les revues scientifiques [3] montrent des aspects de ces difficultés. Les catalogues ont été traités en se basant sur des propriétés purement visuelles tels des séparateurs et des symboles de ponctuations. Le modèle de la structure logique est fourni manuellement au système. Pour les tables des matières une approche linguistique a été adoptée se basant sur le contenu textuel extrait par OCR et l'extraction du modèle de la structure logique est réalisée automatiquement.

Pour les sommaires de revues, nous partons de l'hypothèse qu'aucune contrainte n'est posée sur leur mise en page. Les difficultés qui peuvent se présenter au niveau physique risquent d'engendrer des irrégularités dans l'information physique soumise au processus de reconnaissance. C'est la raison pour laquelle nous avons décidé d'aborder le problème de reconnaissance de la structure logique en adoptant une approche de modélisation hybride combinant un aspect probabiliste pour s'adapter à ces irrégularités, et un aspect structurel pour la modélisation de la structure logique.

La structuration des sommaires de revues étant principalement basée sur des propriétés visuelles, nous avons adopté une approche perceptuelle qui consiste à extraire de telles propriétés directement à partir des images de documents. Et pour couvrir toute catégorie de documents différents dans leur organisation spatiale et typographique, nous avons choisi de concevoir un système générique

de reconnaissance qui soit interactif en se basant sur un apprentissage supervisé. Nous proposons un modèle hybride combinant un classifieur bayésien, un automate probabiliste et une structure hiérarchique arborescente, chaque partie du modèle devant opérer selon les besoins d'un certain niveau de structuration.

La section 2 est consacrée, d'une part, à la description de la structuration de sommaires de revues, et d'autre part, à la présentation de leur traitement pour l'extraction et la caractérisation de la structure physique nécessaire à la phase de reconnaissance. La modélisation du processus de reconnaissance est présentée à la section 3, en décrivant chaque modèle utilisé, son processus d'apprentissage ainsi que son utilisation en reconnaissance. L'expérimentation de notre système sur une base de documents est décrite à la section 4.

2 Les sommaires : des documents à typographie riche et récurrente

Les sommaires de revues sont des documents à typographie riche et récurrente, ceux-ci constituant une grande famille de documents ayant en commun le caractère visuel (typographique et spatial) des éléments de structuration de leur contenu textuel [7, 16]. La richesse typographique signifie la présence de différents attributs ou marquages typographiques visibles permettant de distinguer les catégories de texte à mettre en relief. L'aspect récurrent se traduit par la répétitivité des enchaînements des attributs typographiques permettant de faire ressortir des éléments de structuration logiques. Ces documents peuvent être associés à un mode de lecture dite guidée par la recherche selon la classification de Richaudeau [15]. Il s'agit de documents qui ne nécessitent pas forcément une lecture intégrale de leur contenu pour repérer l'information d'intérêt. Cette famille de documents couvre notamment les catalogues, les annuaires, les dictionnaires, les sommaires et tables des matières, etc. (des exemples de ces documents sont présentés à la figure 1).

2.1 Modélisation de la structure logique

Pour modéliser les structures logiques des documents à typographie riche et récurrente, nous nous sommes basés sur les propriétés qui les caractérisent de manière invariable, à savoir la richesse et la

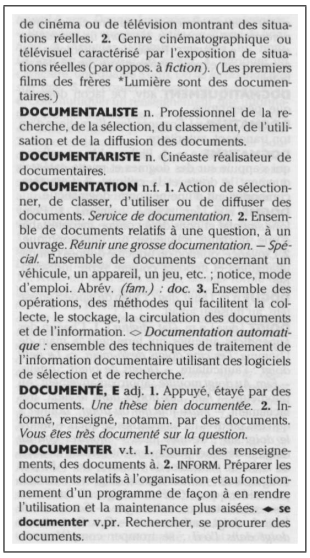


FIG. 1 – Exemples de documents à typographie riche et récurrente (sommaire de revue, catalogue de livres, dictionnaire).

récurrence typographiques. La richesse typographique est exprimée à l'aide de propriétés physiques des blocs de texte mises en relation avec les catégories de texte correspondantes constituant ainsi des *entités logiques élémentaires*. Quant à la récurrence, elle se présente sous forme de répétitions de séquences des propriétés typographiques, ces séquences correspondant à des *entités logiques composées*. Ainsi, on peut définir la structure logique comme étant une hiérarchie d'entités logiques organisées sur plusieurs niveaux et constituées par imbrication des groupes de blocs de texte qui leur sont associés. D'une manière formelle, la structure logique est caractérisée par une hiérarchie d'entités logiques organisées sur un nombre k de niveaux notés de 1 à $k-1$ pour les entités composées et le niveau 0 pour les entités logiques élémentaires (voir figure 2). Les entités composées d'un niveau i se composent d'entités des niveaux inférieurs (0 à $i-1$). Les entités élémentaires sont bâties par regroupement de *blocs élémentaires de texte* consécutifs, un bloc élémentaire étant la plus petite entité physique pouvant constituer une entité logique. Dans l'ensemble de notre travail nous considérons les mots comme blocs élémentaires de texte.

Dans le cas des sommaires de revues, la structuration logique comporte en règle générale trois niveaux définis autour de la notion d'*article*. Le niveau le plus bas contient les composantes d'un article qui sont les entités logiques élémentaires (*titre*, *numéro de page*, *auteur*, etc.). Le niveau suivant correspond à la catégorie d'article dans laquelle chaque entité logique est obtenue par regroupement

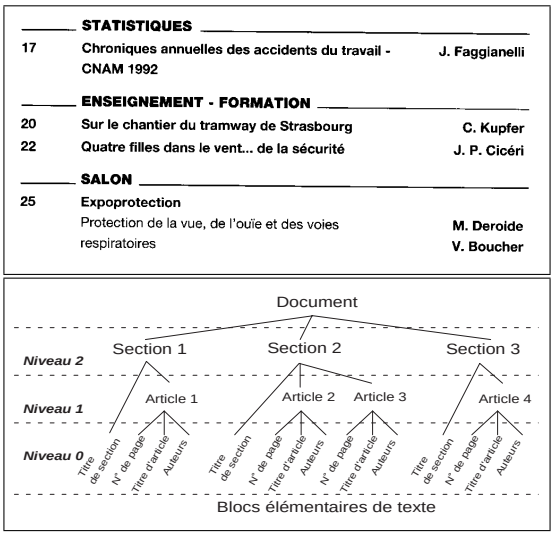


FIG. 2 – Zone d'un sommaire de revue et structure logique spécifique

d'entités du niveau inférieur. Le troisième niveau, s'il y a lieu, représente les regroupements d'articles autour d'une *rubrique* ou d'un *thème* constituant ainsi une *section* dont le début est éventuellement marquée par un *titre*. Nous montrons à la figure 2 un exemple d'une zone de sommaire et la structure logique correspondante. Celle-ci est organisée sur trois niveaux hiérarchiques comportant les entités logiques suivantes : *section* au niveau 2, *article* au niveau 1 et les entités élémentaires (niveau 0) *titre de section*, *titre d'article*, *numéro de page*, *nom d'auteur*.

La représentation présentée à la figure 2 constitue une *structure logique spécifique* à l'exemple de document considéré. Or, dans la famille des documents à typographie riche et récurrente, la structuration logique n'est pas régulière d'un document à l'autre. En effet, même à l'intérieur de la famille des sommaires de revues, la structuration logique reste très variable tant dans l'organisation des entités logiques (niveau logique) que dans la description spatiale et typographique (niveau physique). Nous avons par conséquent décidé de regrouper par sous-familles les documents ayant une structuration basée sur des entités logiques similaires et mise en relief par des éléments typographiques identiques. Dans le cas des sommaires, il s'agit de considérer chaque revue séparément et de lui faire correspondre une *structure logique modèle* construite par apprentissage automatique. Pour l'exemple présenté ci-dessus, la figure 3 montre la *structure logique modèle* associée à la revue à laquelle il appartient. Cette structure montre uniquement les entités logiques types et les liens qui les interconnectent.

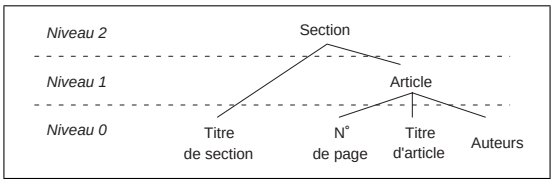


FIG. 3 – Structure logique hiérarchique modèle

2.2 Extraction de la structure physique

La reconnaissance de la structure logique repose sur la richesse et la récurrence typographiques, qui caractérisent les documents que nous considérons. Mais avant de l'aborder, il est nécessaire d'extraire la structure physique d'une part, et de décrire les blocs de texte au moyen de caractéristiques géométriques, typographiques et topologiques, d'autre part. Par ailleurs, pour segmenter nos documents, il faut envisager la présence d'éléments non textuels, l'utilisation de couleurs et de textures variées. Pour tenir compte de tels éléments, nous avons travaillé sur des images de documents en 256 niveaux de gris numérisées à une résolution de 600ppp. Nous avons adopté une segmentation permettant de localiser les lignes de texte directement dans les images en niveaux de gris et procédant à une binarisation locale à chaque ligne avec un seuillage adaptatif [12]. Cette technique de segmentation permet d'obtenir une bonne qualité au niveau des formes des caractères, ce qui est utile notamment pour l'extraction de leurs caractéristiques typographiques. A l'issue de cette étape, nous obtenons une structure physique composée d'une hiérarchie de blocs de texte des niveaux caractère, mot et ligne (figure 4).

Bien que nous ayons utilisé une méthode de segmentation robuste, dans certains cas, des erreurs demeurent inévitables. Nous avons décidé d'aborder le problème de la reconnaissance de la structure logique malgré la présence éventuelle de telles erreurs pour deux principales raisons : d'abord pour nous adapter aux conditions réelles, qui font que les erreurs sont inévitables, mais aussi pour observer les répercussions de ces erreurs et voir s'il est possible de les détecter et les identifier de manière automatique. En effet, les résultats du traitement au niveau physique ne pouvant être toujours parfaits, il serait intéressant de se servir de la connaissance acquise au niveau logique pour le valider ou le corriger. Dans ce qui suit, nous décrirons les caractéristiques physiques que nous avons choisi d'extraire dans nos documents. Elles sont organisées en deux catégories : typographiques et spatiales.

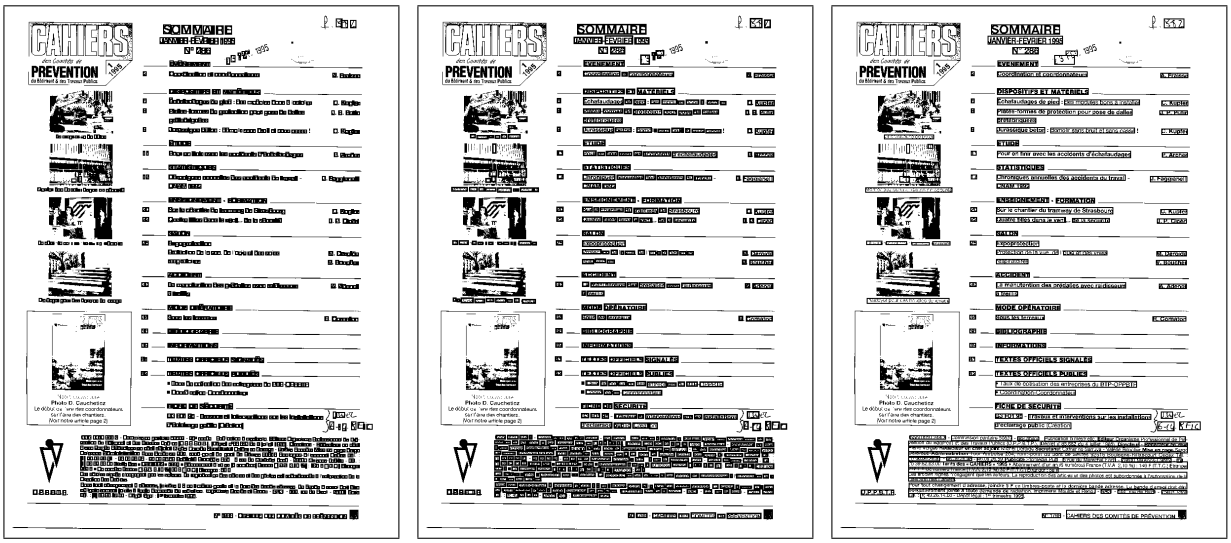


FIG. 4 – Extraction de la structure physique (blocs de texte à trois niveaux : caractères, mots et lignes).

2.2.1 caractéristiques typographiques

Les sommaires de revues comportent au minimum trois polices de caractères d'où l'utilité des propriétés typographiques des mots. Une méthode utilisée pour l'extraction de cette typographie sépare les mots écrits dans la même police sans avoir à reconnaître celle-ci [12]. Cette méthode procède dans un premier temps à un prototypage des caractères contenus dans les documents traités en utilisant une technique d'appariement de leurs formes. Dans un second temps, les caractères étant représentés par leurs prototypes, les mots ayant des prototypes en commun sont regroupés en familles dites typographiques. Cette étape associe à chaque mot un numéro représentant la famille typographique à laquelle il appartient. Nous avons ajouté à cette propriété des informations supplémentaires que sont la hauteur de chaque mot et l'alignement de la ligne qui le contient.

2.2.2 caractéristiques spatiales

La dispersion des blocs de texte dans une page de documents joue également un rôle significatif dans l'identification de leurs fonctions logiques. Cette dispersion peut être représentée par les relations spatiales entre les blocs de texte dont nous avons considéré différents types au niveau des mots. Le voisinage horizontal comporte les voisins se trouvant sur la gauche et sur la droite d'un

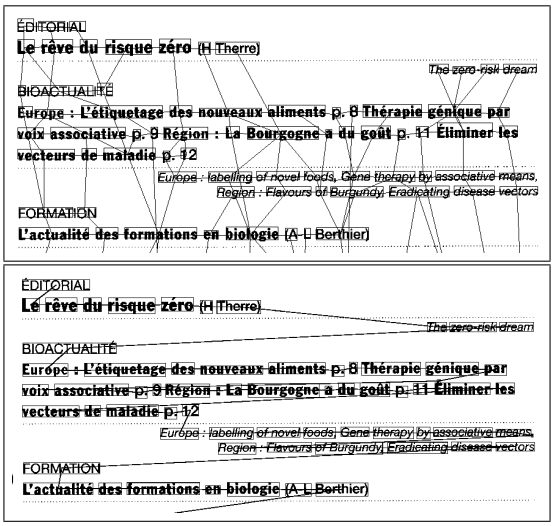


FIG. 5 – Calcul du voisinage.

mot sur le même alignement. Le voisinage vertical considère les voisins les plus proches verticalement pour un bloc donné ayant le plus grand recouvrement au niveau horizontal. Et enfin pour les mots se trouvant en début et en fin de ligne, on considère les mots qui les précèdent ou les suivent sur les lignes précédentes ou suivantes. Le voisinage "ligne précédente ou suivante" est utilisé pour automatiser une extraction naïve de l'ordre de lecture. La figure 5 illustre par un exemple le calcul des voisinages horizontal et vertical et selon le sens de lecture.

2.2.3 Attributs des blocs de texte

A l'issue de ces traitements, nous avons construit un vecteur de 19 caractéristiques physiques discrétisées et réparties sur trois catégories : les attributs portant sur un mot indépendamment de son voisinage (famille typographique, alignement et hauteur), les attributs communs avec le voisinage (distances dans les 4 directions) et les attributs propres à chacun des 4 voisins (haut, bas, gauche, droit). Ces attributs sont montrés au tableau 1 et sont notés pour chaque bloc élémentaire à étiqueter (mot) m comme suit :

- $X(m)$ est l'ensemble des attributs portant uniquement sur le mot m ;
- $X(m')$, pour tout $m' \in V(m)$ avec $V(m)$ l'ensemble des voisins du mot m. Cet ensemble comporte des attributs portant sur les mots voisins considérés.
- $D(m)$ est l'ensemble des attributs reliant le mot m avec son voisinage, correspondant à de l'in-

Attributs	Description	
Attributs propres au bloc		
A_1	X_1	Famille typographique
A_2	X_2	Alignement
A_3	X_3	Hauteur
Attributs partagés avec le voisinage		
A_4	D_1	Distance horizontale à gauche
A_5	D_2	Distance horizontale à droite
A_6	D_3	Distance verticale au dessus
A_7	D_4	Distance verticale au dessous
Voisinage		
A_8, A_9, A_{10}	V_g	bloc voisin à gauche
A_{11}, A_{12}, A_{13}	V_d	bloc voisin à droite
A_{14}, A_{15}, A_{16}	V_h	bloc voisin au dessus
A_{17}, A_{18}, A_{19}	V_b	bloc voisin au dessous

TAB. 1 – Liste des attributs physiques utilisés classés par catégorie.

formation partagée avec les mots se trouvant dans son voisinage. Il s'agit plus particulièrement des distances avec les mots voisins.

2.3 Architecture du système de reconnaissance

Les traitements effectués dans la phase d'analyse sont intégrés dans un système complet d'analyse et de reconnaissance dont le but final est la reconnaissance de la structure logique. Le système que nous proposons est schématisé dans la figure 6. Il se décompose en trois principaux modules : l'analyse, l'apprentissage et la reconnaissance. L'analyse fournit les blocs de texte qui doivent constituer la structure à reconnaître, et leur description au moyen de caractéristiques physiques. L'apprentissage permet de construire une modélisation de la structure logique. Quant à la reconnaissance, en se servant du modèle appris, elle doit extraire automatiquement la structure logique spécifique à un document donné.

Plus concrètement, le but est de concevoir un système de reconnaissance qui permet l'extraction automatique d'entités logiques à plusieurs niveaux hiérarchiques et la reconstruction de la structure logique spécifique à chaque document traité. Cette tâche a été décomposée en plusieurs étapes présentées brièvement comme suit :

– étiquetage logique des blocs élémentaires de texte : associer à chaque bloc sa fonction logique dans le texte ;

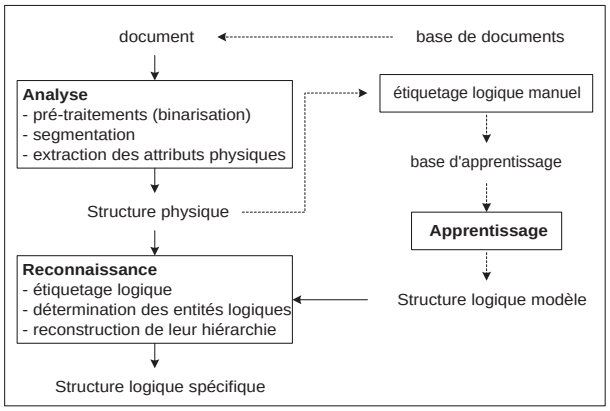


FIG. 6 – Architecture du système de reconnaissance proposé.

– regroupement des séquences de blocs étiquetés ayant la même étiquette afin de constituer les entités logiques du plus bas niveau hiérarchique, que nous appelons entités élémentaires ou simples ;
– détermination des entités logiques des autres niveaux par regroupement récursif ;
– parallèlement à l'étape précédente, reconstitution de la structure logique hiérarchique.

Le modèle de la structure logique est bâti pour chaque revue considérée et doit combiner l'information logique nécessaire à chaque étape de reconnaissance aux données physiques extraites en phase d'analyse. Cette information logique comporte deux parties distinctes :

– l'association d'étiquettes logiques avec les blocs élémentaires de texte qui réalisée manuellement par un opérateur sur les documents de la base d'apprentissage ;
– le regroupement des blocs de texte étiquetés en entités logiques sous forme d'une structure hiérarchique, qui est effectué sans connaissance a priori en se basant la propriété de récurrence typographique des documents considérés.

3 Modélisation du processus de reconnaissance

La modélisation du processus de reconnaissance est basée sur la description physique (typographique et spatiale) des blocs élémentaires de texte d'une part, et de leur ordonnancement logique traduisant le sens de lecture, d'autre part. Nous proposons un modèle hybride adapté aux différentes étapes du processus de reconnaissance qui combine :

– Un classifieur bayésien naïf pour représenter le lien entre les caractéristiques physiques des blocs élémentaires et contribuer à leur étiquetage logique ;
– Un automate probabiliste qui joue un rôle dans toutes les étapes de reconnaissance : (1) dans l'étiquetage logique des blocs élémentaires de texte en combinaison avec le classifieur bayésien dans le but de faire respecter l'ordre logique, et (2) dans la détermination des entités logiques et dans leur regroupement en structure hiérarchique en combinaison avec la troisième partie du modèle ;
– Une structure logique hiérarchique qui représente l'aspect structurel du processus de reconnaissance ; entièrement définie sur la base de l'automate, en combinaison avec celui-ci, elle permet la reconstruction de la structure logique.

Nous décrivons dans ce qui suit chaque étape de reconnaissance et le modèle ou la partie du modèle impliqués dans sa réalisation.

3.1 Etiquetage logique contextuel : combinaison classifieur bayésien et automate probabiliste

Tout d'abord les relations entre les attributs physiques (typographiques et spatiaux) décrivant les blocs élémentaires de texte à étiqueter et leurs fonctions logiques dans le document sont représentées au moyen d'un classifieur bayésien. Nous avons utilisé un classifieur bayésien naïf (CNB) [6, 14] qui est basé sur l'hypothèse d'indépendance entre les attributs en connaissance de la classe. Celle-ci correspond dans notre cas à la notion d'étiquette logique à associer à un bloc élémentaire de texte. Notons Ω la variable aléatoire représentant l'étiquette logique et $A_1, A_2, \ldots, A_n$ les variables qui correspondent aux n attributs considérés. Le rôle du classifieur est de déterminer la classe, c'est-à-dire l'étiquette logique, à partir des valeurs prises par les attributs. Ceci est traduit par le calcul de la probabilité conditionnelle $P(\omega|a_1, a_2, \ldots, a_n)$, pour tout $\omega \in \Omega$, avec $a_1, a_2, \ldots, a_n$ étant le vecteur d'attributs décrivant un bloc élémentaire de texte donné. Le critère de choix classique consiste à retenir la classe ayant la plus grande valeur de cette probabilité conditionnelle. Mais dans notre cas, nous avons ajouté une étape supplémentaire exploitant l'ordre logique des blocs de texte qui n'est pas représenté au niveau du classifieur bayésien.

Outre l'aspect linguistique, la nature linéaire du texte peut servir d'information logique. Cette

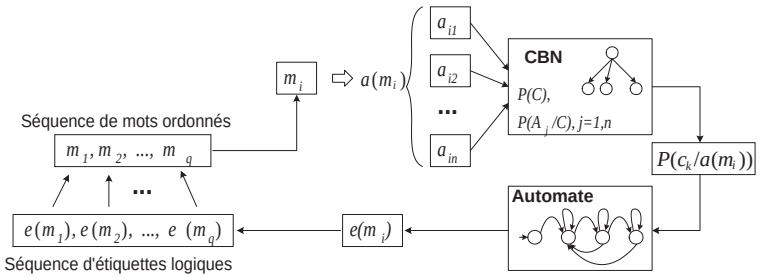


FIG. 7 – Etiquetage logique des blocs élémentaires de texte utilisant la combinaison CBN-Automate.

linéarité se traduit au niveau de la structure physique par l'enchaînement ou la succession des blocs de texte. L'information logique qu'elle véhicule, que l'on peut qualifier de contextuelle, apporte des éléments supplémentaires pouvant contribuer à l'étiquetage logique. C'est dans cette optique que nous avons choisi d'utiliser un automate représentant cette information contextuelle qui a été combiné au classifieur pour réaliser l'étiquetage logique. La structure de base de l'automate utilisé est celle d'un automate fini déterministe car nous n'autorisons pas plus d'une transition à partir d'un état avec un symbole donné. Par ailleurs, il est étendu de façon à l'adapter aux différentes tâches de reconnaissance dans lesquelles il est impliqué. Pour l'étiquetage logique et la combinaison avec le classifieur bayésien, l'automate est augmenté d'une distribution de probabilité associée à la fonction de transition. Ces probabilités sont estimées en phase d'apprentissage de l'automate, en se basant sur les nombres d'occurrences des transitions correspondantes à partir des documents de base d'apprentissage. Pour chaque état, la somme des probabilités de transitions sortantes doit être égale à l'unité. Il ne s'agit donc pas de la définition classique d'un automate stochastique [13].

A ce niveau, le but est d'étiqueter les blocs de texte en tenant compte du contexte dans lequel ils se trouvent, ce contexte étant l'ordre logique représenté par l'automate. L'étiquetage logique est réalisé par la combinaison du classifieur et de l'automate (voir figure 7). Chaque bloc est soumis dans un premier temps au classifieur bayésien qui fournit un vecteur de *probabilités d'étiquetage* sur l'ensemble des étiquettes Ω. Des heuristiques simples sont établies pour l'étiquetage logique combinant les probabilités d'étiquetage fournies par le classifieur et les probabilités de transitions de l'automate. Pour un bloc de texte (un mot) donné, à partir du vecteur de probabilités fourni par le classifieur, nous retenons l'étiquette ayant la probabilité maximum, à condition qu'elle permette une transition ayant elle-même une probabilité non nulle.

Si cette condition n'est pas vérifiée, nous faisons intervenir les probabilités de transition en prenant l'étiquette ayant une probabilité d'étiquetage non nulle avec la probabilité de transition maximale. Sinon, nous nous servirons d'une prédiction d'étiquetage du symbole suivant dans la séquence à analyser. Nous choisissons une étiquette qui n'aboutirait pas à une transition impossible à l'étape suivante tout en prenant en considération les probabilités de transition maximales. Enfin, si aucune condition posée n'est satisfaite, nous décidons de rester dans le même état et donc d'affecter à l'élément en cours l'étiquette associée à la transition en boucle. Il s'agit dans ce dernier cas d'une solution de secours qui consiste en quelque sorte à reporter le traitement du problème rencontré à l'élément suivant dans la séquence.

L'idéal serait qu'il existe une transition de l'état en cours avec l'étiquette ayant la meilleure probabilité fournie par le classifieur, ce qui signifie que la réponse du classifieur convienne à l'automate. Mais malheureusement, ce n'est pas toujours le cas pour des raisons pouvant provenir des niveaux physique et logique :

- niveau physique : il peut s'agir d'une erreur d'étiquetage par le classifieur à cause de perturbations au niveau de la segmentation ou de changement au niveau des attributs physiques. Pour y remédier, il faudrait définir des heuristiques permettant de les repérer et éventuellement de les corriger.
- niveau logique : l'étiquetage réalisé par le classifieur peut être correct mais à cause d'un changement au niveau de la structuration logique qui nécessiterait des transitions non apprises au niveau de l'automate, celui-ci se trouve face à une transition impossible qui sera évidemment refusée. Cette situation correspond à une évolution de la structure logique et le seul moyen d'en tenir compte est de réaliser un apprentissage incrémental du modèle qui permettrait d'ajouter les nouveaux éléments de structuration.

3.2 Reconnaissance de la structure logique hiérarchique

L'enchaînement des blocs de texte sous l'hypothèse de récurrence dans les documents considérés, peut faire ressortir les entités logiques aux niveaux hiérarchiques supérieurs qui se caractérisent par la répétition de sous-séquences similaires d'étiquettes logiques. Afin de repérer ces sous-séquences

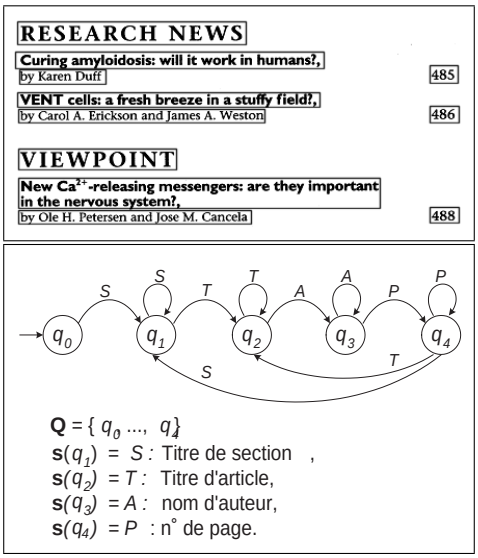


FIG. 8 – Graphe de transitions de l'automate avec la zone de texte qui a servi à son inférence.

qui délimitent au niveau des blocs de texte les entités logiques à déterminer, l'automate utilisé est doté d'une contrainte spécifique supplémentaire. Cette contrainte est traduite par une fonction **s** qui permet d'associer à chaque état un symbole d'entrée unique, ce qui signifie que toutes les transitions vers un état donné q se font avec le même symbole $\mathbf{s}(\mathbf{q})$. Nous montrons à la figure 8 un exemple d'automate où toutes les transitions entrantes vers chaque état sauf l'état initial se font avec le même symbole donné par la fonction **s**. Cette propriété est définie pour servir au repérage des entités logiques à partir du graphe de transition de l'automate.

A l'issue de la phase d'étiquetage logique des blocs élémentaires de texte, on dispose d'une séquence d'étiquettes qui sert de base à la suite du processus de reconnaissance, à savoir l'extraction de la structure hiérarchique spécifique au document traité. Cette structure est composée d'un ensemble d'entités logiques à déterminer à partir de cette séquence de blocs étiquetés. Chaque entité logique dans la *structure logique spécifique* est une instanciation d'une entité type dans la *structure logique modèle* (voir figure 2 et figure 3 à la section 2). Ces entités font également l'objet d'une relation d'inclusion qui correspond à son tour à une instanciation de la relation d'inclusion entre les entités types associées. L'ensemble des entités de la structure spécifique, qui est représentée par un arbre, est complété par les entités élémentaires (de niveau 0) ainsi qu'une entité représentant toute la séquence de blocs de texte à traiter pour un document donné. Cette dernière entité correspondra à la racine de cet arbre, et les entités élémentaires se situent au niveau des feuilles.

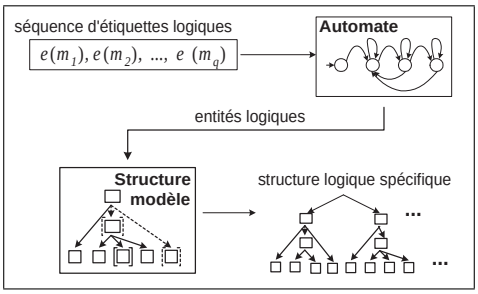


FIG. 9 – Reconstruction de la structure logique hiérarchique spécifique à partir d'une séquence de blocs de texte étiquetés utilisant l'automate et la structure logique modèle appris.

La séquence d'étiquettes initiale subit une phase de réduction qui est effectuée en remplaçant chaque succession d'une même étiquette par une seule occurrence de celle-ci et en conservant les positions de début de fin de cette sous-séquence pour garder la correspondance avec les blocs physiques associés. La séquence d'étiquettes sera ainsi constituée d'*entités logiques élémentaires* (du niveau 0). Ces entités sont ensuite regroupées afin de reconstituer des entités des niveaux supérieurs, en se servant de l'automate et de la structure logique modèle (voir figure 9).

3.3 Phase d'apprentissage

L'automate est construit automatiquement à partir des séquences des étiquettes des blocs de texte extraites dans les documents de la base d'apprentissage. Dans une phase d'initialisation, les séquences d'étiquettes logiques sont réduites en remplaçant la succession d'une même étiquette par une seule occurrence et en conservant le nombre réel d'occurrences. L'ensemble des états de l'automate est initialisé à l'état initial et sa mise à jour consiste à parcourir chaque séquence d'étiquettes en se positionnant sur l'état initial et en ajoutant des états et/ou des transitions. Cette mise à jour doit respecter la contrainte traduite par la fonction **s** qui signifie que chaque état reçoit des transitions avec le même symbole. Le calcul des probabilités de transition est réalisé à partir des nombres d'occurrences comptés au fur et à mesure de la mise à jour des transitions de l'automate. La figure 8 montre un exemple d'automate avec la zone de texte qui a servi à sa construction. Lors de la construction de l'automate, différents cas de figures peuvent se présenter :

– l'état en cours possède une transition avec le symbole en cours ; dans ce cas, aucune modification ne sera apportée à la fonction de transition et seule la probabilité de transition sera mise

à jour ;

- le symbole en cours est déjà associé à un état ; dans ce cas il faut d'abord observer le symbole suivant dans la séquence considérée et vérifier si cet état possède une transition avec ce symbole auquel cas une transition est créée entre l'état en cours et cet état et la probabilité correspondante est initialisée ;
- si les conditions ci-dessus ne sont pas vérifiées, alors un nouvel état est créé, le symbole en cours lui sera associé et une transition avec ce symbole sera créée entre l'état en cours et ce nouvel état ; la probabilité de transition correspondante sera initialisée.

La structure actuelle de l'automate permet de représenter l'enchaînement des étiquettes des blocs élémentaires de texte qui fait ressortir les entités élémentaires de texte. Pour la représentation des entités composées et leur hiérarchie, nous complétons cet automate par une structure arborescente, les éléments de cette structure étant définis sur la base de l'automate. Chaque entité composée appartient à un niveau hiérarchique et est caractérisée au niveau de l'automate par un état de début et un état de fin. Les transitions se trouvant entre les deux états délimitant une entité composée donnée permettent de retrouver les séquences d'étiquettes logiques pouvant la constituer.

L'apprentissage de l'automate est basé sur l'ordonnancement des blocs de texte extrait automatiquement des images de documents, et sur leur étiquetage logique fourni manuellement pour les documents de la base d'apprentissage. Par ailleurs, la construction de la structure hiérarchique est entièrement basée sur la structure de l'automate car aucune information n'est fournie sur les entités logiques de niveaux supérieurs et ni sur leur hiérarchie. Le choix de se limiter à l'étiquetage logique des blocs élémentaires de texte est fait pour exploiter la propriété de récurrence des documents traités et pour reconstituer automatiquement les entités logiques et leur structure hiérarchique. C'est dans cette optique que nous avons défini la contrainte posée sur l'automate qui est traduite par la fonction $\mathbf{s}$. Elle associe à chaque état un symbole entrant unique. De cette façon, nous pouvons observer dans le graphe de transition les débuts et fins de séquences d'étiquettes constituant une entité logique. Les relations d'inclusion entre les entités peuvent également être déterminées de la même manière.

L'automate traduit les récurrences permettant de repérer les entités logiques par des retours en arrière des transitions (voir figure 10). Afin de déterminer ces retours en arrière, nous avons défini une relation d'ordre sur l'ensemble des états qui est basée sur l'ordre de création des états ainsi que sur l'accessibilité d'un état à partir d'un autre état. Par conséquent, la présence d'une transition d'un état vers un autre état qui le précède dans l'ordre défini signifie qu'une sous-séquence se répète, ce qui donne lieu à une entité logique. Chaque entité logique e est caractérisé par un niveau hiérarchique noté $\texttt{niv}(e)$ et est marquée au niveau de l'automate par un état de début $\texttt{début}(e)$ et un état de fin $\texttt{fin}(e)$ et à partir desquels on peut extraire l'ensemble des séquences étiquettes (d'entités élémentaires) qui la composent. La détermination de ces séquences consiste à retrouver les symboles se trouvant sur tous les chemins (en avant) possibles entre les états de début et de fin de l'entité. L'apprentissage de la structure modèle est donc réalisé selon les étapes suivantes.

- On augmente l'automate appris d'un relation d'ordre notée $<$ qui est définie comme suit. Pour deux états $q_i, q_k \in \mathbf{Q}$, avec $\mathbf{Q}$ étant l'ensemble des états de l'automate, on dit que $q_i < q_k$, c'est-à-dire que q_i précède q_k, s'il existe un chemin qui mène de q_i, à q_k mais pas de q_k à q_i, ou encore, si ce dernier existe, que sa création lors de l'apprentissage ait eu lieu après celle du chemin de q_i, à q_k chronologiquement lors de l'apprentissage de l'automate. L'ordre de création des états de l'automate est par conséquent indispensable à la détermination de cette relation d'ordre.
- On extrait des entités logiques peut ainsi être réalisée en repérant les états marquant leurs débuts et fins. Soient deux états $q_i, q_k \in \mathbf{Q}$ tels que $q_i < q_k$. S'il existe une transition de q_k vers q_i, alors q_i et q_k correspondent respectivement au début et à la fin d'une entité logique.
- On détermine les relations d'imbrication entre les entités logiques extraites. Soient deux entités e_i, e_j. On peut dire que e_i est incluse dans e_j si $\texttt{début}(e_j) < \texttt{début}(e_i)$ et $\texttt{fin}(e_i) < \texttt{fin}(e_j)$ ou $\texttt{fin}(e_i) = \texttt{fin}(e_j)$, ce qui signifie que les états marquant l'entité e_i doivent se trouver entre ceux délimitant l'entité e_j. Enfin, pour chaque entité, l'entité parent correspondra à la plus petite entité selon la relation d'inclusion définie.
- On attribue ensuite à chaque entité un niveau hiérarchique numéroté à partir de 1. Le principe consiste à affecter le niveau 1 à toutes les entités, et à augmenter de 1 les niveaux des entités parents par rapport à leurs descendants jusqu'à mise à jour complète construisant ainsi un de

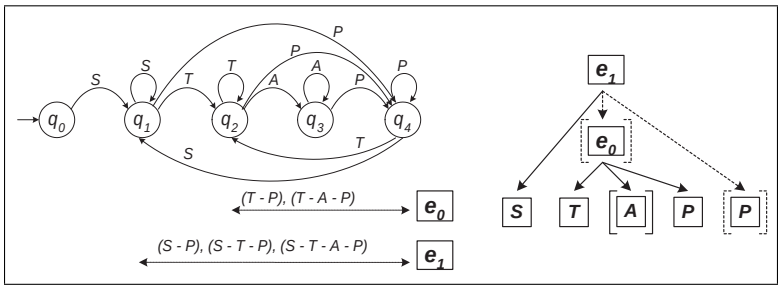


FIG. 10 – Détermination des entités logiques sur plusieurs niveaux hiérarchiques à partir du graphe de transitions de l'automate appris.

la structure hiérarchique.

4 Expérimentation

Nous avons implémenté un prototype du système de reconnaissance proposé, composé des trois modules d'analyse, d'apprentissage et de reconnaissance. Nous l'avons expérimenté sur une base de sommaires de 6 revues (4 en langue française et 2 en langue anglaise). Pour chaque revue, une partie des documents a servi à l'apprentissage et l'autre au test du système conçu. La répartition de la base sur les deux ensembles d'apprentissage et de test est montrée au tableau 2. Nous soulignons le choix de travailler sur un nombre réduit de documents en apprentissage dans le but de rendre le système utilisable dans un contexte réel. Le choix des numéros servant à l'apprentissage est, pour la plupart des revues, effectué selon l'ordre chronologique de leur apparition. La figure 11 montre des images binarisées de documents pour l'ensemble des revues. Différents cas de figures ont été observés dans ces documents, ce qui permet d'étudier des formes variables de structures et de problèmes qui s'y attachent. Pour chaque revue la liste des étiquettes logiques qu'elle contient est établie au tableau 3. Ces étiquettes serviront à définir les symboles de l'automate qui représentera chaque revue.

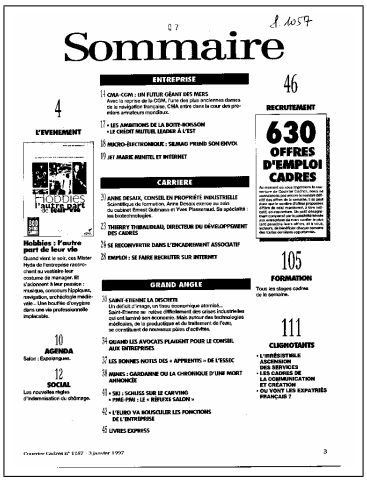
L'étude expérimentale effectuée sera présentée en trois parties :

- l'étiquetage logique contextuel utilisant la combinaison du CBN et de l'automate ;
- l'apprentissage automatique de la structure logique modèle ;
- la reconnaissance de la structure logique spécifique.

		Apprentissage		Test	
N°	Revue	#pages	#mots	#pages	#mots
[1]	Biofutur	3	492	7	1 217
[2]	Cadres	3	615	6	1 168
[3]	Cahiers des préventions	3	474	7	1 326
[4]	NewsWeek	3	641	11	2 445
[5]	TINS	3	682	9	2 312
[6]	M/S	3	1 136	12	4 305
	Total	18	4 040	52	12 773

TAB. 2 – Répartition de la base de documents pour l'ensemble des revues considérées.

[1] : *Biofutur*



[2] : *Cadres*

[3] : *Cahiers des préventions*

[4] : *NewsWeek*

[5] : *TINS*

[6] : *M/S*

FIG. 11 – Exemples de tables des matières dans les revues de la base de documents.

[**1**] : *Biofutur*

Symbole	**Etiquette**
S	Titre de section
T	Titre d'article
A	Auteurs
P	N° de page
R	Titre en anglais

[**2**] : *Cadres*

Symbole	**Etiquette**
T	Titre d'article
A	Auteurs
R	Résumé

[**3**] : *Cahiers de préventions*
[**4**] : *NewsWeek*
[**5**] : *TINS*
[**6**] : *M/S*

Symbole	**Etiquette**
S	Titre de section
T	Titre d'article
A	Auteurs
P	N° de page

TAB. 3 – Etiquettes logiques présentes dans les revues de la base d'exemples. Nous faisons correspondre aux étiquettes logiques des lettres qui seront considérées comme étant les symboles des automates. Les revues [**3**], [**4**], [**5**] et [**6**] ont les mêmes étiquettes.

4.1 étiquetage logique

Pour chaque revue un classifieur bayésien et un automate ont été construits à partir des documents de la base d'apprentissage correspondante. Par la suite, l'étiquetage logique est réalisé sur les documents de la base de test à l'aide de cette combinaison classifieur-automate. Afin d'évaluer l'apport de l'automate dans l'opération d'étiquetage, nous avons également mesuré le taux de reconnaissance obtenu en utilisant le classifieur bayésien seul. Les résultats de performance de cet étiquetage sont résumés à la figure 12 qui montre les taux de reconnaissance $\mathbf{R_{CBN}}$ et $\mathbf{R_{CNB-Auto}}$ correspondant respectivement au classifieur bayésien naïf (CBN) et à la combinaison du CBN avec l'automate. Les symboles (+) et (−) indiquent respectivement l'amélioration et la diminution du taux de reconnaissance avec l'automate par rapport au CBN seul. Ces taux sont calculés à partir des nombres de blocs non correctement étiquetés par le CBN seul et correctement étiquetés avec l'automate, et inversement.

D'après la figure 12, globalement la différence entre les taux de reconnaissance $\mathbf{R_{CBN}}$ et $\mathbf{R_{CBN-Auto}}$ est très faible. Nous pouvons constater que, notamment pour les revues *Cahiers de préventions*, *Newsweek*, et *TINS*, les taux de reconnaissance étaient déjà assez bons avec le CBN seul, et l'utilisation de l'automate n'a pratiquement pas d'effet. Les revues qui font l'exception sont la revue *Biofutur* qui a fait l'objet d'une amélioration très nette en utilisant l'automate et la revue *M/S* qui a

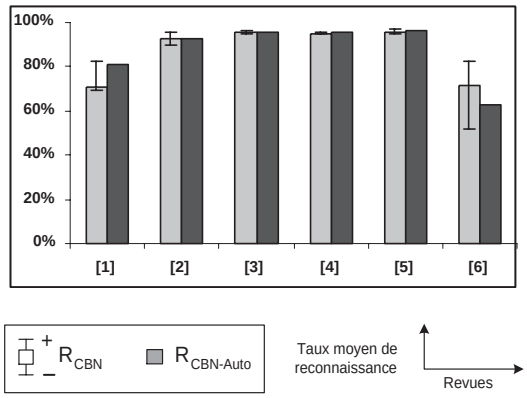


FIG. 12 – Comparaison des taux reconnaissance moyens $\mathbf{R_{CBN}}$ et $\mathbf{R_{CBN-Auto}}$ sur l'ensemble des revues.

donné un résultat inverse. Pour cette revue, les documents pour lesquels les taux de reconnaissance ont diminué ont globalement des taux déjà faibles avec le CBN.

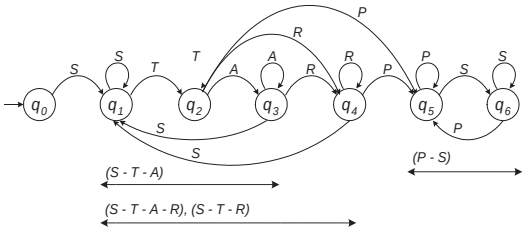
Nous pouvons constater à l'issue de cette expérience, que la phase d'étiquetage avec le CBN est très importante pour la suite du processus de reconnaissance. Cependant, la diminution des taux de reconnaissance avec l'automate n'était pas due uniquement à de faibles taux de reconnaissance obtenus avec le CBN seul. Outre les problèmes liés au niveau physique, de nombreux cas dans la base de test présentaient des modifications dans la structure logique elle-même qui donnaient lieu à des transitions inconnues au niveau de l'automate. Il est à souligner que l'algorithme de reconnaissance utilisant l'automate est ouvert à toute modification afin de considérer d'autres heuristiques qui pourraient être extraites automatiquement, dans le but de s'adapter aux problèmes autant au niveau physique qu'au niveau logique. Nous avons pour l'instant défini ces règles qui sont communes aux documents que nous traitons tout en essayant de maintenir le caractère générique de notre approche.

4.2 apprentissage de la structure hiérarchique

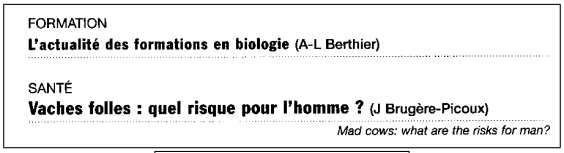
Dans la phase d'apprentissage de la structure hiérarchique, différents cas de figures ont été observés dans les 6 revues traitées. Pour les revues *Cahiers des préventions* et *M/S*, nous avons rencontrés des problèmes de même nature, qui sont liés à une structuration complexe nécessitant une extension de notre représentation de la structure logique à un graphe au lieu d'un arbre.

Les revues *Biofutur* et **[2]** :*Cadres* comportent uniquement des entités logiques d'un seul niveau

[1] : Biofutur



Entité	Niveau	Séquences
e_0	1	*(S-T-R), (S-T-A-R)*
e_1	1	*(S-T-A)*
e_2	1	*(P-S)*


FORMATION
L'actualité des formations en biologie (A-L Berthier)

SANTÉ
Vaches folles : quel risque pour l'homme ? (J Brugère-Picoux)
Mad cows: what are the risks for man?

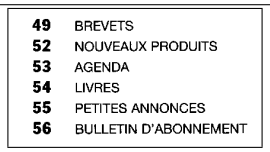
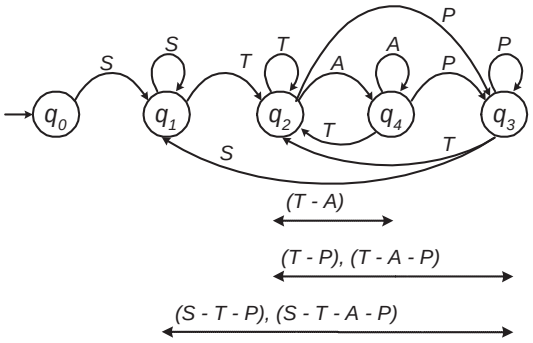

49 BREVETS
52 NOUVEAUX PRODUITS
53 AGENDA
54 LIVRES
55 PETITES ANNONCES
56 BULLETIN D'ABONNEMENT

FIG. 13 – Apprentissage de la structure logique hiérarchique pour la revue *Biofutur*.

de structuration en plus du niveau 0 qui comporte les entités élémentaires composées de séquences d'étiquettes logiques. Nous montrons à la figure 13 l'automate et les entités logiques extraites pour la revue *Biofutur*, ainsi que des exemples de zones de texte dans les documents d'apprentissage montrant ces entités logiques. Il est à noter que celles-ci ont été correctement extraites malgré une structure de l'automate relativement complexe.

Pour les revues *NewsWeek* et *TINS*, les structures logiques sont également correctement apprises, en revanche elles comportent deux niveaux de structuration en plus du niveau 0 (voir figures 14 et 15). Bien qu'il n'y ait pas erreur au niveau de l'extraction des entités logiques pour la revue *TINS*, nous remarquons une transition supplémentaire entre les états q_2 et q_4 avec le symbole P (numéro de page) qui se trouve concrètement selon cette transition entre les symboles T (titre d'article) et A (nom d'auteur). Cette transition donne lieu à une structure d'article de la forme "*T-P*" alors que dans les documents composants la base d'apprentissage, les articles sont organisés selon la séquence "*T-A-P*". L'origine de cette transition est une erreur au niveau du calcul du voisinage à cause de la taille des caractères utilisés pour les numéros de pages (le symbole P). Bien que le numéro de page

[4] : NewsWeek



Entité	Niveau	Séquences	Parent
e_0	1	*(T-P), (T-A-P)*	e_2
e_1	1	*(T-A)*	e_2
e_2	2	*(S-T-P), (S-T-A-P)*	–

WORLD AFFAIRS

Iraq: Who's In Charge Here? **6**
The Hunt for His Secret Weapons **10**
A New Breed of Killers **12**
Interview: Spending Ted's Money **13**

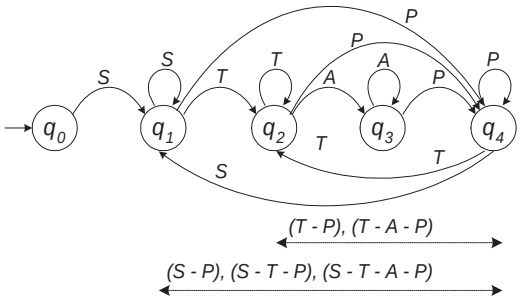
EUROPE

Russia: Get Me Rewrite
by Bill Powell **14**
Belgium: This Week's Horror **16**
Opinion: Working Together?
by Michael Elliott **17**

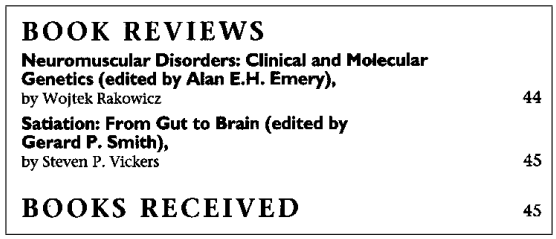
FIG. 14 – Apprentissage de la structure logique hiérarchique pour la revue *NewsWeek*.

soit aligné avec les noms d'auteurs, pour un cas particulier, il a été considéré comme suivant le titre de l'article qui se terminait par un caractère (une virgule) se trouvant légèrement plus bas que les autres caractères de la ligne. Ce genre d'erreur n'est pas facilement détectable lors du traitement du niveau physique, mais des anomalies conséquentes au niveau logique nous permettent de les repérer afin de les considérer en phase d'analyse.

[5] : TINS



Entité	Niveau	Séquences	Parent
e_0	1	*(T-A-P), (T-P)*	e_1
e_1	2	*(S-T-A-P), (S-T-P), (S-P)*	–


BOOK REVIEWS

Neuromuscular Disorders: Clinical and Molecular Genetics (edited by Alan E.H. Emery), by Wojtek Rakowicz 44

Satiation: From Gut to Brain (edited by Gerard P. Smith), by Steven P. Vickers 45

BOOKS RECEIVED 45

FIG. 15 – Apprentissage de la structure logique hiérarchique pour la revue *TINS*.

4.3 Reconnaissance multi-niveaux de la structure

L'apprentissage des parties du modèle impliquées dans la reconnaissance de la structure hiérarchique est une tâche très complexe pour de nombreuses raisons : le nombre réduit de documents dans les bases d'apprentissage, les erreurs provenant du traitement au niveau physique à cause de la complexité de la mise en page des documents traités, et le fait qu'aucune connaissance sur les entités logiques et leur structure n'est fournie pour l'apprentissage. En effet, nous n'utilisons que les étiquettes logiques des blocs élémentaires de texte et le reste de l'information logique utilisée dans le processus d'apprentissage est extraite de manière entièrement automatique. Notre objectif dans cette démarche est de concevoir un système utilisable dans des conditions réelles, allégeant la charge de l'opérateur tout en étant interactif. Malgré toutes ces difficultés, les résultats de l'apprentissage des automates et des structures hiérarchiques associées sont très encourageants. Pour la majorité des revues, les structures modèles extraites sont correctes, et les cas où nous avons observé des erreurs mineures peuvent nous guider dans l'amélioration, notamment au niveau des règles établies pour la construction de l'automate et l'extraction des entités logiques.

Nous présentons dans ce qui suit les résultats de la phase de reconnaissance de la structure hiérarchique utilisant le modèle appris, et ce pour chaque revue de notre base de documents. Le but de cette phase expérimentale est de mesurer la performance de notre modèle d'une part, et d'observer les répercussions des erreurs accumulées des phases précédentes de traitements, d'autre part.

Il est cependant difficile de réaliser une évaluation quantitative de cette phase de reconnaissance car aucune connaissance n'est fournie en ce qui concerne ce niveau de reconnaissance. En d'autres termes, nous ne disposons pas de la structure logique réelle des documents traités. Nous avons donc dû nous baser uniquement sur l'étiquetage logique des blocs élémentaires de texte, qui est la seule information logique disponible. Notre expérience consiste à réaliser, pour chaque revue et chaque document de la base de test, le processus de reconnaissance de la structure logique en se basant sur la structure modèle apprise et en utilisant : (1) la séquence d'étiquettes logiques qui correspond à la réponse de la combinaison du classifieur bayésien et de l'automate, (2) la séquence d'étiquettes réelles qui sont fournies manuellement par l'opérateur et qui sont utilisées notamment pour mesurer

N°	Revue	(a)	(b)
[1]	*Biofutur*	2	5
[2]	*Cadres*	6	0
[3]	*Cahiers des préventions*	4	3
[4]	*NewsWeek*	11	0
[5]	*TINS*	8	1
[6]	*M/S*	11	1

TAB. 4 – (a) : le nombre de documents de la base de test dont l'étiquetage logique est reconnu par l'automate appris, (b) : le nombre de documents comportant une évolution de la structuration logique et qui ne sont pas reconnus par l'automate.

les taux de reconnaissance lors de l'étiquetage logique.

Le but est de comparer les résultats de la reconnaissance basée sur l'étiquetage logique réalisé automatiquement, à ceux obtenus avec l'étiquetage réel qui a servi de repère jusqu'à présent. Cependant, nous avons rencontré un problème au niveau de la reconnaissance utilisant l'étiquetage réel. Il s'agit des séquences d'étiquettes, dans les documents de la base de test, qui ne sont pas reconnues par l'automate. Ceci est dû à l'évolution de la structure logique dans certains documents qui se traduit par la présence de transitions qui ne sont pas apprises au niveau l'automate. Par conséquent, nous présentons les mesures effectuées sur les documents dont l'étiquetage réel s'accorde avec l'automate appris. Nous récapitulons dans le tableau 4 pour chaque revue le nombre de documents (de la base de test) qui ont fait l'objet d'une évolution dans la structuration logique. La détection de cette évolution est réalisée de manière automatique par l'automate à partir des séquences d'étiquettes réelles. Les documents en question ont fait apparaître des transitions inconnues par l'automate.

Les mesures de comparaison et d'évaluation des résultats de cette phase de reconnaissance sont basées sur les nombres d'entités logiques extraites pour chaque document dans les deux processus de reconnaissance lancés. Soit pour chaque document à un niveau donné, N_E le nombre d'entités extraites en utilisant le résultat de l'étiquetage automatique, et N_R le nombre d'entités obtenues à partir de l'étiquetage réel. Pour chaque revue, l'évaluation est réalisée sur l'ensemble des documents de la base de test dont l'étiquetage réel répond à l'automate appris. Les différentes erreurs de reconnaissance de la structure hiérarchique se présentent sous forme de fragmentation d'une entité en deux ou plusieurs entités, de fusion d'entités voisines en une seule entité ou de recouvrement d'entités voisines. Il est à noter que les entités logiques obtenues en utilisant l'étiquetage logique

[2] *Cadres* – 2 niveaux

	Niveau 0		Niveau 1	
Document	$\mathbf{N_E}$	$\mathbf{N_R}$	$\mathbf{N_E}$	$\mathbf{N_R}$
1	39	19	13	8
2	31	29	14	13
3	30	25	13	11
4	31	29	14	13
5	23	23	10	10
6	30	27	13	12

TAB. 5 – Reconnaissance multi-niveaux sur la base de test de la revue *Cadres*. $\mathbf{N_E}$ le nombre d'entités extraites à partir de l'étiquetage réalisé à l'aide de la combinaison classifieur-automate ; $\mathbf{N_R}$ le nombre d'entités extraites à partir de l'étiquetage réel.

réel ne correspondent pas forcément aux entités réelles qu'il faudrait extraire.

La revue *Biofutur* ne compte que deux documents dans la base de test qui n'ont pas subi une évolution de la structure logique. Nous ne pouvons pas donner une évaluation globale en se basant sur ces deux documents. En revanche, en testant la reconnaissance sur la base d'apprentissage, les résultats sont quasiment parfaits ce qui montre que la structure a été correctement apprise. Pour les revues *Cahiers des préventions* et *M/S* les erreurs constatées au niveau de l'apprentissage de la structure hiérarchique ne nous permettent pas d'analyser la phase de reconnaissance de manière significative. Nous présenterons dans ce qui suit les résultats des revues *Cadres*, *NewsWeek* et *TINS*.

Dans la revue *Cadres*, tous les documents de la base de test sont conformes à la structure logique apprise, malgré les erreurs de segmentation importantes que comportent ces documents. Les résultats présentés dans le tableau 5 sont globalement assez bons. La différence entre la reconnaissance au niveau 1 avec les étiquettes réelles et celles extraites automatiquement est très faible malgré la présence d'erreurs d'étiquetage logique. L'exemple présenté à la figure 16 permet d'illustrer cette situation. Les erreurs d'étiquetage sont montrées par des blocs rectangulaires noirs dans l'image du milieu, les autres blocs étant correctement étiquetés. Nous remarquons également que les titres de sections ne sont pas considérés de même que pour les documents de la base d'apprentissage, ceci provient du fait qu'ils sont écrits en vidéo-inverse, qui n'est pas traitée par l'outil de binarisation utilisé. Ces titres ont donc été ignorés dans l'ensemble des documents de cette revue.

Les documents de la base de test de la revue *NewsWeek* sont tous conformes à la structure

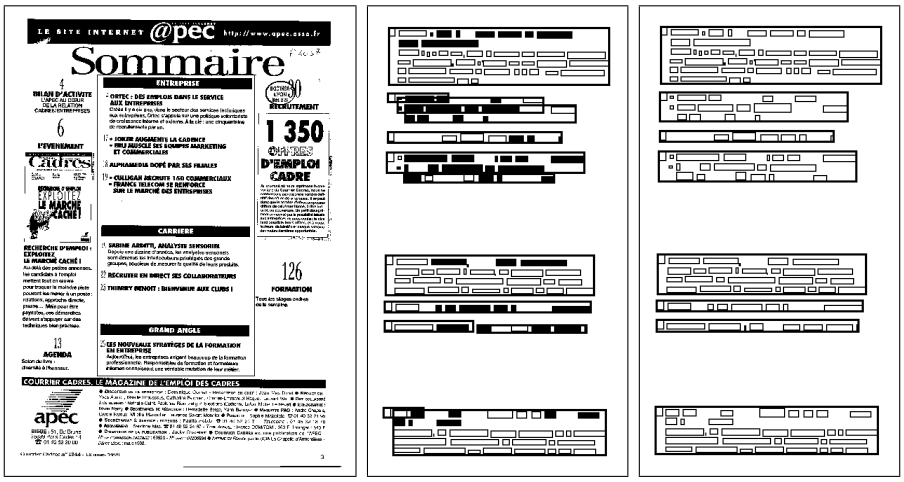


FIG. 16 – Résultats de l'extraction des entités logiques du niveau 1 dans un document de la base de test de la revue *Cadres* : au milieu, à partir de l'étiquetage réalisé automatiquement, et à droite, à partir de l'étiquetage réel.

logique apprise (voir tableau 6). Ceci nous a permis de tester notre processus de reconnaissance sur un nombre raisonnable de documents pour cette revue. Les résultats de reconnaissance obtenus sont très satisfaisants. Nous montrons à la figure 17 les entités logiques des niveaux 1 et 2 dans deux documents de la base de test. Nous faisons remarquer que pour cette revue, les taux de reconnaissance obtenus par le classifieur étaient déjà très bons.

Comme la revue *NewsWeek*, la revue *TINS* comporte 3 niveaux de structurations et tous les documents de sa base de test gardent une structuration conforme à celle apprise dans notre modèle. Elle affiche également de très bons résultats de reconnaissance (voir le tableau 7). Les résultats de l'extraction des entités des niveaux 1 et 2 dans un document sont montrés à la figure 18, et ce avec chacun des étiquetages calculé et réel.

D'après les résultats obtenus dans cette phase de reconnaissance, nous pouvons constater d'une manière générale que notre système nous a permis d'atteindre les objectifs fixés. Cependant, il demeure dépendante de la phase d'étiquetage logique. D'autres erreurs de reconnaissance sont dues à l'évolution de la structure logique.

[**4**] *NewsWeek* – 3 niveaux

Niveau 0		Niveau 1		Niveau 2	
$\mathbf{N_E}$	$\mathbf{N_R}$	$\mathbf{N_E}$	$\mathbf{N_R}$	$\mathbf{N_E}$	$\mathbf{N_R}$
72	72	26	26	7	7
68	68	25	25	7	7
76	71	29	27	7	7
65	65	23	22	7	7
58	60	22	22	6	5
71	73	24	26	7	6
74	74	26	26	7	7
77	75	28	27	7	7
74	73	30	26	7	7
69	69	25	25	7	7
77	75	27	26	7	7

TAB. 6 – Reconnaissance multi-niveaux sur la base de test de la revue *NewsWeek*. $\mathbf{N_E}$ le nombre d'entités extraites à partir de l'étiquetage réalisé à l'aide de la combinaison classifieur-automate ; $\mathbf{N_R}$ le nombre d'entités extraites à partir de l'étiquetage réel.

[**5**] *TINS* – 3 niveaux

Niveau 0		Niveau 1		Niveau 2	
$\mathbf{N_E}$	$\mathbf{N_R}$	$\mathbf{N_E}$	$\mathbf{N_R}$	$\mathbf{N_E}$	$\mathbf{N_R}$
53	57	15	16	7	7
43	39	12	10	8	8
54	49	15	13	8	8
56	57	17	16	7	8
59	59	16	16	9	9
51	49	14	14	8	7
36	36	10	10	5	5
53	48	15	14	7	6

TAB. 7 – Reconnaissance multi-niveaux sur la base de test de la revue *TINS*. $\mathbf{N_E}$ le nombre d'entités extraites à partir de l'étiquetage réalisé à l'aide de la combinaison classifieur-automate ; $\mathbf{N_R}$ le nombre d'entités extraites à partir de l'étiquetage réel.

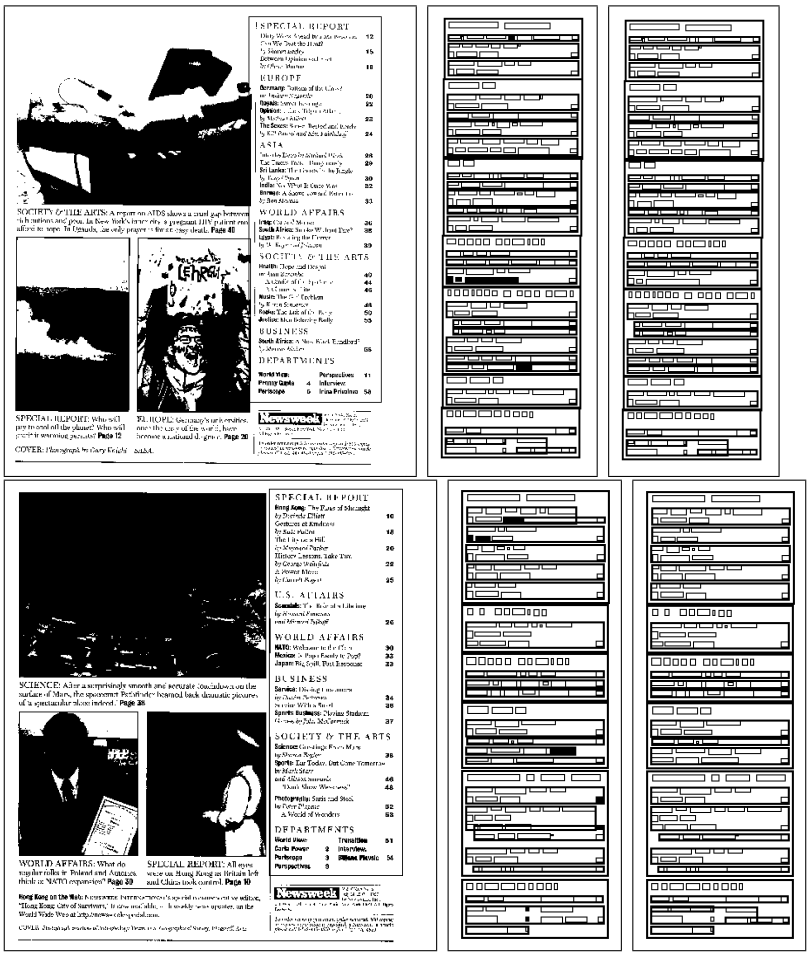


FIG. 17 – Résultats de l'extraction des entités logiques des niveaux 1 et 2 dans deux documents de la base de test de la revue *NewsWeek* : au milieu, à partir de l'étiquetage réalisé automatiquement, et à droite, à partir de l'étiquetage réel.

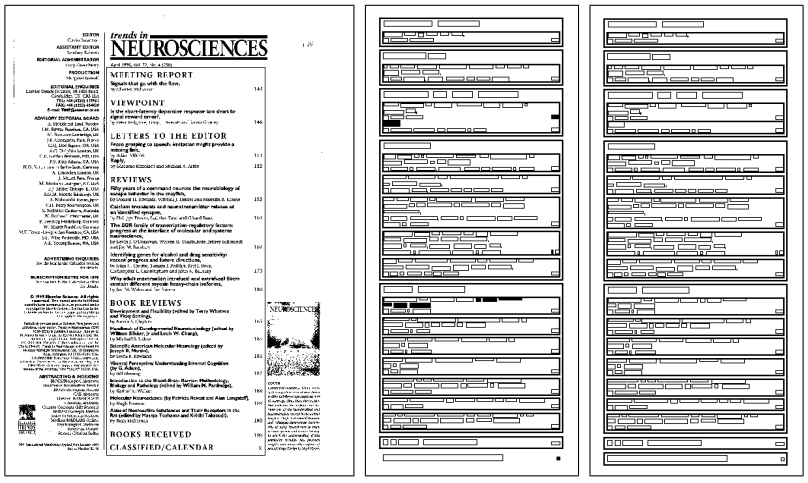


FIG. 18 – Résultats de l'extraction des entités logiques des niveaux 1 et 2 dans un document de la base de test de la revue *TINS* : au milieu, à partir de l'étiquetage réalisé automatiquement, et à droite, à partir de l'étiquetage réel.

5 Conclusion

A travers cet article nous avons présenté un système de reconnaissance automatique de structures logiques hiérarchiques pour les sommaires de revues. Pour la modélisation du système nous avons utilisé un classifieur bayésien naïf, un automate à états finis probabiliste et une structure arborescente représentant les entités logiques de différents niveaux qui constituent la structure du contenu d'un document. Ce système a été testé sur une base de documents caractérisés à la fois par une mise en page relativement complexe et par une structuration très variable.

L'étude des différentes parties composant notre modèle a été conduite en trois principales étapes, visant chacune un objectif particulier :

- l'étiquetage logique en combinant le classifieur et l'automate qui représente l'enchaînement des blocs de texte à étiqueter : les résultats de cette expérience ont montré l'importance de la réponse du classifieur pour l'étiquetage logique qui est déterminante dans la tâche de validation par l'automate, laquelle est à son tour déterminante dans la suite du processus de reconnaissance.
- l'extraction des différentes entités logiques à partir de la structure de l'automate : nous avons examiné le comportement de cette phase d'apprentissage du modèle pour chaque revue de notre base de documents. Ce module a correctement fonctionné pour certaines revues, mais nous avons cependant relevé certaines anomalies dues à l'insuffisance des règles que nous avons établies pour l'extraction des entités logiques et des relations qui les lient.
- L'extraction de la structure logique spécifique à un document d'une revue donnée en utilisant le modèle appris : dans cette partie nous avons soulevé différents types de problèmes que nous avons également tenté d'expliquer. En plus des difficultés provenant des étapes de traitements antérieures, nous avons également rencontré des cas d'évolution de la structure logique dans les documents de la base de test par rapport à celle apprise à partir de la base d'apprentissage. Enfin, la troisième catégorie de problèmes rencontrés, est due à l'étape d'apprentissage de la structure logique. Elle concerne les revues qui ont fait l'objet de quelques erreurs dans la détermination des entités logiques. Il faut donc considérer ces problèmes au niveau de l'étape d'apprentissage citée dans le point précédent.

De plus, nous avons pu voir concrètement les conséquences des problèmes du niveau physique sur le niveau logique. L'adoption d'une approche probabiliste avait pour but l'adaptation à ce genre de problèmes, mais certains de ces problèmes nécessitent un retour vers le niveau physique afin de procéder à une correction guidée par l'information du niveau logique qui est extraite automatiquement ou fournie manuellement.

Les difficultés rencontrées sont également dues aux contraintes que nous nous sommes posées afin de concevoir un système de reconnaissance automatique qui soit opérationnel en tenant compte de conditions de travail réelles. Ces contraintes se traduisent notamment par le fait que nous ayons choisi d'utiliser un nombre réduit de documents pour la phase d'apprentissage, et que l'intervention de l'opérateur est minimisée car elle se limite à l'étiquetage logique des blocs élémentaires de texte ne donnant aucune information sur les entités logique et leur structure hiérarchique.

Références

[1] O. Altamura, F. Esposito, D. Malerba. Transforming paper documents into XML with WISDOM++. *IJDAR : International Journal on Document Analysis and Recognition*, 4(1) :2–17, August 2001.

[2] A. Belaïd. *Cours INRIA : Le traitement électronique du document*, pages 49–92. Collection ADBS, Aix–en–Provence, Octobre 1994.

[3] A. Belaïd. Recognition of table of contents for electronic library consulting. *IJDAR : International Journal on Document Analysis and Recognition*, 4(1) :35–45, August 2001.

[4] R. Brugger, A. Zramdini, R. Ingold. Modeling documents for structure using generalized N–Grams. 4^{th} *ICDAR : International Conference on Document Analysis and Recognition*, volume 1, pages 56–60, Ulm, Germany, August 1997.

[5] A. Dengel, F. Dubiel. Clustering and Classification of Document Structure –A Machine Learning Approach–. 3^{th} *ICDAR : International Conference on Document Analysis and Recognition*, volume 2, pages 587–591, Montréal, Canada, August 1995.

[6] R. O. Duda, P. E. Hart. *Pattern Classification and Scene Analyis* . John Wiley and Sons, New York, 1973.

[7] L. Duffy. *Recherche d'information logique dans les documents à typographie riche et récurrente, application aux sommaires*. Thèse de doctorat, INSA de Lyon, Lyon, France, Décembre 1997.

[8] R. Haralick. Docment Image Understanding : A Geometric and Logical Layout. *CVPR'94 : Computer Vision and Pattern Recognition*, pages 385–390, Seattle, USA, June 1994.

[9] J. Higashino et al. A knowledge–based segmentation method for document understading. 8^{th} *ICPR : International Conference on Pattern Recognition*, volume 1, pages 745–748, Paris, France, October 1986.

[10] A. Jacquesson, A. Rivier. *Bibliothèques et documents numériques : Concepts, composants, techniques et enjeux*. Electre - Éditions du Cercle de la Librairie, Paris, 1999.

[11] A. K. Jain, B. Yu. Document representation and its Application to page decomposition. *PAMI : Pattern Analysis and Machine Intelligence*, 20(3) :294–308, 1998.

[12] F. LeBourgeois, H. Emptoz. Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies. 5^{th}*ICDAR : International Conference on Document Analysis and Recognition*, pages 177–180, Bangalore, India, September 1999.

[13] L. Miclet. *Méthodes structurelles pour la reconnaissance de formes*. Eyrolles, Paris, France, 1984.

[14] W. Iba P. Langley, K. Thompson. An analysis of Bayesian classifiers. *Proceedings of the Tenth Annual Conference on Artificial Intelligence*, pages 223–228, Menlo Park, CA, USA, 1992. AAAI Press.

[15] F. Richaudeau. *Manuel de typographie et de mise en page*. Éditions Retz, Paris, 1989.

[16] S. Souafi-Bensafi. *Contribution à la reconnaissance des structures des documents écrits : Approche probabiliste*. Thesis, INSA de Lyon, France et Université Laval, Québec-Canada, 2002.

[17] Y. Y. Tang, C.D. Yan, M. Cheriet, C. Y. Suen. *Handbook of Pattern Recognition and Computer Vision*, chapter 3.6 : Automatic analysis and understanding of documents, pages 625–654. World Scientific Pub., Singapore, 1993.

[18] S. L. Taylor, M. Lipshutz. Document understanding system for multiple document representations. *DAS : Document Analysis Systems*, pages 155–171, Malvern, Pensylvania, October 1996.

[19] J. Toyada et al. Study of extracting Japanese newspaper article. 6^{th} *ICPR : International Conference on Pattern Recognition*, volume 2, pages 1113–1115, Munich, Germany, October 1982.

[20] H. Walischewski. Automatic Acquisition for Spatial Document Interpretation. 5^{th} *ICDAR : International Conference on Document Analysis and Recognition*, pages 317–320, Bangalore, India, september 1999.

[21] T. Watanabe, X. Huang. Automatic Acquisition of Layout Knowledge for Understading Business Cards. 4^{th} *ICDAR : International Conference on Document Analysis and Recognition*, volume 1, pages 216–220, Ulm, Germany, 1997.

[22] C. L. Yu, Y. Y. Tang, C. Y. Suen. Document Architecture Language (DAL) Approach to Document Processing. 2^{th} *ICDAR : International Conference on Document Analysis and Recognition*, pages 103–106, Tsukuba Science City, Japan, October 1993.