Indexation Sémantique des Vidéos : cas des Journaux Télévisés Arabes.

Sadek Mansouri

REPUBLIQUE TUNISIENNE
Ministère de l'Enseignement Supérieur
Et de la Recherche Scientifique

UNIVERSITE DE SFAX
Faculté des Sciences Économiques et de Gestion de Sfax
Laboratoire de recherche : RLANTIS

THÈSE DE DOCTORAT

Pour L'obtention du Titre de Docteur en

Informatique

Indexation Sémantique des Vidéos : cas des Journaux Télévisés Arabes.

Présentée et soutenue publiquement le 28/12/2021 par :

Sadek MANSOURI

Membres du jury :

M. Kais HADDAR	**Professeur d'Ens. Sup – FSS- Sfax**	**Président**
M. Mounir ZRIGUI	**Professeur d'Ens. Sup – FSM- Monastir**	**Directeur de thèse**
M. Abdelmajid BEN HAMADOU	**Professeur d'Ens. Sup –ISIMS-Sfax**	**Rapporteur**
Mme. Fadoua DRIRA	**Maître de Conférences – ENIS-Sfax**	**Rapporteur**
Mme.Emna JAMMOUCI	**Maître de Conférences – FSS-Sfax**	**Membre**

Année Universitaire : 2021-2022

Dédicaces

A ma chère mère Saida, A mon cher père Fethi

Que DIEU vous garde

Trésor de bonté, de générosité et de tendresse,

En témoignage de ma grande affection et ma profonde gratitude pour leurs dévouements tout au long de mes études.

Que ce modeste travail soit le meilleur témoin de mon respect, de la profondeur de mon amour et de mon infinie reconnaissance. Que DIEU vous préserve et vous accorde santé, bonheur et prospérité.

A ma chere femme Hanen et mes chers enfants Adem et Assil

A mes chers frères.

A mes chères sœurs

Source inépuisable de générosité et d'affection, en témoignage de mon grand amour et l'étendue de mes sentiments à votre égard

Vous n'avez jamais cessé à m'apporter le soutien moral, l'encouragement et l'aide Que DIEU sauvegarde notre solidarité et notre attachement familial.

Mes collègues

Tous mes chers amie(s)

Vous m'avez aimé, soutenu, aidé et vous avez cru en moi dans la voie que j'ai choisi, que notre amitié dure et que vous trouvez dans ce travail le témoignage de mon profond amour.

Remerciements

Je remercie DIEU, le tout puissant, le tout miséricordieux, qui m'a donné l'opportunité de mener à bien ce travail.

Aux termes de ce modeste travail, j'exprime mes sincères remerciements à tous ceux qui ont bien voulu me porter leurs concours précieux à la réalisation de cette étude.

Je tiens tout particulièrement à remercier mon directeur de thèse M. Mounir ZRIGUI, pour ses pertinentes directives, ses suggestions et ses critiques constructives en termes d'encadrement, qu'il trouve ici l'expression de ma profonde gratitude

Mes respectueux remerciements s'adressent bien vivement à mon encadrant Mr Mbarek CHARHAD qui m'a fait partager son expérience et m'a aidé à élargir mes connaissances, pour sa contribution en termes d'encadrement, sa disponibilité malgré toutes ses occupations, avec son dynamisme intellectuel et ses idées créatrices, pour son aide et ses encouragements, ainsi que l'intérêt continu qu'il a porté à mon travail et à la correction de ce manuscrit

Je remercie vivement mes rapporteurs Mr.Abelmajid BEN HAMADOU et Mme. Fadoua DRIRA qui ont accepté d'évaluer le présent travail et m'ont aidé beaucoup à améliorer la qualité de mon rapport de thèse.

Je remercie également le président de jury M. Kais HADDAR et le membre de jury Mme. Emna JAMMOUCI pour l'honneur qu'ils m'accordent en jugeant ce travail.

Par la même occasion, je voudrais exprimer mes remerciements à tous les membres du laboratoire RLANTIS qui m'ont aidé par leurs conseils et leurs critiques.

Une pensée toute spéciale à ma famille qui m'a apporté le soutien dont j'avais besoin pour mener à bien ce travail. Je remercie mes parents pour leur confiance et pour leurs encouragements. Ma gratitude s'adresse également à mes frères, à ma sœur et qui ont cru en moi.

J'adresse un très grand merci s'adresse à ma femme Hanen que je ne cesserai de remercier. Son aide, son amour et ses encouragements durant ces années de thèse m'ont été des plus précieux.

Mes derniers remerciements s'adressent à mes enfants qui m'ont offert les moments les plus agréables dont j'avais besoin tout au long de cette thèse.

En témoignage de ma reconnaissance, je vous prie de trouver, dans ce travail, l'expression de ma profonde gratitude.

Résumé

Les travaux menés dans cette thèse s'intègrent dans le cadre de l'indexation sémantique des vidéos et s'intéressent plus particulièrement aux journaux télévisés en langue arabe.
Notre contribution au niveau théorique consiste à proposer une approche d'indexation permettant de passer d'une représentation bas niveau (signal) vers une description sémantique de contenu vidéo en exploitant le texte incrusté comme une source d'information. Cette approche est composée de deux modules. Un module de traitement bas niveau introduit une phase d'extraction des images clés et une étape de détection et de reconnaissance du texte incrusté. Le deuxième module permet d'extraire les informations sémantiques selon une modélisation multi-facettes : conceptuelle, événementielle et thématique.
Au niveau expérimental, notre contribution consiste à développer le système SISAVIN. Celui-ci regroupe tous les modules nécessaires pour l'indexation sémantique des journaux télévisés, allant de l'extraction des images clés jusqu'à la génération des indexes sous forme des fichiers xml.
Les résultats obtenus lors de l'évaluation sont globalement satisfaisants pour les différentes méthodes proposées et montrent que le système SISAVIN est capable de fournir une indexation efficace et pertinente pour les journaux télévisés arabes.

Mots clés : indexation sémantique, texte incrusté, modélisation multi-facette, vidéos arabes.

Abstract

The work carried out in this thesis is part of the semantic indexing of videos and focuses more particularly on Arabic news videos. Our contribution at the theoretical level consists in proposing an indexing approach allowing to go from a low-level digital representation to a semantic description of video content by exploiting the embedded text as a source of information. This approach is composed of two modules. A low-level processing module introduces a phase of the extraction of the key frames and a step of detection and recognition of the embedded text. The second module allows the extraction of semantic information according to multi-faceted modeling : conceptual, event and theme. At the experimental level, our contribution is manifested by the development of the SISAVIN system. This includes all the modules needed for the semantic indexing of television news, ranging from the extraction of key frames to the generation of final indexes in the form of xml files. The results obtained during the evaluation are generally satisfactory for the various methods proposed and show that the SISAVIN system is capable of providing effective and relevant indexing for Arabic news videos.

Keywords : semantic indexing, embedded text, multifaceted modeling, Arabic news videos.

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

ANNIE	*A Nearly-New Information Extraction System*
ASR	*Automatic Speech Recognition*
ACL	*Association of Computational Linguistics*
ACE	*Automatic Content Extraction*
BOE	*Bag of entity*
BOW	*Bag of word*
CRF	*Conditional Random Fields*
CRR	*Character Recognition Rate*
CoNLL	*Conference on Computational Natural Language Learning*
DARPA	*Defense Advanced Research Projects Agency*
DCT	*Discrete Cosine Transform*
ESTER	*Évaluation des Systèmes de Transcription d'Émissions Radiophoniques*
EMM	*Europe Media Monitor*
GATE	*General architecture for text engineering*
HD	*Haute definition*
ICDAR	*International Conference on Document Analysis and Recognition*
KNN	*k-nearest neighbors*

LBP *local binary patterns*

MSER *Maximally stable extremal regions*

MUC *Message Understanding Conference*

OCR *Optical Character Recognition*

POS *Part of speech*

RENAR *Repérage des Entités Nommées Arabes*

REN *Reconnaissance des entités nommées*

SISAVIN *Semantic Indexing System for Arabic Video News*

SRI *Système de Recherche information*

SVM *Support vector machine*

SWT *Stroke width transform*

SD *Standard Definition*

TRECVid *TREC Video Retrieval Evaluation*

WRR *Word Recognition Rate*

I

Chapitre

1 Introduction Générale

1.1 Contexte

La multiplication des chaînes de télévision et le progrès rapide des supports de stockage ont permis d'archiver des grandes collections des journaux télévisés qui ne cessent d'augmenter chaque jour. Pour assurer un accès rapide et pertinent à ces collections, il est nécessaire de mettre en place des outils d'indexation permettant de modéliser et de représenter efficacement le contenu des vidéos.

La description du contenu d'un document vidéo à travers le processus d'indexation est une étape décisive. En effet, l'indexation se présente en amont de toute démarche du traitement des données vidéo. Elle consiste à extraire une signature numérique (bas niveau) ou sémantique (haut niveau) qui décrit le contenu d'une manière précise. Les descripteurs de bas niveau décrivent les caractéristiques physiques d'une vidéo comme la couleur, la texture et la forme. Alors que les indexes de type sémantique visent à fournir une vue symbolique de contenu sous forme de descripteurs de haut niveau.

Les textes incrustés dans les séquences vidéo font partie de ce dernier type des indexes et représentent une source d'information très riche pour les applications d'indexation sémantique, notamment pour les journaux télévisés. Cependant, la détection et la reconnaissance de ce type de texte restent encore une tâche difficile à réaliser à cause de la variabilité du texte(en style, en taille, en position) et les mauvaises conditions d'acquisition comme la faible résolution, le fond complexe, la luminosité non uniforme.

Un autre problème important pour l'efficacité du processus d'indexation est lié à l'extraction des informations sémantiques à partir des données numériques afin de fournir une description pertinente du contenu vidéo. Ceci est traduit par la question suivante : Com-

ment passer d'une représentation au niveau signal (bas niveau) vers une représentation sémantique (haut niveau) ? Autrement dit, il s'agit de minimiser l'écart (fossé sémantique) entre la réponse de la machine et l'interprétation humaine dans le cadre d'extraction des concepts sémantiques (Figure 1.1).

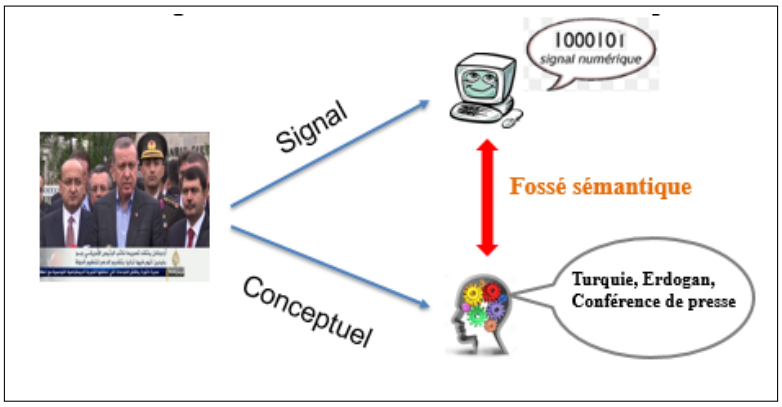

Figure 1.1: Problème de fossé sémantique

1.2 Positionnement

Nos travaux de thèse visent à concevoir et à développer un système d'indexation sémantique par la mise en œuvre d'une approche d'extraction de descripteurs de haut niveau. L'objectif est de réaliser un modèle capable de représenter efficacement le contenu d'un journal télévisé arabe en utilisant le texte incrusté comme une source d'information.

L'accent est mis sur la langue arabe pour plusieurs raisons. Premièrement, cette langue est classée à la quatrième position parmi les langues les plus parlées dans le monde (plus de 538 millions locuteurs en 2017). De plus, les archives de journaux télévisés arabes sont aujourd'hui en croissance exponentielle grâce à l'apparition de plusieurs chaînes télévisées arabes. Deuxièmement, le texte arabe dispose de plusieurs caractéristiques spécifiques qui rendent sa détection et sa reconnaissance très difficiles. Nous notons à titre d'exemple, le texte cursif, des formes plus variées que les textes latin et chinois. De même, son interprétation sémantique demeure un défi intéressant à cause de la richesse et la complexité morphologique de la langue arabe.

1.3 Organisation du mémoire

La suite de ce rapport est organisée en quatre chapitres en plus d'une introduction générale et d'une conclusion générale.

Le **chapitre 2 : Généralités** détaille le contexte général de notre travail à travers la description des caractéristiques audiovisuelles d'un document vidéo. Ainsi, nous présentons un aperçu sur l'architecture du système d'indexation et de la recherche d'information audiovisuelle en se focalisant sur ses différents modules. Nous décrivons ensuite les différentes approches proposées pour l'indexation sémantique des journaux télévisés.

Le **chapitre 3 : État de l'art**, nous passons en revue les approches rencontrées dans la littérature pour l'extraction d'information en langue arabe. Nous détaillons, en particulier, les approches de reconnaissance des entités nommées et de détection de texte dans les images. Nous enchaînons par la présentation des approches d'extraction des évènements et des approches d'extraction et de détection du texte.

Le **chapitre 4 : Contributions**, décrit nos contributions pour l'indexation sémantique des journaux télévisés en langue arabe. Nos contributions se résument dans :

— La proposition d'une méthode de détection de texte arabe incrusté dans les images issues des journaux télévisés.

— La proposition d'une modélisation multi-facette utile à extraire les informations sémantiques pour une description pertinente du contenu vidéo.

— La proposition d'une méthode de classification orientée thématique du texte court. Cette méthode reflète mieux la sémantique en utilisant les entités nommées et les évènements comme des marqueurs thématiques.

Le **chapitre 5 : Expérimentation et évaluation**, décrit la réalisation de notre système d'indexation SISAVIN ainsi qu'une étude expérimentale pour évaluer l'ensemble des méthodes proposées.

Finalement, la conclusion générale présente une synthèse de nos différentes contributions et présente un aperçu sur les futures perspectives.

Chapitre

2 Généralités

Introduction

Grâce aux progrès technologiques ainsi que la multiplication des chaînes télévisées, les collections audiovisuelles ne cessent d'augmenter y compris les journaux télévisés, les émissions sportives, les documentaires ... Cette grande quantité de données audiovisuelles représente actuellement les principaux défis des recherches menées sur le traitement automatique de l'information, notamment dans l'indexation et la recherche des vidéos. Dans ce chapitre, nous présentons les caractéristiques d'un document audiovisuel ainsi que les traitements associés(segmentation et indexation).

2.1 Notion et spécificités de document audiovisuel

2.1.1 Composition

Un document audiovisuel est défini comme une combinaison de flux d'information. Principalement, deux sources d'information composent ce document (l'image et le son). Un troisième flux généralement associé aux documents audiovisuels et présente l'information textuelle. Il provient soit d'un flux séparé, soit il est dérivé de deux sources : source audio et source visuelle (Figure 2.1).

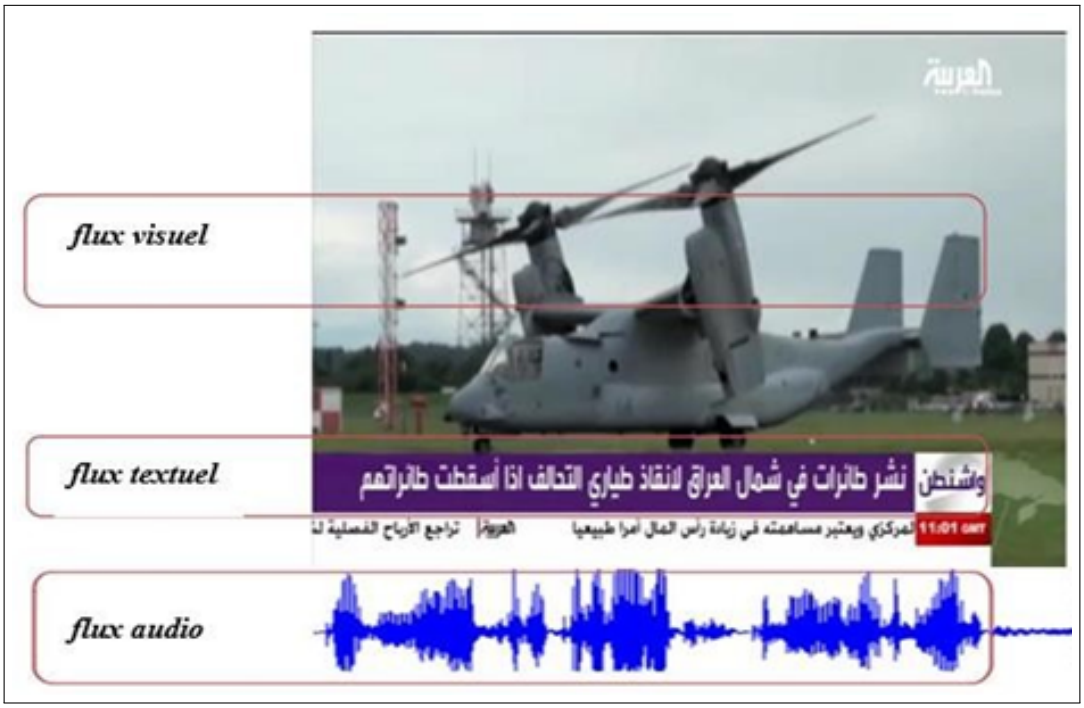

Figure 2.1: Les composants d'un document audiovisuel.

Flux visuel : Le flux visuel comporte une séquence d'images fixes qui apparaissent animées selon l'axe temporel à une fréquence de 25 à 30 images par seconde [MCh05].

Flux audio : L'audio constitue une composante importante du document audiovisuel ayant un contenu riche en information sémantique. Parmi les caractéristiques qui sont largement utilisées, nous pouvons citer le volume. Ce dernier est un très bon indicateur de silence et utile pour la segmentation audio. Le taux de passage par zéro est aussi très largement utilisé, il est notamment très efficace pour discriminer le dialogue de la musique. Typiquement, la voix est caractérisée par un faible volume et un taux important de passage par zéro. Le pitch ou la fréquence fondamentale du signal audio est un paramètre important pour l'analyse et la synthèse de la voix et de la musique. D'un point de vue du contenu sémantique, la parole est plus sollicitée que d'autres types d'informations, tels que la musique ou le bruit.

Flux textuel : Le flux textuel d'un document vidéo contient une information riche en éléments sémantiques et relativement facile à exploiter, de façon similaire aux documents textuels. Le flux textuel peut-être produit des diverses sources. Par exemple, à la télévision, certains programmes sont diffusés avec le sous-titrage qui est issu de la tech-

nologie télétexte. De même, une analyse des flux audio et visuel permet aussi d'extraire des données sous forme textuelle. Les systèmes de transcription de la parole (ASR) visent à reconnaître des phonèmes ou des mots par un processus de classification automatique depuis la modalité sonore. De même, les systèmes de reconnaissance optique de caractères (OCR) permettent d'extraire le texte présent dans l'image.

2.1.2 Structure

Il existe une variété de catégories des documents audiovisuels (journaux télévisés, documentaires, films, publicité, vidéo-surveillance, etc.). Ces documents ont une ou plusieurs structures spécifiques qui sont généralement décomposées en trois éléments : plans, scène, et séquence (suite de scènes). La Figure 2.2 illustre la structure hiérarchique d'un journal télévisé.

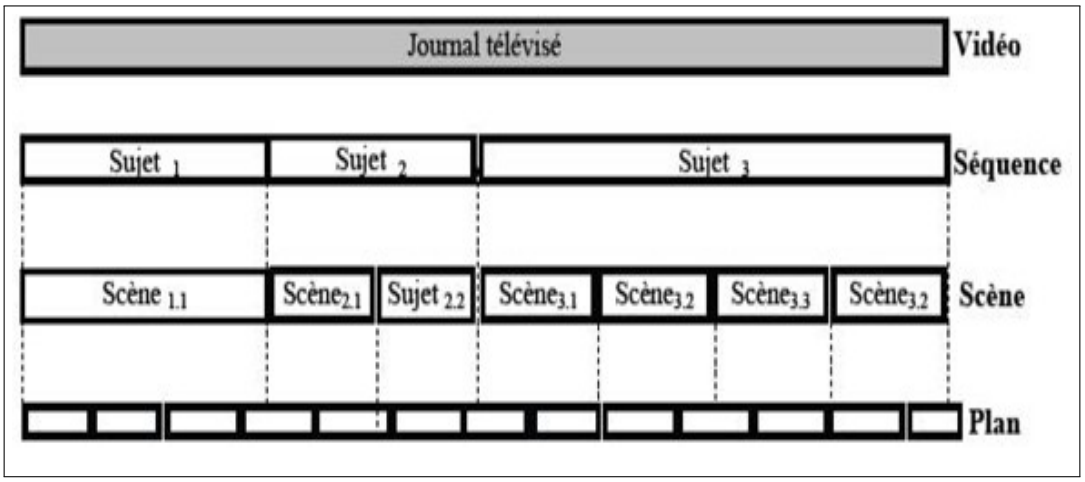

Figure 2.2: Structure du journal télévisé.

Séquence : C'est la plus grande unité que la vidéo pourrait avoir. Elle évoque généralement un thème bien déterminé. Par exemple, dans le cas d'un journal télévisé, un reportage concernant un sujet de l'actualité forme une séquence.

Scène : La scène est composée d'un ensemble de plans ayant une même unité de lieu. En pratique, la scène correspond généralement à une prise d'une ou de plusieurs caméras d'un seul contexte (lieu et personnages).

Plan : Un plan est simplement défini dans un cadre de montage vidéo à partir d'une série d'images acquises par une seule caméra. Cet ensemble d'images représente une action continue dans le temps et dans l'espace.

2.1.3 Segmentation

La segmentation des documents audiovisuels est une tâche incontournable dans le cadre d'un système d'indexation. Il existe deux types de segmentation : segmentation en plans et segmentation en scènes.

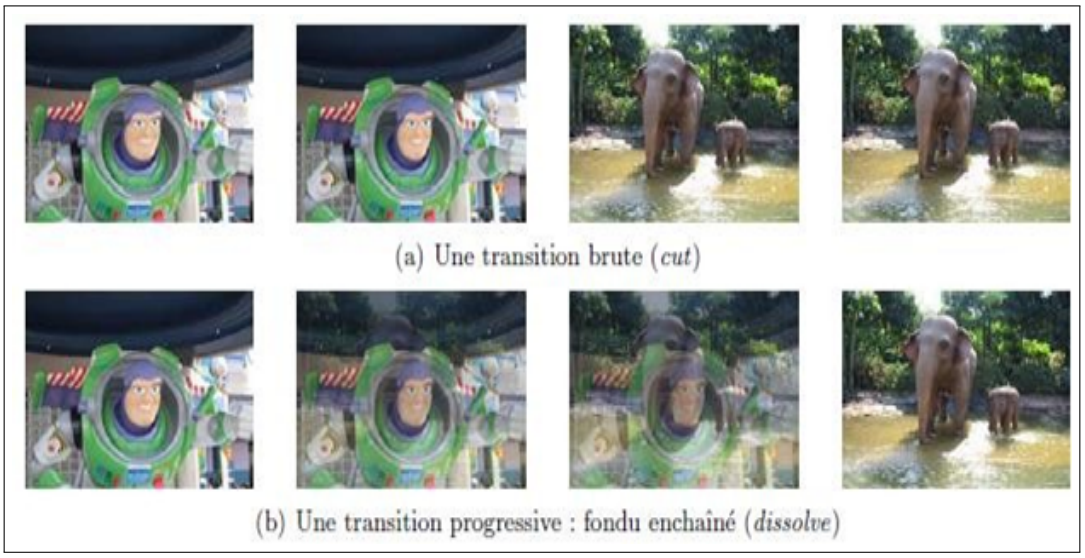

Figure 2.3: Exemples de transitions entre deux plans [Dem00].

Segmentation en plans : La segmentation en plans est la première étape à effectuer dans n'importe quel processus de traitement d'une vidéo. Il s'agit de détecter la transition d'un plan à un autre. Nous distinguons deux principaux types de transitions à savoir les cuts et les dissolves. Les cuts sont les transitions rapides entre deux plans sans l'utilisation d'effets. Ce type de transition est brusque. Il possède l'avantage qu'il est facile à détecter. Un dissolve ou fondu enchaîné est le mélange progressif de la dernière image d'un plan avec la première image du plan suivant. Ce type de transition est très difficile à détecter (Figure 2.3). Les méthodes de segmentation en plans consistent essentiellement à extraire les caractéristiques de chaque image dans la vidéo et calculer ensuite une distance (ou mesure de similarité) entre deux images successives. L'application de cette distance engendre un signal unidimensionnel, dans lequel, les pics correspondent aux frontières de plans vidéo

[ZKP03].

Segmentation en scènes : La scène est composée de plans étroitement liés (Figure 2.4). La détermination des scènes est utile pour la navigation, la visualisation et aussi pour l'analyse sémantique de contenu audiovisuel. Il existe deux catégories de méthodes de segmentation en scène : la première catégorie comprend les approches utilisant les algorithmes de regroupement. Les plans sont regroupés en fonction de leur similarité et éventuellement de contraintes temporelles [WZ04]. Un graphe des transitions est ensuite construit et permet de capturer la logique présentée dans la séquence des plans. Dans la deuxième catégorie se trouve les méthodes séquentielles qui regroupent au fur et à mesure les plans. Un ensemble des règles permet de déterminer si un plan appartient à la scène courante ou à une nouvelle scène [SW00 ; Eri+05 ; BPP08].

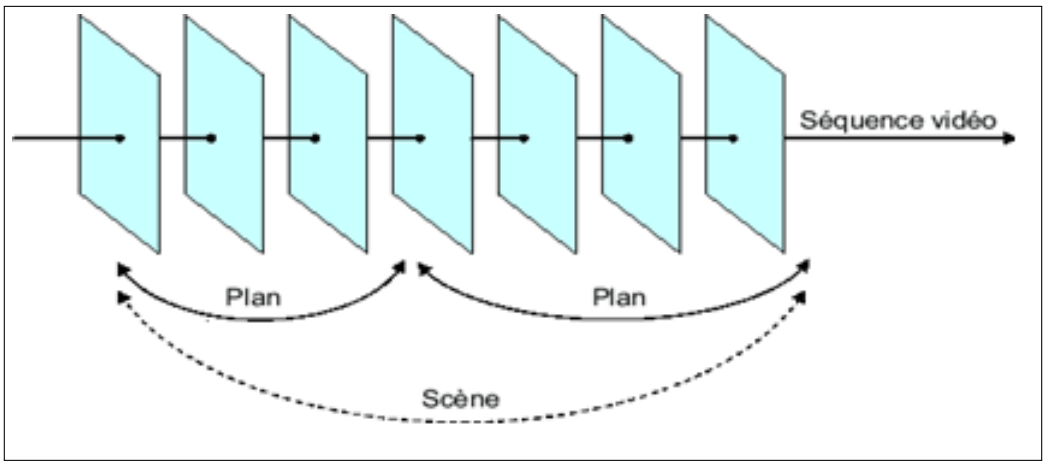

Figure 2.4: Segmentation de la vidéo en plans et en scènes.

Sélection des images représentatives : Une fois les plans sont identifiés, il devient possible de les représenter par une ou plusieurs images représentatives. Ces images ont un intérêt important au niveau de l'interaction homme machine puisqu'elles fournissent un résumé visuel d'une vidéo, qui permet à l'utilisateur d'accéder directement au plan, sans subir de surcoût temporel en navigation. En plus, elles ont un intérêt au niveau de temps d'exécution dans le processus d'indexation d'une vidéo. En effet, l'extraction d'images clés permet de réduire le nombre d'images fixes à analyser. Plusieurs chercheurs sélectionnent simplement la première, la dernière ou l'image médiane du plan [HM00]. Toutefois lorsqu'un plan contient des mouvements, il est intéressant de sélectionner les

images représentatives en fonction de l'intensité ou des variations du mouvement [FDi00].

2.2 Indexation des documents audiovisuels

La recherche d'information est un domaine des technologies de l'information qui consiste à trouver, dans une collection volumineuse de documents, l'information recherchée à partir des besoins des utilisateurs exprimés sous forme d'une requête écrite en langue naturelle. Dans un système de recherche d'information (SRI), la similarité entre les informations recherchées et le besoin de l'utilisateur correspond à un calcul de correspondance entre les supports d'information et la requête. Cette correspondance s'applique à une diversité de supports d'informations, tels que les images et les vidéos.

2.2.1 Principe

L'indexation du contenu est l'un des principaux enjeux pour la recherche des vidéos dans les collections audiovisuelles. Cette indexation peut être faite au niveau du signal (la couleur, la texture et le mouvement) ou bien au niveau sémantique. L'indexation au niveau du signal peut être utile dans certains domaines, tels que le domaine médical. Toutefois, dans les vidéos publiques, comme les journaux télévisés ou les vidéos personnelles, les utilisateurs sont souvent plus intéressés par le contenu sémantique (évènement, lieu, personnes et objets) et recherchent des séquences particulières qui répondent efficacement à leurs besoins en informations.

Dans ce sens, l'indexation sémantique vise à fournir une description symbolique du contenu qui soit proche de l'interprétation humaine. Il s'agit d'une nouvelle perspective qui est apparue récemment et qui a fait l'objet de plusieurs travaux dont l'objectif est de proposer des systèmes intelligents qui doivent imiter le processus de perception humaine afin d'extraire des informations et des interprétations sémantiques.

2.2.2 Architecture

L'architecture d'un système d'indexation et de recherche d'information audiovisuelle est composée essentiellement des modules suivants : interrogation, correspondance, ana-

lyse et indexation (Figure 2.5).

2.2.2.1 Interrogation

C'est l'expression du besoin d'information de l'utilisateur par une requête dans un formalisme propre au système. Le formalisme de spécification de la requête peut être proposé en divers types de langages d'interrogation (langage naturel, langage graphique).

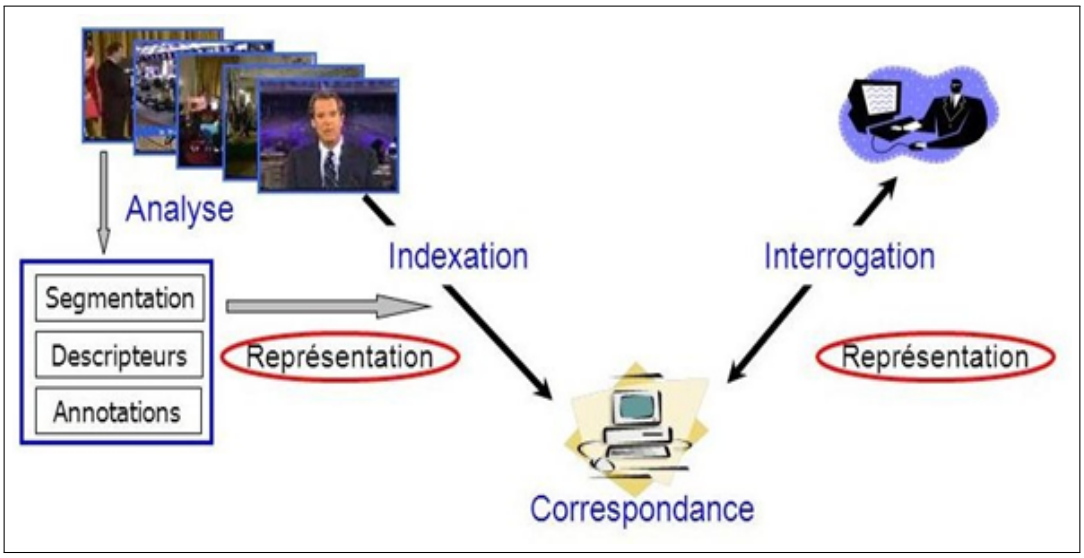

Figure 2.5: Architecture du système de recherche d'information audiovisuelle [MCh05].

2.2.2.2 Correspondance

Le processus d'appariement requête-réponse tente de retrouver les vidéos qui correspondent à la requête d'utilisateur. Il permet d'associer à chaque vidéo une valeur de pertinence vis-à-vis d'une requête. La mesure de pertinence est calculée à l'aide d'une fonction de similarité. Elle tient en compte les poids des indexes déterminés dans l'étape d'analyse et de l'indexation. Le processus de correspondance est étroitement lié aux indexes des vidéos et des requêtes.

2.2.2.3 Analyse et indexation

L'indexation vidéo consiste à extraire, représenter efficacement le contenu des bases audiovisuelles. Pour cela, les documents audiovisuels sont tout d'abord représentés par une description numérique à travers une phase d'analyse. Cette opération est réalisée en deux étapes. La première étape implique la segmentation spatio-temporelle de la vidéo

(segmentation en plan) afin de sélectionner une image représentative de chaque plan. La seconde étape permet éventuellement d'indexer le contenu des images clés à travers une sélection adéquate des descripteurs. Cette indexation peut être faite au niveau du signal ou au niveau sémantique.

Indexation au niveau du signal Ce type d'indexation se base sur l'extraction des descripteurs bas niveau (couleur, texture, forme). Plus particulièrement, la couleur est la caractéristique visuelle la plus utilisée. Elle est relativement robuste face à la complexité de l'arrière-plan et indépendante de l'orientation de l'image.

En outre, la texture se réfère aux motifs visuels homogènes et contient des informations importantes sur l'organisation structurelle des surfaces et de leurs relations avec l'environnement au alentour. La forme fournit aussi des informations pertinentes sur le contenu de l'image notamment la forme des objets.

Indexation par le contenu sémantique L'indexation sémantique des vidéos peut être réalisée selon deux méthodes : indexation par concepts visuels et indexation par entités sémantiques (lieu, personne, événement, organisation).

2.3 Indexation des journaux télévisés

Depuis plusieurs années, les chercheurs proposent des approches pour l'extraction automatique d'informations dans les journaux télévisés afin d'optimiser la manipulation des archives audiovisuelles dont la taille et le nombre augmentent régulièrement.

Les méthodes impliquées incluent la transcription automatique de la parole, l'extraction des informations sémantiques et l'identification des locuteurs. L'objectif principal consiste à analyser les archives des journaux télévisés pour répondre aux quatre questions décrivant les différentes facettes d'un sujet d'actualité : **Qui ? Où ? Quand ? Quoi ?** La suite de cette partie décrit ces différentes approches.

2.3.1 Approches basées sur l'identification des locuteurs

Savoir « qui parle », dans des larges collections des vidéos est très utile pour fournir un accès efficace à l'information dans tout type de documents audio-visuels tels que les vidéos du web et les journaux télévisés . Par conséquent, l'identification des personnes fournit des descripteurs sémantiques facilitant la recherche et la navigation dans ce type de contenu(Figure 2.6).

Pour atteindre cet objectif, la modélisation du locuteur a donné lieu à des nombreux travaux et le regroupement en locuteur lui-même est un sujet de recherche depuis longtemps. Plusieurs sources d'informations peuvent nous renseigner sur les personnes présentes dans une vidéo. On peut utiliser les méta-données liées à une vidéo, comme les tags pour une vidéo du web ou les noms dans le programme télévisé de l'émission. Toutefois, ces ressources sont souvent incomplètes et ne spécifient pas le moment d'intervention d'une personne dans une vidéo.

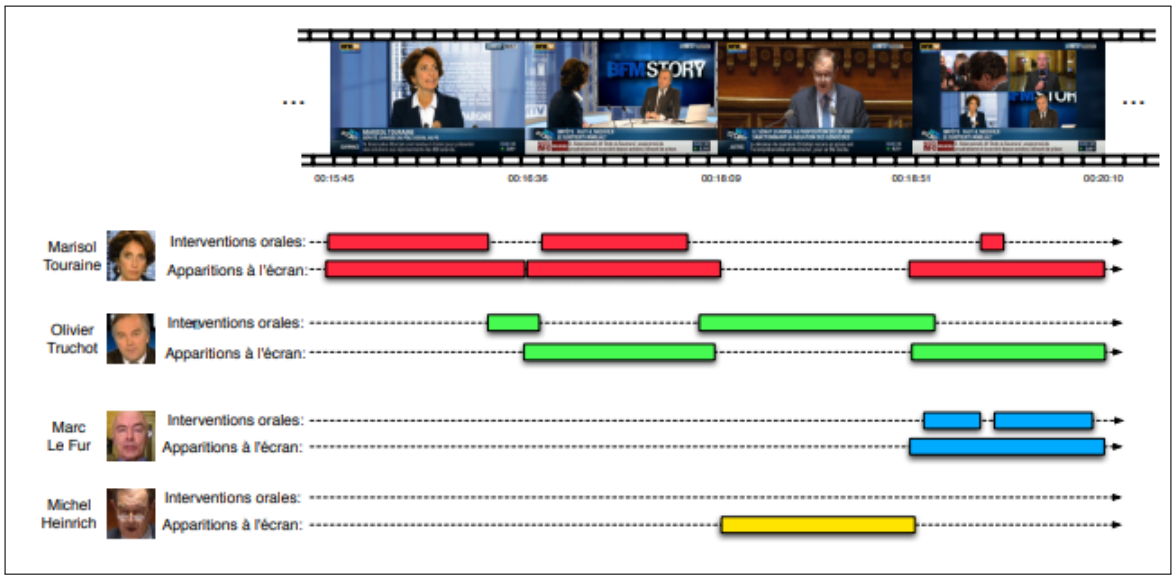

Figure 2.6: Identification des locuteurs dans les journaux télévisés [Gay15].

D'autres ressources sont intégrées dans le contenu vidéo et s'avèrent intéressantes pour déterminer l'identité de locuteurs. En premier lieu, nous citons les noms prononcés : le présentateur se charge d'introduire les invités en citant leurs noms.

Une autre source d'informations est utilisée par les émissions de télévision pour introduire une personne : les noms écrits à l'écran dans une cartouche. Une cartouche corres-

pond à l'emplacement utilisé par une émission pour écrire un nom, en vue de présenter la personne correspondante. (Figure 2.7)

Figure 2.7: Le personnage interviewé dans l'image de droite (Jean ARTHUIS) peut être identifié de deux manières : soit par la cartouche où il est écrit son nom et sa fonction , soit par la transcription du présentateur qui annonce qui va parler[Gay15].

Utilisation de la transcription Les premiers travaux ont été proposés par Canseco et al. dans [LG04]. Ils ont utilisé des patrons linguistiques définis manuellement pour déterminer l'identité de locuteur : au locuteur courant, suivant ou précédent. La Figure 2.8 décrit quelques exemples de patrons. Nous notons cinq segments de locuteurs de A à E non identifiés dans la partie haute de la figure.

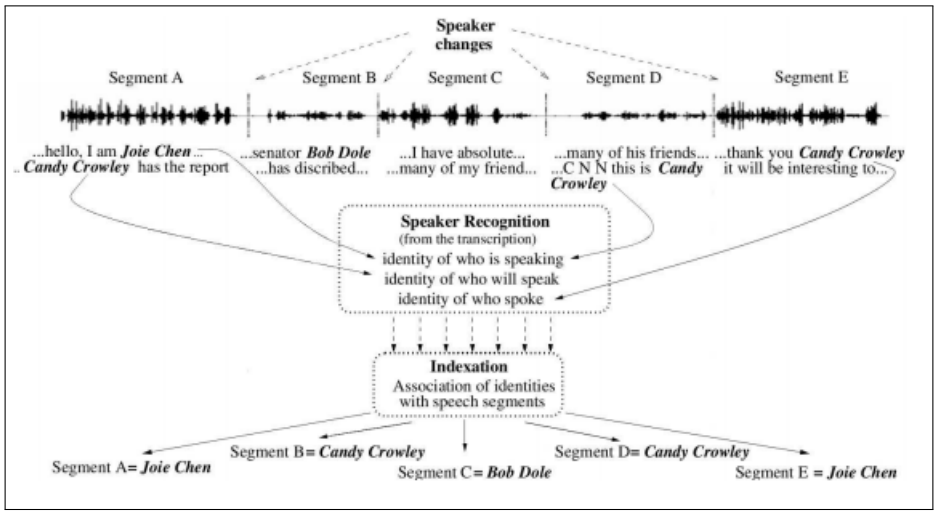

Figure 2.8: Exemple de patrons linguistiques utilisés pour le nommage des locuteurs [Cha05].

Grâce aux noms prononcés dans la transcription de la parole associée à ces segments, nous pouvons les nommer comme suit :

1 Le premier segment « I am Joie Chen » défini clairement que le locuteur courant s'appelle « Joie Chen ».

2 Le segment « This is Candy Crowley » défini que le segment D est prononcé par « Candy Crowley ». Cela est renforcé par la phrase « thank you Candy Crowley » prononcée dans le segment E.

3 Le segment B correspond à « Candy Crowley » aussi puisqu'il a été introduit par le locuteur du segment A avec la phrase « Candy Crowley has the report ».

4 Le nom du locuteur du segment C « Bob Dole » est introduit par « Candy Crowley » dans le segment B.

5 Le segment E peut-être nommé grâce au résultat de la diarization qui l'associe au segment A (« Joie Chen »).

Pour la détection des noms dans les transcriptions, les auteurs ont utilisé des dictionnaires de noms. Cependant, Ils n'ont pas pris en compte le contexte linguistique autour des noms, ce qui peut produire des erreurs de détection. A titre d'exemple, un lieu peut avoir le même nom qu'une personne).

Dans leurs travaux suivant [LG05], Canseco et al. ont a proposé de détecter le rôle du locuteur pour lui assigner un nom. Par exemple, si un tour de parole a été détecté comme provenant du présentateur, il suffit de lui donner le nom du présentateur.

Le système proposé par Charhad et al [Cha05] exploite la sortie d'un ASR et définit un ensemble de patrons linguistiques pour déterminer l'identité de locuteurs dans les journaux télévisés. Les patrons employés sont basés sur l'extraction des expressions régulières dans la transcription de la parole. Les informations liées aux locuteurs sont ensuite intégrées dans un système d'indexation sémantique de journaux télévisées pour la langue anglaise.

Dans [Cha05], les auteurs ont employé la même méthode de patrons linguistiques mais avec des systèmes automatiques et transcription automatique de la parole issue du LIMSI [JA02]). Ils ont pu identifier 53% du temps de parole, avec une précision de 82%, sur deux heures de journaux télévisés (CNN et ABC, TRECVID 2003).

En 2006, Tranter [Tra06] a remplacé les règles définies manuellement par une phase d'apprentissage de séquences de n-grammes avec des probabilités associées. Sur le corpus Hub-4, elle a montré que moins de locuteurs sont nommables avec les systèmes automatiques (47.3%) qu'avec les annotations manuelles (76.8%), sur l'ensemble de test (4,2 heures, 138 personnes). Cette réduction abaisse le temps de parole correctement identifiée de 38% à 26% avec une précision réglée à 95%.

Dans [CM07], Ma et al ont remplacé le système de calcul des règles de [JA02] par l'utilisation d'un modèle à maximum d'entropie. Ils ont, en plus, enrichi la méthode avec les informations de position du nom dans la phrase et utilisé la correspondance du genre (masculin/féminin) entre le nom et le locuteur.

Utilisation des noms écrits La plupart des systèmes d'identification de la campagne REPERE [Poi13] utilisent en priorité l'information issue des cartouches. Dans [J13], les auteurs ont construit un graphe complet (Figure 2.9) entre les tours de parole et les cartouches. Chaque arc entre deux tours de parole est pondéré par une mesure de similarité fondée sur la distance.

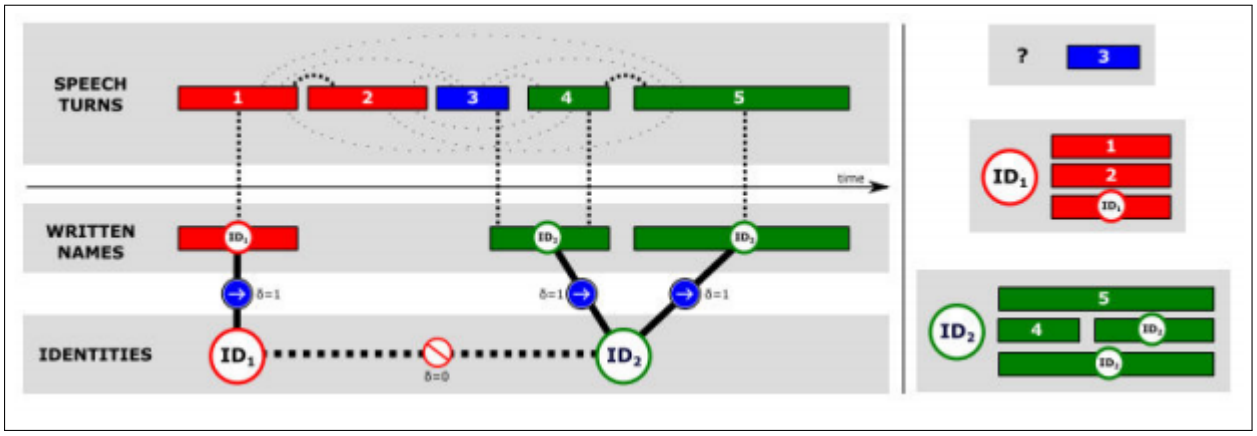

Figure 2.9: À gauche : exemple de graphe probabiliste multimodal incluant les tours de parole, les cartouches et les variables d'identités. À droite, les résultats attendus : les classes de locuteurs et les identités assignées [J13].

Les similarités entre les tours de parole et les cartouches sont calculées à partir des fréquences sur un ensemble d'apprentissage. Une variable d'identité ID est ajoutée pour chaque nom extrait des cartouches. Ensuite, l'algorithme de regroupement ILP est utilisé sur ce graphe. Des contraintes sont ajoutées dans l'algorithme pour empêcher deux

variables d'identités portant des noms différents d'apparaître dans la même classe. Ainsi le regroupement obtenu peut être utilisé directement pour le nommage car une classe n'est associée qu'à un seul nom. Ce système est évalué sur la partie TEST0 des données REPERE, il obtient une précision de 90.6% et un rappel de 58.2%.

2.3.2 Approches basées sur l'extraction d'information

Les approches linguistiques exploitent les techniques d'extraction d'information en langue naturelle pour extraire des descripteurs sémantiques tels que le nom des personnes, le lieu, l'activité, l'événement et la date dans le but de fournir une description sémantique du contenu vidéo.

Le flux textuel de la vidéo représente une source d'information très riche pour l'extraction des descripteurs linguistiques. En particulier, l'extraction des entités nommées permet de décrire les séquences vidéo par des éléments sémantiques précis tels que les personnes, organisations et lieux géographiques. En outre, certains travaux impliquent une phase d'analyse linguistique profonde pour assurer l'extraction de entités nommés et les relations qui les relient afin d'identifier les événements qui caractérisent le contenu vidéo.

Dans [RD05], les auteurs ont mis en place un système d'indexation pour gérer les archives de chaînes de télévision RAI en Italie. Ce système utilise un outil ASR pour obtenir les transcriptions textuelles de la parole. Il intègre également des techniques d'extraction d'information telle que la reconnaissance des entités nommées dans le but de fournir une description sémantique du contenu des vidéos.

Dans [YL07], les auteurs proposent une approche sémantique de détection d'événement pour les vidéos de sport. L'approche est basée sur l'analyse des textes incrustés afin de regrouper et de détecter les événements sémantiques.

Küçük et al [KDY13] proposent un système automatique pour l'annotation sémantique et la recherche des journaux télévisés en langue Turque. Ce système exploite plusieurs techniques d'extraction d'information (EI) en utilisant les textes vidéo comme une source d'information. Les techniques d'EI employées incluent la reconnaissance d'entité nommée, l'extraction d'entité de personne avec la résolution de coréférence et l'extraction

d'événement. Le système utilise les sorties de module d'extraction d'information comme des annotations sémantiques pour faciliter la recherche et la navigation dans les archives des journaux télévisés (Figure 2.10).

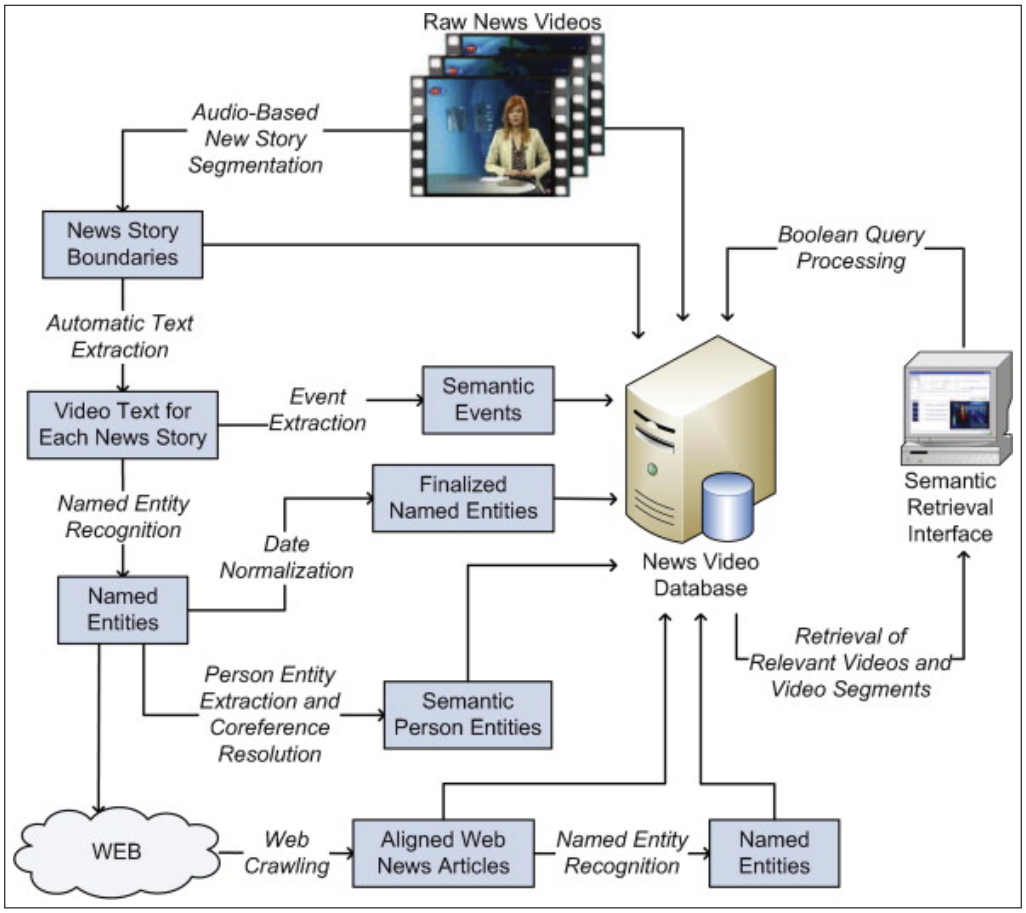

Figure 2.10: Architecture du système proposé par Küçük et al [KDY13].

2.3.3 Synthèse

Nous avons bien présenté dans la partie précédente, les différentes approches proposées dans le cadre de l'indexation des journaux télévisés. Le tableau 2.1 décrit les avantages et les limites de chaque approche.

Approches	**Identification des locuteurs**	**Extraction des informations**
Avantages	— Fournit des informations sur les personnes parlant dans la vidéo. — Facilite la recherche et l'accès au contenu de journaux télévisés.	— Fournit des informations pertinentes sur le contenu vidéo selon plusieurs facettes : nom de personnes, évènements, nom de lieux et organisations. — Répondre efficacement aux requêtes des utilisateurs.
Limites	— Les systèmes de reconnaissance de la parole engendre des erreurs dans la transcription des noms proches. Le nom peut ne pas faire référence à quelqu'un présent dans la vidéo. — Chaque émission utilise un gabarit avec des emplacements spécifiques pour écrire les textes. — La difficulté de la détection des noms écrits pour introduire la personne correspondante réside dans la détection des positions spatiales des cartouches.	— Absence des approches basées sur le texte incrusté dans la vidéo. Toutes les méthodes proposées sont basées sur l'analyse de parole. — Présence des erreurs de transcriptions. — Difficulté de sélectionner les informations pertinentes dans les cas de présence de plusieurs évènements et plusieurs participants dans le même discours.

Tableau 2.1: Avantages/Limites des approches d'indexation

Suite à cette synthèse, nous constatons que les méthodes basées sur les transcriptions

des paroles ne sont pas les plus adéquates pour l'indexation de journaux télévisés, notamment pour les langues qui sont riches en lexique et complexes en morphologie, comme la langue arabe. Pour cette raison, nos recherches s'orientent vers le texte incrusté, non seulement pour identifier les locuteurs mais aussi pour analyser le contenu textuel et extraire les informations pertinentes dans la vidéo.

Pour la langue arabe, la plupart de travaux de recherche sont proposés dans le contexte d'indexation sémantique de documents textuels [SZ13 ; Sou15 ; RZ15]. Selon notre connaissance, aucun système n'existe pour l'indexation sémantique des vidéos. Ceci constitue une motivation intéressante pour la proposition d'une nouvelle approche permettant de modéliser et indexer le contenu vidéo en langue arabe.

Conclusion

Dans ce chapitre, nous avons évoqué quelques spécificités des documents audiovisuels notamment les aspects de multimodalité et la structure hiérarchique du contenu ainsi que les approches proposées pour l'indexation sémantique de journal télévisé qui représente le cadre applicatif de notre travail de thèse. Le chapitre suivant détaille un état de l'art sur les techniques d'extraction d'information en langue arabe et les méthodes de détection de texte dans les images.

Chapitre

3 État de l'art

Introduction

Dans ce chapitre, nous mettons l'accent sur les spécificités de la langue arabe. Plus particulièrement, nous détaillons les caractéristiques géométriques et linguistiques du texte arabe[Moh17 ; Cha17], ainsi que les problèmes et les difficultés rencontrées au niveau d'extraction de l'information. Nous présentons également une analyse performante des études proposées dans la littérature pour la détection de texte dans les images.

3.1 Extraction d'information en langue Arabe

Le texte incrusté, comme les noms des personnes et les titres des évènements dans les journaux télévisés, représente une source d'information sémantique très riche pour les applications d'extraction d'information et l'indexation des documents audiovisuels. Cependant, ce type de texte, notamment en langue arabe est considéré comme un texte difficile à maîtriser et à manipuler aussi bien de mettre en œuvre dont son traitement automatique nécessite la maîtrise de ses spécificités géométriques et linguistiques[**b103** ; Moh16 ; Emn20].

3.1.1 Caractéristiques géométriques

A la différence des langues indo-européennes, les chaînes alphabétiques arabes s'écrivent de droite à gauche et elles sont composées de vingt-huit lettres de base (voir Annexe 6.2). Plusieurs lettres sont similaires par leur squelette et ne se diffèrent que par des points utilisés comme diacritiques au-dessus ou au-dessous de la ligne d'écriture (Figure 3.1. c). La plupart des lettres s'attachent entre elles et leur forme peut changer selon

leur position dans le mot (en position initiale, médiane, finale, isolées) (Figure 3.1.a). Par ailleurs, six lettres ne s'attachent jamais à la lettre suivante. Ceci provoque une segmentation du mot en une séquence des entités connexes entièrement disjointes nommées pseudo-mots, qui est à son tour formé d'un ou plusieurs caractères (3.1.b). De plus, le texte arabe est cursif, c'est-à-dire que les caractères d'un mot sont reliés entre eux par une ligne horizontale centrale appelée ligne de base. En outre, il y a des lignes qui apparaissent au-dessus et au-dessous de la ligne de base, appelées ascendantes et descendantes, comme le montre la Figure 3.1.d.

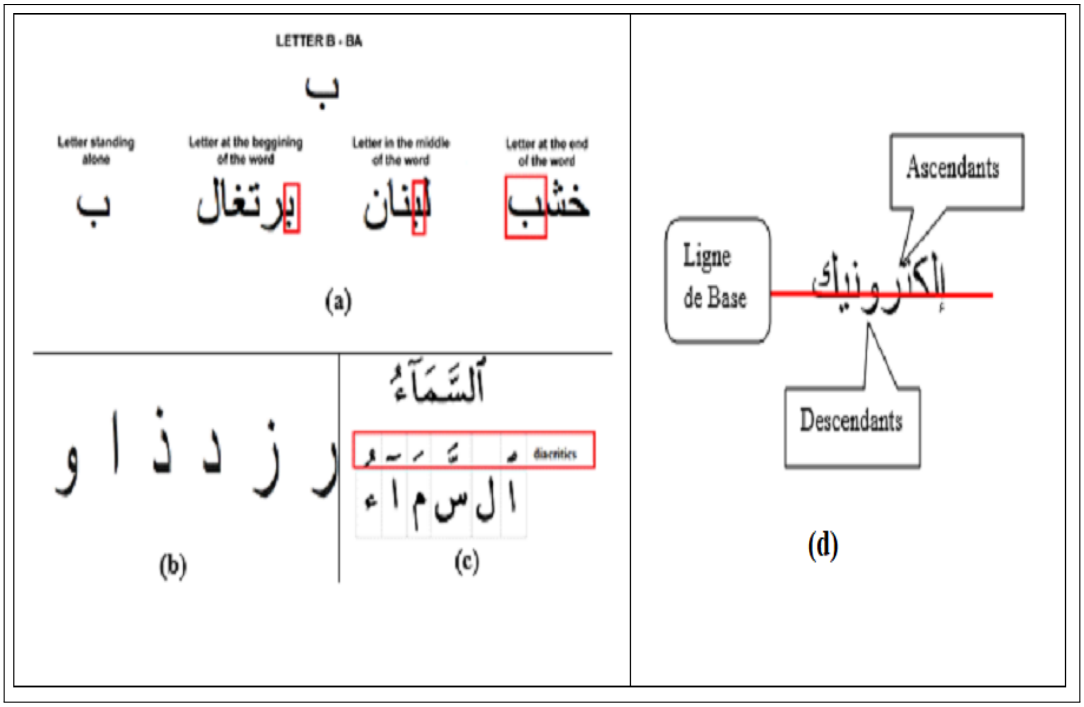

Figure 3.1: Caractéristiques géométriques de texte arabe.

3.1.2 Caractéristiques linguistiques

Les caractéristiques linguistiques sont liées essentiellement à la richesse du lexique et la complexité morphologique de la langue arabe[Adn19a ; Adn18 ; Nai15].

3.1.2.1 Lexique

Le lexique arabe comprend trois catégories de mots : les noms, les verbes et les particules détaillées comme suit :

Les noms : Ils désignent une personne ou un objet et expriment un sens indépendant du temps. Ainsi, nous distinguons trois principales catégories des noms en langue arabes [S08] :

- Les noms dérivés ou déverbaux qui peuvent être dérivés à partir d'une racine verbale, le nombre et la nature de ces formes varient selon le statut du verbe auquel il se rattache.
- Les noms primitifs qui ne peuvent pas être rattachés à une racine verbale.
- La dernière catégorie correspond aux nombres qui peuvent être des numéros simples ou composés.
- La morphologie des noms en langue arabe varie selon la fonction. En effet, il existe plusieurs fonctions. Nous citons à titre d'exemple : le lieu (endroit), le nom d'action (désigne l'action), l'objet (celui qui subit l'action), agent (qui fait l'action), etc.

Les verbes : Ce sont des entités exprimant un sens dépendant du temps. En langue arabe, la majorité des verbes sont formés par trois ou quatre consonnes.

Les particules : Les particules sont classées selon leurs fonctions dans la phrase, parmi lesquelles nous pouvons citer :

- Particules de préposition : exemple MaEa, ILA, Fi, Ka, Bi(ب،ك،في،إلى،مع).
- Particules de coordination : exemple Wa,Voma,Fa,Aaw و،ثم، ف،أو.
- Particules interrogatives : exemple Hal هل.
- Particules d'affirmation : exemple LaA, NaEam, Bala, Ajal لا، نعم ،بلا،أجل.
- Particules de négation : exemple Lame, LaA, Lane لا،لن،لم.
- Particules distinctive : exemple Aye أي.
- Particules de future : exemple Sa, Sawefa, س،سوف.

3.1.2.2 Morphologie

Un mot arabe peut être composé d'une tige à laquelle sont ajoutés des affixes et des clitiques. La tige se compose d'une racine consonantique. Les affixes flexionnels comprennent des marqueurs pour exprimer le temps, le sexe et/ou les chiffres. Les clitiques comprennent

les prépositions, conjonctions, déterminants, pronoms possessifs et les pronoms.[Cha13 ; Moh20 ; Emn17] Les clitiques attachés au début d'une tige sont appelés proclitique et ceux fixés à l'autre extrémité sont appelés enclitiques. La plupart des morphèmes arabes sont définis par trois consonnes, auxquels divers affixes peuvent être attachés pour créer un mot. De plus, un mot en langue arabe peut correspondre à plusieurs mots en anglais. Cela rend la segmentation des données textuelles arabes plus difficile que les langues latines [R16]. Le mot arabe peut donner des significations différentes en ajoutant des affixes (Préfixes, infixes ou suffixes) comme montre le Tableau 3.1.

sens	suffixe	infixe	préfixe	mot
scientifiques	يّة	***	***	علميّة
nous appris	نَا	***	***	علمنَا
sa science	ه	***	***	علمه
scientifiques	اء	***	***	علماء
enseignement	***	ي	تـ	تعليم
Sciences	***	و	***	علوم
informatif	يّة	ا	استـ	استعلاميّة

Tableau 3.1: Versions du mot et sa signification lors de l'ajout d'affixes [R16]

3.1.3 Problèmes d'extraction d'information en langue arabe

L'extraction automatique d'information consiste à analyser les textes pour fournir des informations pertinentes et structurées en vue d'une application précise. Avec la diffusion de la langue arabe sur le web, beaucoup de problèmes d'extraction de l'information liés au traitement automatique de la langue naturelle (TALN) ont apparu. Ces problèmes étaient causés par le fait que la forme de lettres change selon la position dans le mot (au début, au milieu ou à la fin du mot) et que la plupart des lettres se rattachent les unes aux autres. Mais, le problème majeur de la langue arabe réside dans sa complexité morphologique, notamment l'absence des voyelles et l'agglutination. Dans ce qui suit, nous allons détailler ces problèmes[Emn18 ; Naf20b ; Moh18].

3.1.3.1 Agglutination

Un des problèmes majeurs de la langue arabe est causé par l'agglutination car un mot peut signifier toute une phrase. La structure du mot arabe est décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique. Le Tableau 3.2 présente la structure générale d'un mot arabe complexe. La lecture du tableau se fait de droite à gauche.

Enclitique	Suffixe	Corps schématique	Préfixe	Proclitique
ها	ون	تفكر	ت	س
Pronom suffixe	Suffixe verbale exprimant le pluriel	Dérivée de la racine (ف ك ر) selon le schème تفكر	Préfixe verbale exprimant l'aspect inaccompli	Conjonction d'interrogation

Tableau 3.2: Structure morphologique du mot arabe [Lhi16]

Pour détecter la racine d'un mot, il faut connaître le schéma à partir duquel il a été dérivé et supprimer les éléments flexionnels qui ont été ajoutés. Ceci pose un problème réel pour l'extraction de l'information dans le texte arabe, notamment pour la reconnaissance des entités nommées. En effet, nous trouvons souvent des particules qui s'attachent aux radicaux en empêchant leurs détections.[Ram16]

3.1.3.2 Absence des majuscules

La notion de majuscule est absente dans l'écriture de l'arabe. Cette notion a une grande utilité lors des processus d'extraction de l'information telle que la reconnaissance des noms propres. D'ailleurs, contrairement aux langues latines, son absence pour la langue arabe nécessite l'ajout d'un dictionnaire des mots déclencheurs et des règles[**b115** ; Naf20a] grammaticales [Lhi16] permettant de repérer automatiquement les entités nommées.

3.1.3.3 Absence des voyelles

Les voyelles en langue Arabe sont des signes diacritiques ajoutés soit au-dessus ou au-dessous des lettres. Leur absence, notamment dans le texte arabe moderne, engendre de

nombreuses ambiguïtés au niveau du sens ou de la fonction syntaxique. [Adn19b ; Adn17] Par exemple, le mot **ذهب** peut-être un verbe (aller) ou un nom propre (Or). Dans le contexte de reconnaissance d'entité nommée, **قطر** peut désigner un nom de lieu ou bien un mot déclencheur d'entité nommé de type mesure. De même, le mot , **مؤسسة** réfère à deux types d'entités nommées, **مُؤَسَّسَة** pour le nom d'une organisation et **مُؤَسِسَة** pour le nom de la personne fondatrice.

3.1.3.4 Manque des ressources

Les grandes collections de documents annotés (corpus) ainsi que les dictionnaires (listes prédéfinies des entités nommées) sont d'excellentes ressources pour mettre en œuvre et tester les performances d'un système d'extraction de l'information. Malheureusement, les ressources disponibles en langue arabe ont souvent une capacité et / ou une couverture limitée [BM12]. De plus, Peu de ces corpus ont été rendus publics et librement disponibles à des fins de recherche, mais d'autres sont disponibles en vertu des accords de licence [MC09].

3.1.4 Plateformes d'extraction d'information en langue arabe

3.1.4.1 Plateforme GATE

L'Architecture générale pour le traitement de texte ou GATE est une plateforme open source écrite en JAVA et elle est très utilisée par de nombreuses communautés (scientifiques, entreprises, enseignants, étudiants) dans le domaine de traitement du langage naturel pour différentes langues. GATE dispose d'un système d'extraction d'information nommé ANNIE, lui-même est formé par un analyseur lexical, une base de toponymes (gazetteer), un analyseur syntaxique (segmentation de phrases, avec désambiguïsation) et un étiqueteur. Cela facilite le développement des systèmes REN en fournissant à l'utilisateur la possibilité de mettre en œuvre des règles grammaticales en tant que transducteur d'état fini en utilisant le langage JAPE. Un certain nombre des chercheurs ont utilisé la plateforme GATE dans leurs travaux de recherche notamment pour la reconnaissance d'entités nommées arabe, y compris [MD09 ; OS13 ; Emn16].

3.1.4.2 Plateforme NooJ

Nooj est un environnement de développement linguistique qui permet de construire et de gérer des dictionnaires et des grammaires à large couverture, afin de formaliser divers niveaux des langues naturelles : orthographe, morphologie flexionnelle et dérivationnelle, lexique de mots simples, mots composés et expressions figées, syntaxe locale et désambiguïsation, syntaxe structurelle et transformationnelle, sémantique et ontologies. NooJ est utilisé dans des applications variées du Traitement Automatique des Langues Naturelles (par ex. reconnaissance d'entités nommées [S08], compréhension automatique de la parole [Lhi16]. L'arabe est l'une des langues supportées par NooJ.

3.1.4.3 Plateforme Farasa

Farasa [1] est une boîte à outils de traitement rapide et précise pour le texte en langue arabe. Elle comprend un module de segmentation / tokenisation, un étiqueteur POS, un analyseur de dépendance et un module de reconnaissance des entités nommées. Le module de segmentation est basé sur la méthode d'apprentissage SVM. Les noyaux linéaires utilisés dans le SVM utilisent une variété de fonctions et de lexiques pour trouver les segmentations possibles d'un mot. Le module de reconnaissance des entités nommées utilise la méthode CRF et un large corpus d'apprentissage issu des différentes sources d'information comme Wikipédia et twitter.

3.2 Approches de reconnaissance des entités nommées

Les approches de reconnaissance des entités nommées sont généralement classées en trois catégories :

— L'approche linguistique, nommée aussi symbolique ou à base des règles.

— L'approche statistique ou à base d'apprentissage.

— L'approche hybride qui consiste à combiner les deux premières approches.

1. `https://farasa.qcri.org/`

Nous proposons dans la suite de cette section un survol des approches précitées tout en mentionnant les principaux travaux qui y affèrent.

3.2.1 Approche linguistique

L'approche linguistique, appelée aussi approche à base de règles, s'appuie sur la définition manuelle des grammaires formelles constituées des règles et des patrons linguistiques. Ces patrons exploitent généralement des marqueurs lexicaux et des ressources telles que les dictionnaires de noms propres. Les marqueurs lexicaux sont des mots déclencheurs qui encadrent l'entité nommée et facilitent sa reconnaissance. D'autre part, les dictionnaires de noms propres comportent une liste des noms et des prénoms les plus fréquents, des noms de localisations, et des noms d'organisations selon le domaine envisagé.

Dans la suite de cette section, nous citons quelques systèmes de reconnaissance des entités nommées arabes.

Abuleil [Abu04] a proposé un système à base de règles pour extraire les noms de personne à partir de systèmes de Question/Réponse. Il a défini un ensemble de règles en se basant sur des verbes spéciaux tel que " **أعلن،قال**" pour détecter la position de nom de la personne dans la phrase.

Mesfar [Sli07] a développé un système de reconnaissance des entités nommées (ENs) en arabe. Ce système s'appuie sur des grammaires locales implémentées dans la plateforme linguistique NooJ. L'architecture de ce système est composée par un analyseur syntaxique procédé par un analyseur morphologique. L'analyseur morphologique regroupe le maximum des informations pour les formes de mots reconnus dans le dictionnaire et permet de résoudre certains problèmes liés aux mots arabes comme l'agglutination. L'analyseur syntaxique utilise deux types des ressources linguistiques : gazetteers et les grammaires syntaxiques pour identifier les différents types des entités nommées.

Shaalan et Raza [SR09a] ont développé un système de reconnaissance des ENs à base des règles dont l'objectif est de traiter les problèmes liés à la langue arabe comme la complexité morphologique et l'ambiguïté sémantique. Ce système utilise trois ressources : les dictionnaires de noms propres, une grammaire sous forme des expressions régulières et un processus de filtrage. Ce dernier permet de corriger les résultats initiaux de processus

de reconnaissance à l'aide de métadonnées. De plus, il contribue à la désambiguïsation des entités nommées.

Zaghouani [SR09b] a présenté un système nommé RENAR (Repérage des Entités Nommées ARabes) qui a été inspiré du système EMM (Europe Media Monitor). Le système RENAR a été dédié aux textes écrits en arabe moderne et il est basé essentiellement sur un lexique et un ensemble des règles spécifiques pour chaque classe d'EN. L'architecture de ce système est fondée sur deux étapes. La première étape sert à segmenter le texte et à normaliser le phonème, alors que la deuxième étape permet de repérer les entités nommées en utilisant les dictionnaires et les expressions régulières décrites par les règles.

EN 2011, Ben Hamadou et al [ZS10] ont développé un système de reconnaissance et de traduction des ENs arabes vers le français pour le domaine du sport. L'objectif de ce système était d'une part l'identification des noms des personnes, des lieux et des organisations sportives par le biais de grammaires locales et la représentation lexicale de la phrase et d'autre part, l'intégration d'un module de traduction pour traduire les ENs extraites. L'expérimentation du système développé a été effectuée en utilisant la plateforme linguistique NooJ.

Al-Jumaily et al [AH10] ont proposé un système de reconnaissance de ENs dédié aux applications Web. Ce système a été développé en utilisant les différents dictionnaires disponibles dans la plateforme GATE . De plus, il a fourni une analyse morphologique spécifique à la langue arabe. Les expérimentations effectuées ont été réalisées en utilisant le corpus NERcorp.

Dans le même contexte, Hkiri et al [AG12] ont utilisé la plateforme GATE pour développer un système de reconnaissance des ENs arabes dans le domaine militaire. En premier lieu, les auteurs ont enrichi les dictionnaires de la plateforme GATE par une nouvelle liste de noms de personnes, de lieux et des organisations. En deuxième lieu, ils ont développé des règles linguistiques JAPE pour extraire les entités nommées.

3.2.2 Approche statistique

Les approches statistiques ou à base d'apprentissage visent à apprendre à partir de corpus annotés, les règles d'extraction d'une manière automatique. L'acquisition de ces règles se fait tout d'abord par l'application d'une méthode d'apprentissage pour entraîner le système à exploiter les différentes caractéristiques des entités nommées. Ensuite, le système d'apprentissage génère un modèle d'analyse permettant de reconnaître les entités nommées dans des nouveaux documents. Nous détaillons ci-dessous les systèmes statistiques développés dans le cadre reconnaissance des entités nommés arabes.

Nous commençons par le système ANERsys développé par Benajiba [BR09]. Ce système possède trois versions en utilisant différentes méthodes d'apprentissage. Premièrement, ANERsys [BR09] utilise l'approche du maximum d'entropie avec l'étiquetage morphosyntaxique pour reconnaître les noms propres longs. La performance globale du système en termes de précision, de rappel et de F-mesure était respectivement 70,24%, 62,08% et 65.91%. Dans le but d'améliorer la performance d'ANERsys, Benajiba et Rosso [BR07] ont utilisé un autre modèle probabiliste basé sur les champs aléatoires conditionnels (CAC) (en anglais, Conditional Random Fields (CRF). De plus, des caractéristiques propres à la langue arabe ont été aussi utilisées dans le modèle du CRF, y compris les étiquettes POS et les dictionnaires. Ce système a obtenu des meilleurs résultats par rapport à l'ancien. La performance du système en termes de précision, rappel, et F-mesure était respectivement 86,90%, 72,77%, et 79,21%. L'amélioration ne dépend pas seulement de l'utilisation du modèle CRF mais aussi de l'intégration des descripteurs spécifiques à la langue arabe.

Dans la troisième version de ANERsys, Benajiba et al [BR08] ont attribué un classificateur à chaque type d'entité nommée en utilisant les modèles CRF et la méthode machine à vecteurs de support (SVM). Ensuite, ils ont combiné tous les classificateurs pour le système global d'extraction des ENs. La meilleure performance du système en termes de F- mesure était respectivement 83,5% pour ACE 2003, 76,7% pour ACE 2004, et 81,31% pour ACE 2005.

En 2009, Benajiba a testé les trois versions d'ANERsys en utilisant trois types de modèles d'apprentissage : entropie maximale, MVS et CRF. Les meilleurs résultats ont été obtenus par la combinaison de trois modèles à la fois.

Abdul-Hamid et al [AD10] ont mis en place un système REN arabe basé sur le CRF en utilisant un ensemble de descripteurs pour reconnaître les trois types des entités nommées classiques : la personne, le lieu et l'organisation. L'ensemble de descripteurs proposés comprend : n-gramme de caractères et la longueur des mots sachant que ce système n'utilise aucune ressource lexicale externe. Les modèles n-gramme de caractères tentent de capturer les indices qui indiqueraient la présence ou l'absence d'une entité nommée.A titre d'exemple, les modèles bigramme, trigramme et 4-gramme peuvent être utilisés pour capturer l'attachement de préfixe d'un nom pour une entité nommée candidate tel que le déterminant "ال" (AL), une conjonction de coordination et un déterminant "وال" (w+al) et une conjonction de coordination, une préposition et un déterminant "وبال" (w+b+al). Ce système a été évalué en utilisant le corpus de test ANERcorp. La performance globale du système en termes de précision, rappel et F-mesure était respectivement de 89%, 74% et 81% .

3.2.3 Approche hybride

L'approche hybride consiste à combiner 1'approche à base de règles et 1'approche d'apprentissage pour l'extraction des ENs. Parmi les systèmes admettant cette approche mixte, nous citons le système proposé par Abdul-Hamid [AD10] qui repose essentiellement sur deux étapes : une étape linguistique et une étape statistique.

— L'étape d'analyse lexicale du texte consiste à identifier les listes de déclencheurs de chaque entité. C'est une étape de reconnaissance à l'aide de grammaires qui traduisent les différentes règles de reconnaissance.

— L'étape d'apprentissage sert pour l'étiquetage des séquences isolées afin d'avoir un texte bien annoté. Les résultats expérimentaux ont montré que l'intégration de l'étape d'apprentissage corrige et améliore notablement la performance de système.

Dans le même contexte, nous citons le système de Zribi et al [TZ15]. Ce système utilise une méthode d'apprentissage pour extraire automatiquement les règles permettant l'identification des entités nommées tout en tenant compte de la structure et du contexte d'apparition de chaque NE. En outre, ce système utilise un ensemble de règles extraites manuellement pour corriger et améliorer le résultat de la méthode d'apprentis-

sage. L'évaluation de ce système a montré un taux global de F-mesure égal à 79.24%.

Récemment, Hkiri [EZ17] a développé un système hybride de reconnaissance d'entités nommées arabes. L'architecture de ce système est basée sur une approche mixte combinant les méthodes à base de connaissances avec les méthodes à base d'apprentissage. Le module à base de règles est implémenté sous la plateforme linguistique GATE alors que le module d'apprentissage est fondé sur les champs aléatoires conditionnels (CRF). En comparaison avec les systèmes NERar [AH10] et le système de Benajiba [BR08], le système de Hkiri a obtenu les meilleurs résultats en terme de F-mesure qui dépasse 81%.

3.2.4 Discussion

Les approches à base de règles sont les plus adéquates lorsque la reconnaissance d'entités nommées est appliquée à des textes bien structurés dans de petits corpus. Les règles sont écrites manuellement, ce qui demande une expérience linguistique et un effort humain conséquent. De même, la reconnaissance nécessite des connaissances *a priori* concernant les caractéristiques linguistiques de la langue étudiée.

Bien que ces approches soient très précises, elles sont coûteuses à mettre en œuvre car elles nécessitent la construction manuelle des dictionnaires et des grammaires. Toutefois, les méthodes statistiques nécessitent elles aussi un grand effort de travail pour construire et annoter les corpus de référence. Cependant, ces approches restent les plus flexibles et les plus robustes pour la reconnaissance d'entités nommées dans les textes bruités.

Les approches hybrides utilisent simultanément les techniques linguistiques et les méthodes d'apprentissage. Les règles sont premièrement écrites par un expert linguistique puis elles sont corrigées automatiquement à travers un processus d'apprentissage, ceci améliore progressivement la performance du système de reconnaissance d'entités nommées.

3.3 Approches d'extraction des évènements

Les travaux développés dans le cadre d'extraction automatique d'événement (EAE) sont classés en deux catégories : les approches linguistiques et les approches statistiques.

Beaucoup de chercheurs ont travaillé sur l'EAE en anglais et d'autres langues latines comme le français, l'allemand, l'espagnol. Par contre, ce domaine de recherche est encore pauvre et trés limité en langue Arabe.

3.3.1 Approches linguistiques

Dans le contexte linguistique, nous citons le système REES [AR00] qui vise à extraire les relations et les évènements à grande échelle. Ce système est fondé sur l'utilisation de lexiques et de patrons syntaxiques pour la détection d'événement selon le model ACE [2]. Les lexiques fournissent une description syntaxique et sémantique concernant les arguments de chaque événement. Ensuite, ces informations sont modélisées par des patrons syntaxiques décrivant les différents contextes d'apparition d'un événement.

Partant du même principe, celui de l'approche symbolique, la société américaine SRA spécialisée dans le traitement de l'information [AR98] a développé un système d'extraction nommé IEE (Information Extraction Engine). Ce dernier est composé de six modules dont celui nommé "EventTag" permettant l'annotation d'événement grâce à des règles syntactico- sémantiques élaborées manuellement. Lors de la campagne d'évaluation MUC-7, ce système a obtenu les meilleurs résultats au terme de F-mesure (51%) pour la tâche Scenario Template (ST).

Également, dans ce cadre de méthodologie linguistique, nous citons le système FASTUS [AK95] qui consiste à élaborer des règles nommées FASTSPEC visant à faciliter la tâche d'extraction d'évènement. L'évaluation de ce système a montré de bonnes performances notamment lors de la participation de compagne d'évaluation MUC-6. FASTUS a obtenu une valeur de F-mesure de l'ordre de 51% contre 56% pour le meilleur des systèmes de la campagne.

Parent et al. [Par08] ont présenté également un système d'extraction des évènements en français définis selon le Model TimeML. Ce système est fondé sur un étiquetage syntaxique et des patrons syntaxiques élaborés manuellement dans le but de repérer les expressions adverbiales de localisation temporelle. Le résultat obtenu en terme de F-mesure est égal à 71,8%.

2. `https://www.ldc.upenn.edu/collaborations/past-projects/ace`

3.3.2 Approches statistiques

Parmi les systèmes basés sur l'approche statistique, nous citons le système ALICE [3] [Chi03]. Ce système intègre une analyse des dépendances syntaxiques ainsi que des chaînes de coréférence entre entités nommées. De même, il utilise quatre techniques d'apprentissages issues de l'outil Weka [4]. Les expérimentations effectuées sur les corpus de la campagne MUC-4 ont montré que le classifieur à maximum d'entropie (ALICE-ME) a obtenu le meilleur résultat en comparaison avec tous les systèmes participants.

Dans le même contexte, quelques systèmes ont été proposés dans le cadre de la compétition TempEval-2(2010) [5] selon la norme TimeML. Le système TIPSEM [HN10] est fondé sur un apprentissage artificiel par CRF et des descripteurs de différentes natures : lemmes, parties-du- discours, informations syntaxiques et informations sémantiques tirées de WordNet. Ce système a obtenu les meilleurs résultats à TempEval-2 et a été utilisé comme référence pour TempEval-3.

Arnulphy et al. [BT14] ont développé un système d'extraction d'évènements en anglais et en français. Leur système s'appuie essentiellement sur la combinaison de deux méthodes d'apprentissage CRF et KNN. En premier lieu, Le CRF est employé pour identifier toutes les classes candidates d'évènement ainsi que leurs probabilités. Ensuite, le KNN calcule la similarité entre les instances à classer et les données d'entraînement pour déterminer la classe le plus adéquate. Ce système a été testé et évalué sur le corpus français et Anglais de la campagne d'évaluation TempEval-2 et les résultats obtenus en terme de F- mesure sont respectivement 83% et 86%.

3.3.3 Approches d'extraction d'évènements arabes

En langue Arabe, la plupart de travaux proposés ont été développés dans un cadre spécifique et dehors de normes ACE et TimeML [6].

A titre d'exemple, Dridi et al [Dri14] ont proposé un système d'extraction d'évènements arabes à partir de Twitter. Ce système traite la particularité de dialecte tunisien tel que

3. AutomatedLearning-basedInformationContentExtraction
4. https://www.cs.waikato.ac.nz/ml/weka/
5. http://www.timeml.org/tempeval2/
6. http://www.timeml.org/tempeval2

les abréviations, fautes d'orthographe, mots arabes écrits avec des alphabets latins et des chiffres, etc. La première étape consiste à regrouper les termes similaires discutant le même sujet. Ensuite, les auteurs ont utilisé les clusters obtenus pour déterminer les sujets saillants correspondants aux évènements à l'aide des méthodes statistiques. Les résultats obtenus ont montré la robustesse de système développé vis-à-vis l'hétérogénéité et la complexité des données traitées.

En 2015, Hkiri et al [EZ16] ont développé un système d'extraction des évènements dans le domaine militaire. Leur système repose sur l'utilisation des lexiques et des règles implémentées sous la plateforme linguistique GATE. L'évaluation de ce système relève un taux de rappel 35% et de précision 70%.

3.4 Méthodes d'extraction/détection du texte

Nous distinguons deux types de texte dans l'image : texte de scène et texte incrusté. Le texte de scène apparaît dans la scène capturée par la caméra. Il constitue une partie de l'image. Nous pouvons citer, à titre d'exemple, les logos dans une émission d'un match de football, le numéro sur le T-shirt des joueurs et les panneaux d'affichage. Il existe plusieurs applications pour l'extraction de ce type de texte comme la détection des plaques d'immatriculation des voitures [ZX04], l'identification des logos [FC03] et la détection de blocs d'adresse [WC04].

L'extraction de texte de scène est un sujet important dans le domaine de traitement des images notamment en présence de plusieurs facteurs tels que : fond arbitraire, couleurs non uniformes. De même, les textes peuvent être déformés à cause de de la transformation affine, du mouvement, ou superposé sur des dessins [XH14 ; MP07] (Figure 3.2).

Figure 3.2: Texte de scène

Figure 3.3: Texte incrusté

Le texte incrusté est un texte inséré après l'acquisition de l'image. Ce type de texte est ajouté dans une phase de pré-production de la séquence vidéo pour enrichir le contenu visuel. Il est étroitement lié au contenu sémantique de la vidéo. Les noms des personnes et les événements constituent des exemples du texte incrusté (Figure 3.3). Généralement, les caractéristiques de ce type de texte sont fortement liées aux contraintes de visibilité. Ces caractéristiques concernent les aspects suivants :

— **taille** : vue l'importance de l'information qu'il fournit, le texte artificiel apparaît toujours avec une taille remarquable par les spectateurs. La taille de la zone de texte est mesurée par rapport aux dimensions de l'image.

— **Alignement et orientation** : le texte incrusté est en général aligné horizontalement. Cependant, certains effets spéciaux de mise en page peuvent engendrer des distorsions dans l'alignement.

— **Intensité et couleur** : la couleur est un fort dispositif dans l'usage de l'indexation visuelle de l'information. Les caractères du texte incrusté tendent à avoir toujours une couleur et une intensité uniformes. Dans le cas où la couleur varie à travers la légende, elle varie d'une manière progressive de sorte que les caractères adjacents ou les segments de caractères aient des couleurs très semblables.

— **Mouvement** : il est souvent uniforme (horizontal ou vertical) pour le texte incrusté, tandis que, pour le texte de scène, il peut être totalement arbitraire.

— **Contour** : la plupart des textes incrustés sont définis pour être facilement lisibles. Dans ce but, le contraste avec le fond est donc souvent très élevé et les zones de texte comportent généralement des nombreux contours réguliers grâce à l'unicité du style d'écriture de l'ensemble des caractères.

La détection est l'étape la plus délicate. Elle consiste à déterminer les zones de l'image qui contiennent du texte. A cet effet, de nombreux travaux ont abordé le problème de la détection de texte dans la vidéo. Ces travaux tendent souvent à trouver des descripteurs spécifiques de texte qui les caractérisent par rapport à l'arrière-plan comme les contours, la couleur, les caractéristiques de texture et la géométrie. Ces descripteurs sont utilisés de différentes manières. Ils peuvent être utilisés par apprentissage automatique ou à travers des règles heuristiques qui peuvent être appliquées à ces descripteurs de manière à trouver un modèle plus représentatif des régions du texte. Ils sont aussi utilisés par des techniques hybrides qui combinent les méthodes d'apprentissage automatique et les règles heuristiques.

Dans cette section, nous examinons les méthodes de détection de texte proposées pour le contenu vidéo.

3.4.1 Approches heuristiques

Les approches heuristiques utilisent des descripteurs visuels (couleur, texture, forme) pour identifier toutes les régions candidates de texte dans l'image. Elles appliquent des

techniques de filtrage basées sur des règles heuristiques (rapport hauteur/largeur, taille minimale...) définies manuellement afin de supprimer les régions non texte.

Dans ce contexte, Anthimopoulos [MP07] a créé une carte de contour avec le détecteur de contour Canny. Les régions candidates de texte sont ensuite construites en appliquant des opérations morphologiques (dilatation et ouverture), comme présenté dans la Figure 3.4.

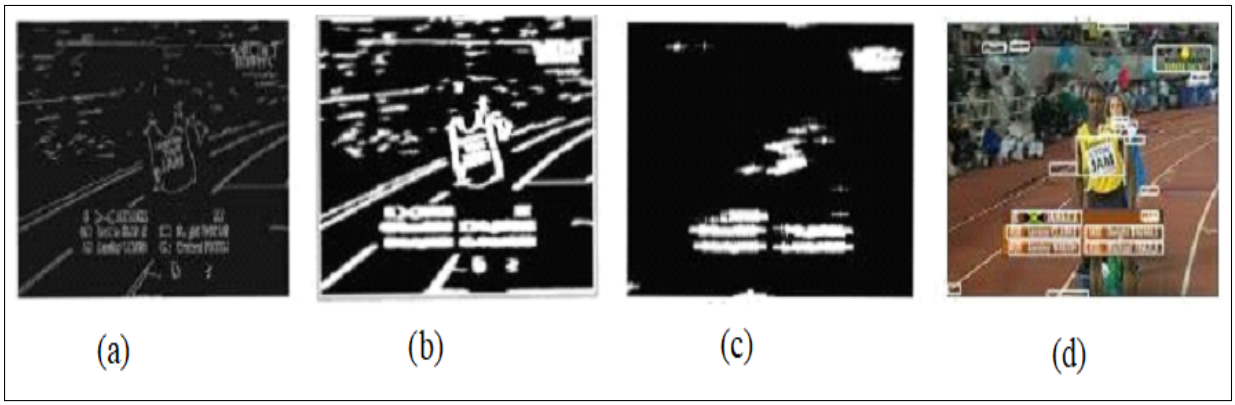

Figure 3.4: Méthode d' Anthimopoulos [MP07] : (a) Carte de contour (b) dilatation (c) ouverture (d) régions candidates

Pour augmenter la précision de la détection, des projections horizontales et verticales de contour sont effectuées sur chaque région afin de supprimer les régions non texte. La même méthodologie a été appliquée dans le travail de Moradi [MO10] pour la détection du texte arabe dans les vidéos.

Epshtein et al. [BW10] ont proposé un nouveau descripteur de texte appelé la transformation de largeur de trait SWT qui quantifie la largeur de trait pour chaque composant dans l'image. Ce descripteur permet de faciliter la détection du texte puisque ce dernier est caractérisé par une largeur de trait fixe. Cette méthode a été appliquée à la détection de texte de scène et elle a fourni d'excellents résultats dans la compagne d'évaluation ICDAR [7] 2003 et 2005 . Cependant, cet opérateur est très sensible à la variation d'éclairage et à la présence élevé de bruit.

D'autres travaux sont basés sur les caractéristiques de la couleur afin de distinguer entre les régions de texte et l'arrière-plan [WK03 ; YT11]. Dans [KK09], une analyse structurelle est effectuée tout d'abord sur les caractères de texte dans le but d'extraire les

7. TheInternationalConferenceonDocumentAnalysisandRecognition(ICDAR).

composantes connectées (CC). Ensuite, des règles heuristiques de regroupement ont été appliquées pour fusionner les CC et former les régions du texte.

D'autres approches heuristiques considèrent les caractéristiques de texture pour caractériser les régions de texte et former des CC. Dans [DK03], des règles heuristiques basées sur différentes descripteurs LBP ont été utilisées pour affiner et vérifier les régions de texte initialement sélectionnées en fonction des caractéristiques de contour. Dans [YL00 ; KH99], les coefficients DCT ont été intégrés en tant que descripteurs textuels dans les méthodes heuristiques. Bassem et al [B B05 ; B B08] proposent une approche de detection de texte basée sur deux étapes : une étape d'extraction des régions potentielles de texte en utilisant la technique quadtree. La deuxième étape consiste à supprimer les fausses détections en employant trois critères de filtrage : le contraste, la géométrie et la persistance temporelle du texte.

3.4.2 Approches statistiques

De nombreuses méthodes basées sur l'apprentissage automatique ont été proposées pour la détection de texte. Ils visent à apprendre des caractéristiques discriminantes à partir d'un corpus d'apprentissage annoté afin de construire un classificateur texte / non-texte. Pour la localisation du texte, ces approches utilisent généralement une technique de fenêtre glissante. Le classificateur scanne l'image à différentes positions et échelles en produisant des régions candidates. Ensuite, un algorithme de regroupement est appliqué pour produire des régions de texte.

Kim et al [KKi96] a appliqué le réseau de neurone multicouche pour détecter+ les zones de texte dans un journal télévisé. La couche d'entrée reçoit les valeurs de niveau de gris de chaque pixel à positions prédéfinies à l'intérieur d'une fenêtre MxM. La sortie du réseau génère deux valeurs : si la première valeur est supérieure à la deuxième valeur, le pixel au centre est classé en classe texte, sinon il est classé en classe non-texte.

Delakis et Garcia [DG08] ont proposé un système de détection de texte horizontale basé sur le réseau de neurones à convolution (ConvNet). Un pipeline de couches de convolution capture les caractéristiques textuelles et les transmet à un perceptron multicouche pour la classification binaire (texte / non-texte). La méthode proposée a fourni les

meilleurs résultats par rapport aux autres méthodes testées dans le cadre de la compagne de d'évaluation de ICDAR'03.

Une autre méthode de Li et al [HK00] est basée sur un Perceptron multicouche qui apprend les moments moyens, de second et de troisième ordre de la décomposition de l'image en ondelettes. Une technique de fenêtre glissante est ensuite appliquée sur les trames vidéo où les opérations de décomposition et d'extraction des caractéristiques sont effectuées pour classer les régions de trames.

Pour le texte arabe, Alimi et al [Ali12] a proposé un système de détection et de localisation basé sur le réseau de neurone. Pour un ensemble des pixels E, le système génère un vecteur de 10 caractéristiques en utilisant cinq fonctions issues de l'image des ;,nbvcxw¡ contours et cinq coefficients issus de l'image en TSV (Teinte Saturation Valeur). Les vecteurs générés sont présentés sous forme de propagation de réseaux de neurones contenant 10 entrées nœuds, 3 nœuds cachés et 1 nœud de sortie. Une étape de classification est ensuite appliquée à deux niveaux : au niveau des lignes et au niveau des colonnes. L'étape de fusion est enfin effectuée pour générer une image contenant les zones de texte. Les autres zones (classe non texte) sont éliminées.

En 2014, Yousfi et al [SG14] a proposé deux types d'approches basées sur l'apprentissage artificiel pour la détection de texte en langue arabe dans un journal télévisé. La première est fondée sur un réseau de neurones convolutifs (ConvNet) composé de six couches. Les premières quatre couches forment une alternance de couches de convolution et de sous échantillonnage effectuant l'extraction des caractéristiques. Les deux dernières couches forment un simple perceptron multicouche pour la classification. La deuxième approche est basée sur la technique de Boosting. Les caractéristiques Multi-Block Local Binary Pattern (MBLBP) ont été utilisées dans le but de déterminer les descripteurs les plus pertinents pour une meilleure distinction entre les zones de texte et de non-texte. Les résultats obtenus montrent que la méthode à base de ConvNet surpasse les autres méthodes de Boosting notamment en terme de taux de rejet de fausses alarmes.

3.4.3 Approches hybrides

Ce type de méthodes combine les approches heuristiques et les approches à base d'apprentissage automatique. En général, une détection des régions candidates du texte est d'abord effectuée en utilisant des règles heuristiques. Ensuite, un processus de filtrage est mis en œuvre pour rejeter les fausses alarmes en utilisant des techniques basées sur l'apprentissage artificiel.

Anthimopoulos et al [MP10] a proposé un schéma en deux étapes pour la détection du texte dans la vidéo. Premièrement, les régions de texte sont déterminées à l'aide de détecteur de contour et de certaines règles heuristiques (dilatation, lissage, projections, etc.). Ensuite, les résultats obtenus sont affinés par une classification SVM basée sur les modèles binaires locaux de contour (eLBP).

Dans [WT14], une méthode hybride a été proposée pour la détection de texte de scène en combinant la méthode des régions extrêmes à stabilité maximale (MSER) [NH08] et les réseaux de neurones ConvNets. L'idée consiste à appliquer tout d'abord le détecteur MSER sur l'image d'entrée. Un classificateur ConvNet est ensuite utilisé pour affecter une valeur de confiance à chaque composant MSER. Un seuil sur les valeurs de confiance résultantes donne les régions de texte finales. Cette méthode a été évaluée sur l'ensemble des données de référence d'ICDAR 2011 et a surpassé les méthodes existantes.

En 2016, Zayene et al [OH16] a développé un système hybride de détection de texte en langue arabe en utilisant l'opérateur SWT et une technique d'apprentissage non supervisé nommée, auto-encoder convolutionel. Tout d'abord, les régions candidates de texte sont extraites en appliquant l'opérateur SWT et certaines règles heuristiques. Ensuite, ce système utilise des auto-encodeurs convolutionnels pour produire automatiquement des caractéristiques textuelles qui ont été apprises à partir des régions étiquetées. Ces caractéristiques sont utilisées par la suite dans la phase de classification.

3.4.4 Discussion

Les approches basées sur l'heuristique remontent aux premières tentatives de la détection du texte dans les vidéos. En effet, elles sont très adéquates pour le texte bien structuré qui possède des caractéristiques constantes, comme le texte incrusté dans la

vidéo d'un journal télévisé. Bien que, ces méthodes sont simples à implémenter et facile à paramétrer avec un temps d'exécution rapide, elles sont très sensibles à la complexité de l'arrière-plan et à la qualité de l'image.

Les méthodes basées sur l'apprentissage automatique ont démontré une bonne capacité de discrimination et de généralisation de descripteurs textuels. Cependant, étant donné qu'elles sont étroitement liées à la technique des fenêtres glissantes qui scanne la totalité de l'image, elles nécessitent beaucoup de temps dans la phase d'apprentissage et de classification.

Conclusion

Dans ce chapitre, nous avons effectué un survol des approches d'extraction d'information en langue arabe ainsi que les techniques de détection de texte dans les images. Dans le chapitre suivant, nous allons nous intéresser à la mise en œuvre de nos contributions relatives à l'indexation sémantique des journaux télévisés arabes.

Chapitre

4 Contributions

Introduction

Le survol de l'état de l'art mené dans le deuxième chapitre nous a orienté vers l'élaboration d'une approche sémantique pour l'indexation de la vidéo en langue arabe. Ce choix a été justifié par l'importance et la richesse de l'information textuelle incrustée dans la vidéo ainsi que la performance des techniques d'extraction des connaissances basées sur les approches linguistiques.

Dans ce contexte, notre contribution consiste à mettre en œuvre une approche d'extraction automatique des descripteurs de haut niveau. Cette approche permet la localisation et la reconnaissance des textes incrustés dans un journal télévisé. Elle permet aussi l'interprétation linguistique de ces textes dans le but d'extraire des informations sémantiques selon une modélisation multi- facette de contenu vidéo.

4.1 Contribution 1 : Extraction du texte de la vidéo

Suite à notre étude menée sur le texte incrusté dans la vidéo, nous constatons que ce type de texte possède des caractéristiques spécifiques qui le distinguent par rapport aux autres objets dans l'image. Ces caractéristiques sont liées à la stabilité de la couleur et de la densité de contours élevée ainsi que les formes géométriques de texte incrusté. Par conséquent, nos recherches s'orientent vers les approches heuristiques dans le but de développer un système capable de détecter et de localiser les zones de texte d'une manière efficace et rapide vis-à-vis à la taille de corpus des vidéos à indexer.

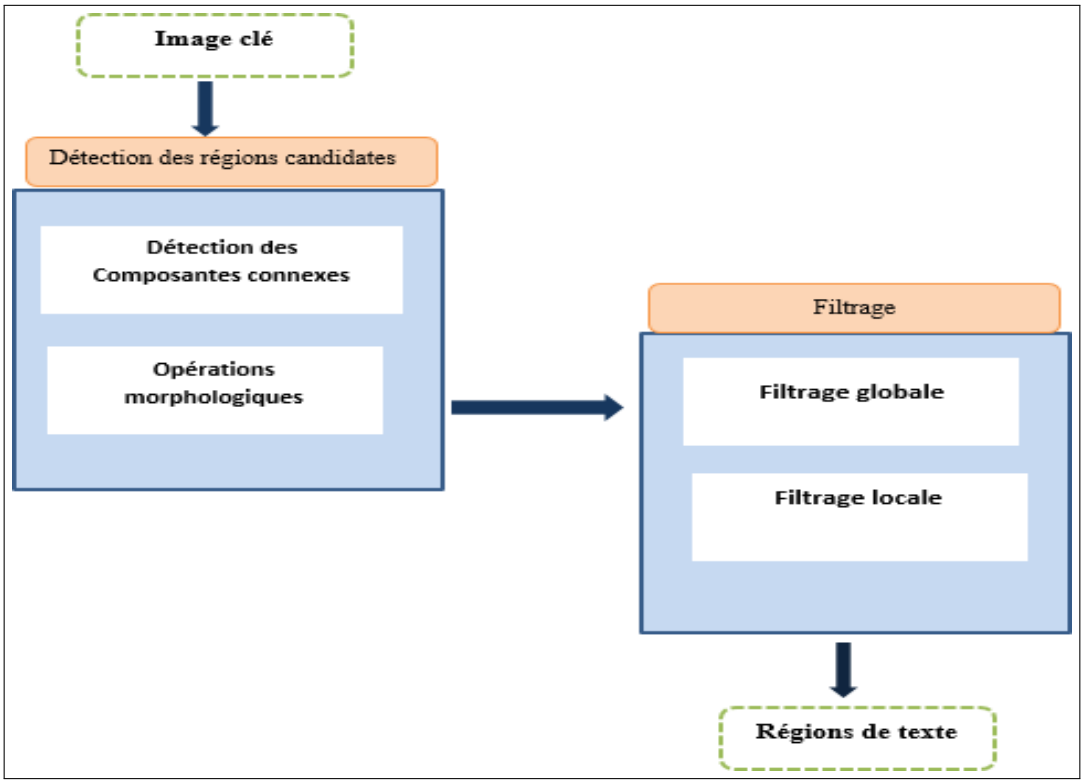

Figure 4.1: Architecture de la méthode proposée.

L'architecture de l'approche proposée est présentée dans la Figure 4.1. Elle est composée essentiellement de deux étapes : Une étape de détection des régions candidates et une étape de filtrage. La première étape sert à déterminer les régions candidates de texte en utilisant la méthode MSER ainsi que des opérateurs morphologiques (ouverture et fermeture). La deuxième étape vise à sélectionner seulement les régions correspondantes au texte en utilisant des règles heuristiques liées aux caractéristiques spécifiques de texte en langue arabe. Ces différentes étapes seront détaillées dans la suite de cette section.

4.1.1 Détection des régions candidates de texte

L'objectif de cette étape consiste à sélectionner toutes les régions candidates qui sont susceptibles de contenir le texte. Elle est composée de deux sous étapes : La détection des composantes connexes et l'application des opérations morphologiques. Ces étapes sont détaillées dans **l'algorithme 1** . ¨

Algorithm 1: détection des régions candidates

Input : D={ICi .....ICm } : liste des images clés d'une vidéo
Output : R={RCi .....RCm } : liste des régions candidates
CC =liste des régions extrêmes
OC= liste des régions candidates après ouverture
FC = liste finale des régions candidates
MaxVariation=0.25
RegionArea_min=100
RegionArea_max =2000
Δ=10
B_1= élement structurant de rayon N_1
B_2= élement structurant de rayon N_2
for each IC $\in$ *D do*
  CC= MSER(IC,MaxVariation,RegionArea_min,RegionArea_max,Δ)
endfor
for each IC $\in$ *D do*
    OC= fermeture(IC,B_1)
    FC= ouverture(FC,B_2)
endfor
retourner FC

4.1.1.1 Détection des composantes connexes

Pour la détection des composantes connexes dans une image, il existe principalement deux détecteurs : SWT (Stroke Width Transform) et MSER (Maximally Stable Extremal Regions).

La méthode SWT [BW10] considère que l'épaisseur des caractères de texte possède des largeurs uniformes. Pour ce faire, la première étape consiste à déterminer la distance séparant les deux gradients opposés tout au long du contour. La deuxième étape consiste à sélectionner les composantes connexes qui ont des faibles variations et les considèrent comme des caractères candidats de texte. Cependant, cette méthode engendre un problème de fragmentation pour le texte arabe : un même caractère est décomposé en un ensemble des composantes connexes vue l'épaisseur du texte arabe qui n'est pas stable (Figure 4.2).

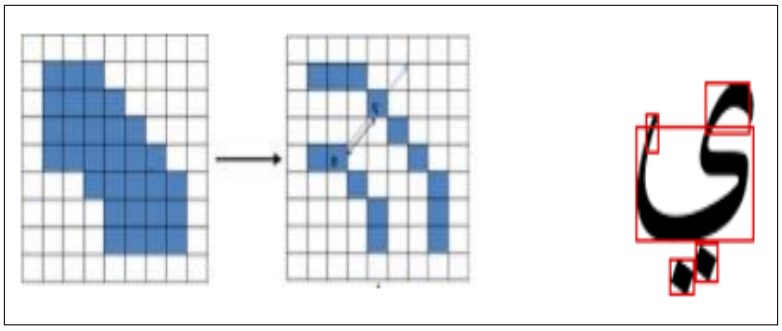

Figure 4.2: Gauche :Méthode SWT. Droite :Problème de fragmentation [OH15]

La méthode MSER a été proposée par Matas et al [JP02] comme une méthode de détection des régions où chaque région MSER est caractérisée par une intensité différente par rapport au fond qui l'entoure. Cette méthode a prouvé son efficacité en comparaison avec les autres descripteurs existants en termes de robustesse et de performance pour la détection de texte comme étant des composantes homogènes et connexes.

Pour toutes ces raisons, nous avons choisi d'utiliser la méthode MSER pour la détection des composantes connexes. Chaque région MSER est caractérisée par la stabilité de niveau de gris vis-à-vis le changement des seuils suivant le paramètre delta Δ. L'idée de base consiste à construire un arbre des composantes connexes où chaque niveau de la hiérarchie correspond à un seuil de niveau de gris Δ et regroupe un ensemble des régions nommées régions extrêmes contrôlés par le paramètre RA (RegionArea).Les MSER sont ensuite identifiées par une analyse de stabilité de chaque région dans l'arbre C à travers le paramètre MAV (MaxAreaVariation).

Pour le calcul de ces paramètres, nous avons testé plusieurs valeurs sur un ensemble des images. Plus particulièrement, nous avons choisi ceux qui donnent les meilleurs résultats. Pour le paramètre RA, nous avons choisi les valeurs *min–max* (150-2000). Ainsi que pour le paramètre MAV, nous avons opté la valeur 0.25. La figure 4.3 montre les résultats de variation de paramètre Δ.

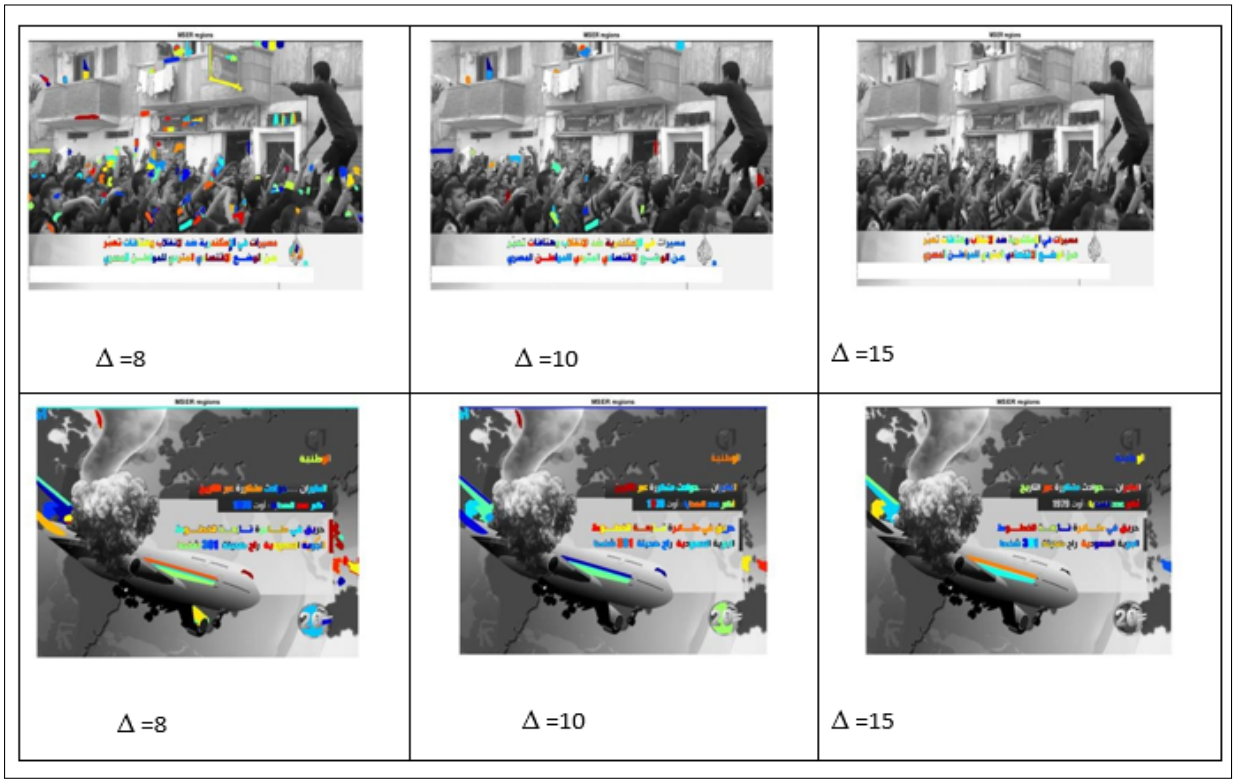

Figure 4.3: les paramètres de la méthode MSER.

Nous remarquons que pour la valeur de Δ=15 quelques caractères commencent à disparaitre , notamment pour les images à moyenne résolution. Pour cela, nous choisissons la valeur 10 car il permet de détecter tous les caractères avec un minimum de fausse détection par rapport à valeur Δ= 8.

4.1.1.2 Prétraitement par opérations morphologiques

Une fois les composantes connexes sont bien extraites, nous appliquons une méthode de regroupement basée sur les opérateurs morphologiques pour fusionner ces composantes dans une seule région. En premier lieu, nous utilisons l'opérateur morphologique (fermeture) et un élément structurant B de type disque et de rayon N (voir annexe 2).

La fermeture de l'image X par B est la composition de la dilatation suivie de l'érosion par B.

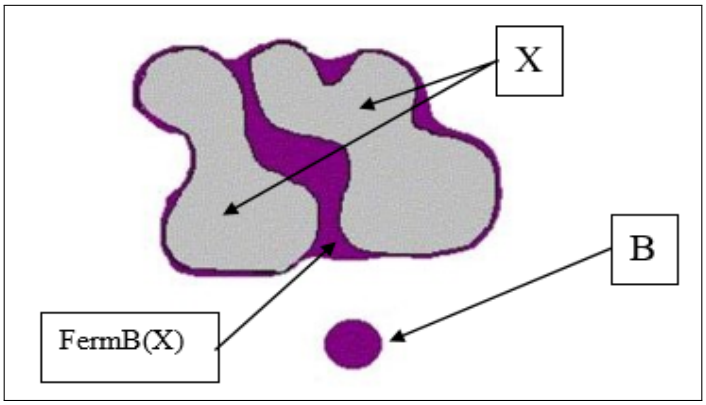

Figure 4.4: Fermeture.

Lorsqu'on effectue une fermeture avec un élément structurant en forme de disque de rayon N, les contours sont lissés, les régions fines sont fusionnées et les petits trous sont éliminés (Figure 5.9). L

Suite à nos expérimentations, nous remarquons que certaines régions de texte sont liées aux autres objets dans l'image. Pour cela, nous choisissons d'appliquer l'opérateur morphologique ouverture pour supprimer ces fausses liaisons. En effet, l'ouverture est définie comme une composition de l'érosion suivie par une dilatation avec le même élément structurant B (voir annexe 2).

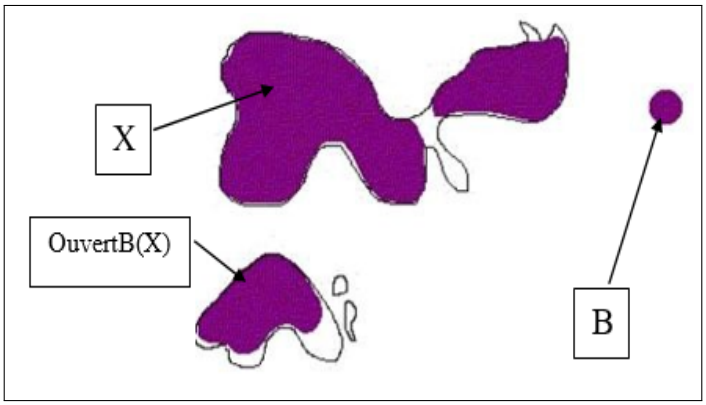

Figure 4.5: Ouverture.

L'ouverture permet de lisser le contour et d'éliminer les petites régions qui sont liées avec de petites liaisons. Pour ce faire, nous avons utilisé un élément structurant de la forme disque et de rayon égal à 5 pixels.
Les Figures (4.6.c) et (4.6.d) illustrent respectivement les résultats de fermeture et d'ouverture. Nous remarquons que tous les mots de texte sont bien regroupés en une seule région. De même, les fausses liaisons (encadrées en rouge), avec les autres objets de l'image,

sont supprimées grâce à l'opérateur ouverture.

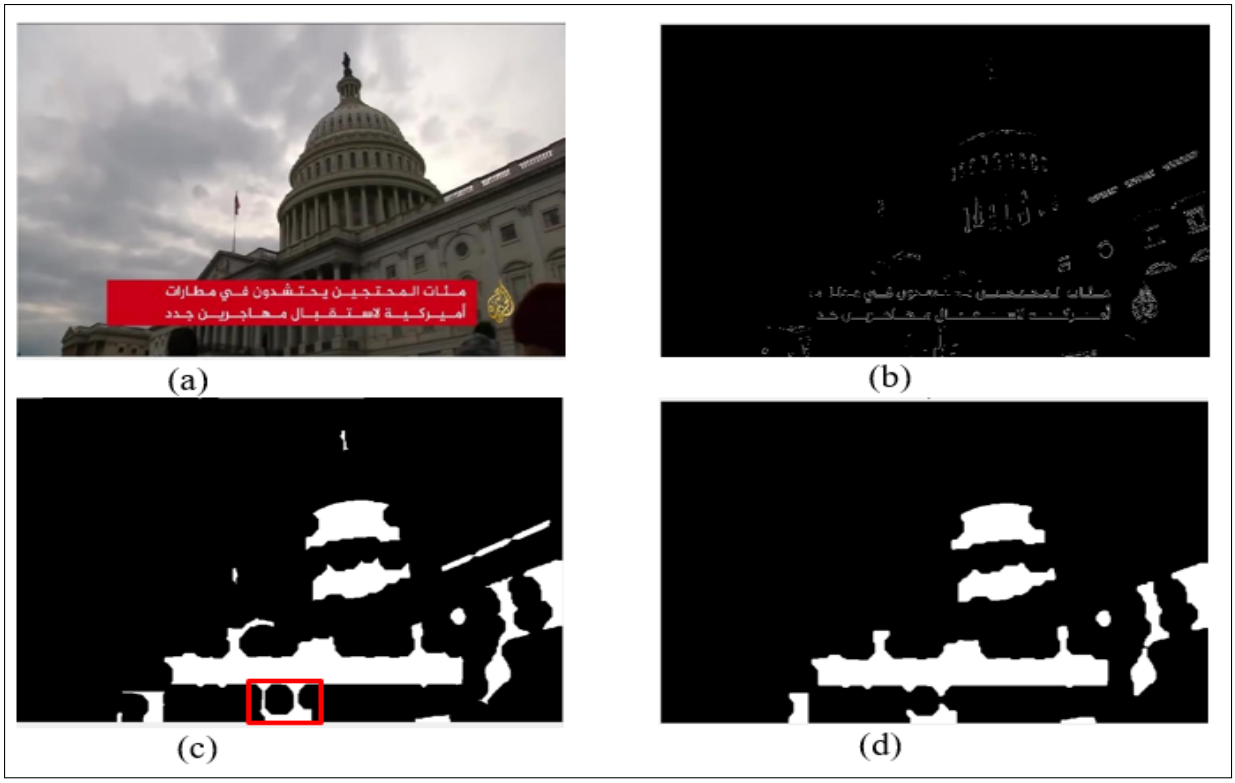

Figure 4.6: Operations morphologiques : (a) image originale (b) MSER (c) résultat de fermeture (d) résultat d'ouverture.

4.1.1.3 Filtrage par des règles heuristiques

Pour améliorer la qualité de la localisation des régions textuelles et supprimer les éventuelles fausses détections, nous précédons à une étape de filtrage basée sur des règles heuristiques. La procédure de filtrage exploite les caractéristiques géométriques du texte arabe. Elle est constituée de deux étapes : filtrage global et filtrage local.

1 **Filtrage global** : Pour que le texte incrusté dans la vidéo soit facilement repérable, il doit être placé dans une région de grande taille. Par conséquent, la région de texte doit avoir une taille qui garantit le critère de lisibilité. Pour une région donnée R_i,W_i, H_i et respectivement la largeur et la hauteur de cette région. La surface est définie par : A_i=W_i*H_i. la région candidate est considérée comme une région de texte si les conditions suivantes sont vérifiées :

$$seuil1 \leq A\_i \leq seuil2 \tag{4.1}$$

2 **Filtrage local** : Le filtrage global est une étape nécessaire mais pas suffisante pour valider une région du texte. En effet, plusieurs objets ayant les mêmes caractéristiques de texte (taille) peuvent être détectés comme étant des régions textuelles.

4.1.2 Détection de la ligne centrale du texte

Pour remédier à ce problème, nous introduisons un nouveau descripteur géométrique, nommé la ligne centrale de texte. Ce descripteur est inspiré de la structure géométrique de texte arabe. Comme il est illustré dans la Figure 4.7.a, le texte arabe suit toujours une ligne centrale (Baseline) qui regroupe les différents mots dans la même direction.

Notre contribution à ce niveau consiste à exploiter ce descripteur comme une signature spécifique pour valider la détection finale des régions de texte. Nous définissons la ligne centrale de texte (Baseline) par une droite qui passe par un ensemble des segments de mots qui sont adjacents et alignés comme il est présenté dans la Figure 4.7.b.

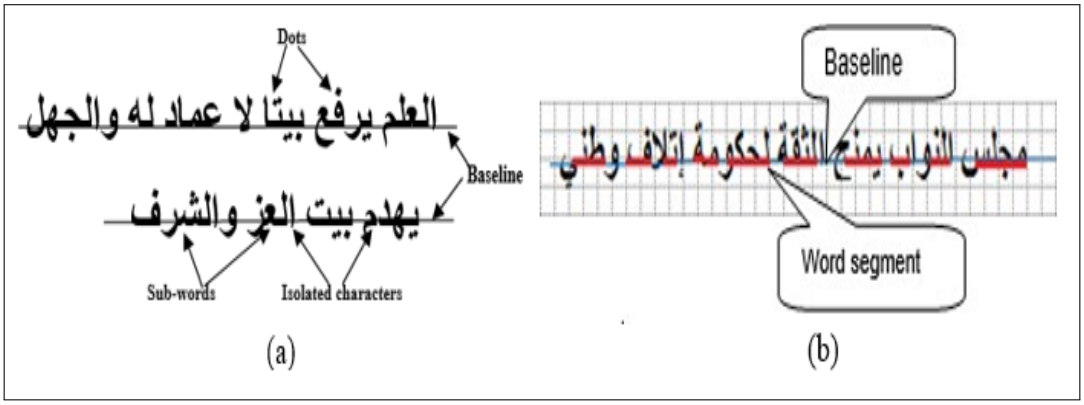

Figure 4.7: La ligne centrale du texte arabe.

Les étapes d'extraction de la ligne de base sont décrites par l'algorithme 2.

4.1.2.1 Extraction des segments de droite

La première étape consiste à extraire les différents segments de droites existants dans l'image. Pour ce faire, nous avons utilisé la méthode de transformée de Hough.

Algorithm 2: détection de la ligne de base

Input :RC une image de région candidate
Output : LB(firstpoint, endpoint) : ligne de base
H : accumulateur initialiser à 0
SM : Liste des segments des mots
droites : liste de tous les droites détectés dans RC
// *****transformer de hough***//**
for all x $\in$ *RC do*
for all y $\in$ *RC do*
if edge point in (x,y) then
for all θ **do**
r =sin(θ**)** $*$ $y + cos(\theta) * x$.
incrémenter la valeur de H (θ **r)**
endfor
endif
endfor
endfor
peaks = max_(H)
droites=chercher_droites(RC,peaks)
// ****sélection des segments de mots*//**
for each droite ((x1,x2), (y1,y2) $\in$ *droites do*
if(y1==y2) do
longueur= sqrt((x1-x2)$^2 + (y1 - y2)^2$)
if seuil1$\preceq$ *longueur* $\preceq$ *seuil*2 *do*
ajouter droite in SM
endif
endif
endfor
if taile_SM ≥ 3
trace la ligne de base LB

Historiquement, la méthode de Hough a été proposée par Paul Hough en 1959 et fait l'objet d'un brevet (IBM) en 1962. Elle est basée sur le principe suivant : Chaque droite est définie en coordonnées polaire par l'équation suivante :

$$r = sin(\theta)y + cos(\theta)x. \tag{4.2}$$

avec r est la distance entre l'origine et la droite et θ est l'angle entre l'axe x et le vecteur r. L'espace paramétré $r - \theta$ ici est borné par $r \in [-d, d]$ avec d est la taille de la diagonale de l'image et $\theta \in [-\Pi/2, \Pi/2]$ ($\pm$ 90°) .

Une droite dans l'image s'exprime comme un point dans l'espace (r, θ) et chaque point de l'image correspond à une sinusoïde dans l'espace de paramètre (r, θ).

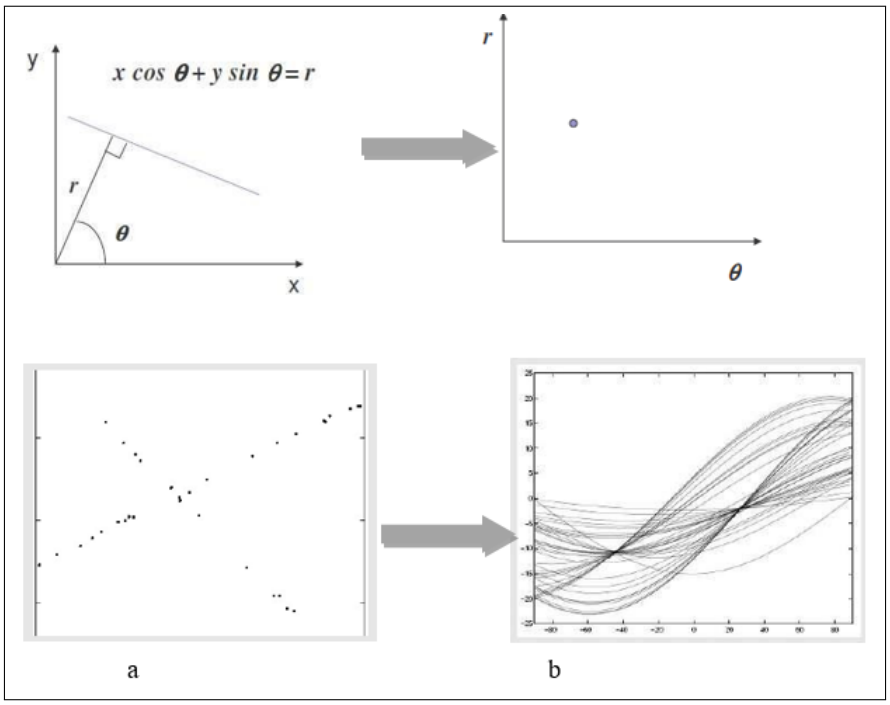

Figure 4.8: Principe de la transformée de Hough. a) espace de l'image. b) espace de paramètres..

Les points d'intersection dans l'espace de paramètre sont utilisés pour trouver les droites dans l'espace image.

Dans notre travail, nous détectons tout d'abord le contour à l'aide de la méthode de Canny [Can86] puis nous utilisons **l'algorithme 3** pour appliquer la transformée de Hough et détecter les segments de droites.

À la fin de l'exécution, les valeurs des cases de l'accumulateur $Vote[r][\theta]$ correspondent au nombre de points « les votes ». Une case de la grille d'accumulation qui possède un score élevé qui correspond, dans l'espace de l'image, à une droite de paramètres.

La Figure 4.9 illustre les résultats de l'application de la méthode de Hough sur une image HD . Dans nos expérimentations, nous remarquons que les segments de mots sont bien détectés dans les images HD alors que pour les images SD, il existe de fausses détections. C'est ce qui nécessite une méthode de filtrage.

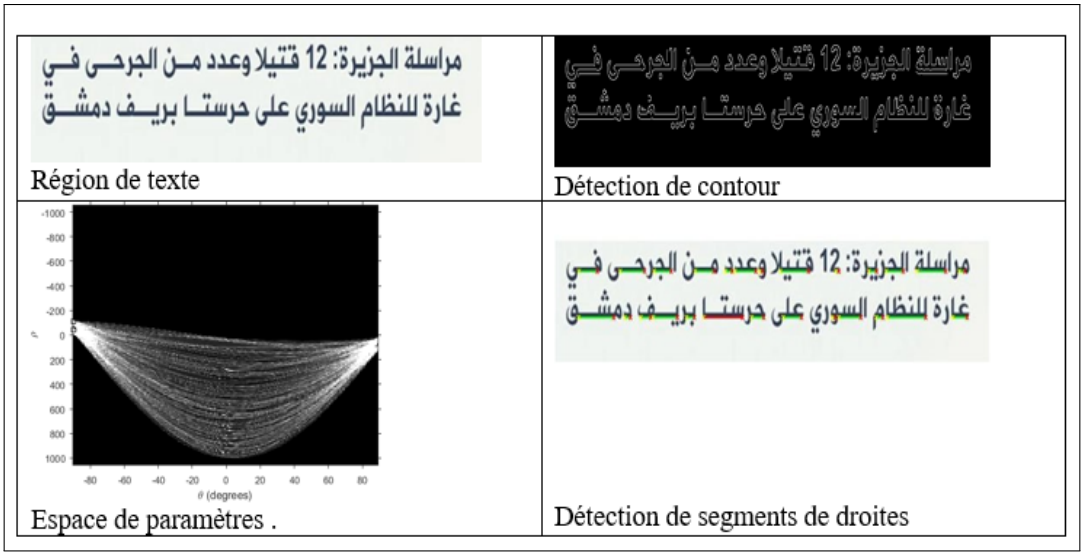

Figure 4.9: Les étapes de détection des segments des droites

4.1.2.2 Identification des segments de mots

La deuxième étape consiste à identifier les segments qui correspondent aux segments des mots en se basant sur deux critères géométriques : l'orientation et la longueur. Chaque segment est défini par un point de départ F (x_i, y_i) et un point d'arrivée E (x_j, y_j). Les segments détectés sont considérés comme des segments de mot, si ces deux conditions sont vérifiées :

1 Orientation : Nous acceptons seulement les segments horizontaux qui vérifient la condition suivante :

$$y_i = y_j \tag{4.3}$$

2 Longueur L :

$$seuil1 \leq \sqrt{(x\_j - x\_i) + (y\_j - y\_i))^2} \leq seuil2 \tag{4.4}$$

Une fois les segments de mots sont bien détectés, nous traçons la droite qui correspond à la ligne centrale et qui passe par le milieu de ces segments. Finalement, la validation de la région de texte se base sur la présence de la ligne centrale. Seules les régions qui possèdent une ligne de base seront acceptées.

4.2 Contribution 2 :Modélisation multi-facette du contenu vidéo

Dans cette section, nous présentons le schéma de modélisation que nous proposons pour l'exploitation de l'information textuelle dans le cadre de l'indexation sémantique d'un journal télévisé en langue arabe. Ce schéma se base sur une modélisation multi-facette permettant de structurer l'information textuelle pour fournir une description abstraite et sémantique du contenu vidéo. La Figure 4.10 illustre l'architecture du modèle proposé.

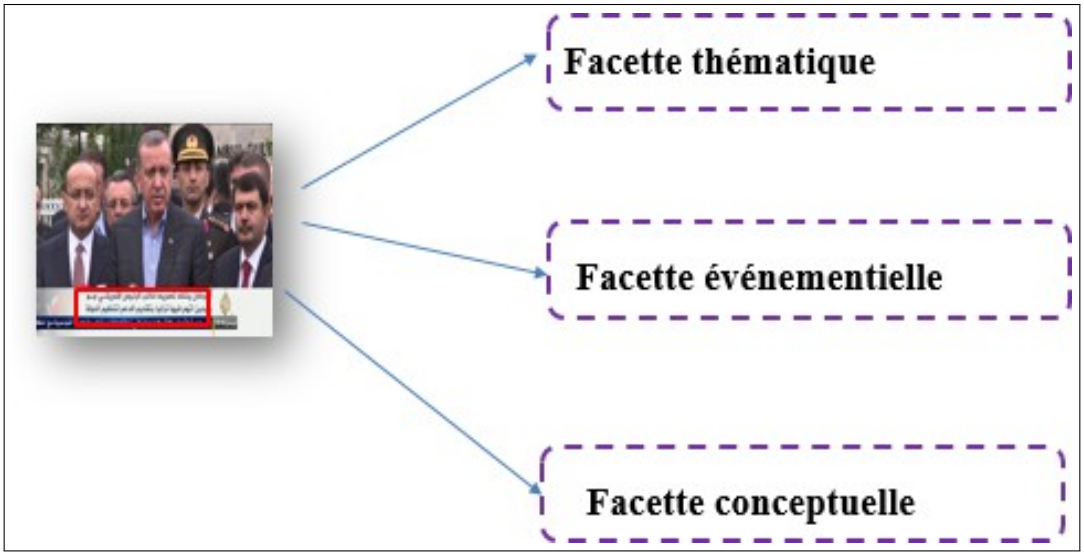

Figure 4.10: Description de contenu vidéo par le modèle multi-facette.

Le modèle proposé suggère trois types de facettes :

0. **Une facette conceptuelle** : cette facette regroupe l'ensemble des concepts sémantique liés au lieu, nom de personne et nom d'organisation
0. **Une facette événementielle** : elle décrit les différents événements contenus dans le journal télévisé.
0. **Une facette thématique** : elle permet d'associer un thème au contenu textuel. Ceci fournit une vue globale sur le contenu vidéo tel que politique, sport, éducation...

Notre proposition se base sur les techniques d'extraction de l'information et le traitement linguistique en langue arabe. Elle permet de repérer les informations pertinentes dans le contenu vidéo. Chaque facette de notre modèle est implémentée par une méthode spécifique d'une analyse linguistique de texte.
Dans la facette conceptuelle, nous utilisons une méthode de reconnaissance des entités nommées pour extraire les concepts liés aux personnes, organisations et lieux. Une méthode d'extraction des évènements est aussi proposée pour repérer les événements dans le texte. Une troisième méthode pour la classification thématique se base sur les deux précédentes méthodes pour fournir une représentation sémantique de chaque thème.

4.2.1 Notion de facette

Une facette est définie comme une interprétation particulière qui regroupe un ensemble des informations permettant une description sémantique du contenu vidéo. Il existe deux facettes principales : une facette physique qui représente l'entité perçue par l'œil humain à travers des descripteurs bas niveau (couleur, texture, forme), et une facette logique qui regroupe des interprétations et des descripteurs plus sémantiques [MCh05]. Dans notre travail, nous nous intéressons plus particulièrement aux facettes logiques. La section suivante présente avec plus des détails les différentes facettes proposées de notre schéma de modélisation.

4.2.2 Facette conceptuelle

La reconnaissance des entités nommées (REN) est une des techniques de traitement automatique de langues naturelles les plus répondu. C'est un axe de domaine de l'extraction de l'information qui permet la reconnaissance des entités linguistiques. Ces entités

rassemblent traditionnellement à l'ensemble des noms propres (noms de personnes, de lieu et d'organisation) et à certaines expressions numériques et temporelles (expressions de dates, de temps, de pourcentages, etc.). Les systèmes d'indexation et de la recherche des vidéos par le contenu sémantique pourraient bénéficier considérablement de la REN, notamment dans la réponse aux requêtes des utilisateurs qui implique le repérage des entités nommées [MCh05]. Par exemple, la reconnaissance des entités nommées de type personne et organisation portent souvent des informations sur les locuteurs et les participants dans les actions déroulés dans la vidéo (Figure 4.11).

Figure 4.11: Exemples des entités nommées.

Dans ce sens, la REN peut être utilisée de manière à reconnaître les éléments sémantiques au sein de contenu vidéo qui nous permet plus tard de trouver les résultats pertinents aux requêtes des utilisateurs. L'objectif de cette facette est d'extraire les concepts liés à la personne, l'organisation et le lieu.

Nous avons utilisé l'outil linguistique Farasa dont le module de reconnaissance des entités nommées de cette outil est basé sur une approche par apprentissage en utilisant la méthode Champs Aléatoires Conditionnels (CAC) pour l'apprentissage supervisé des structures séquentielles des entités nommées. De plus, elle intègre un grand corpus d'apprentissage issu d'Arabic Wikipédia et DBpedia.

4.2.3 Facette événementielle

Par définition, un événement est un fait qui survient à un moment donné. Au sens général, il signifie tout ce qui arrive et possède un caractère spécifique [MCh05]. Vu la diversité de contenu d'un journal télévisé, nous avons définis une liste des évènements couvrant plusieurs thématiques. Les événements représentent des actions réelles telles que les activités d'une personne ou bien un groupe de personnes. Par exemple **انتخابات، مظاهرات** d'autres événements un peu plus génériques sont classés selon un sujet particulier comme par exemple les événements de météo **فيضانات**.

Figure 4.12: Exemple des évènements.

Le tableau 4.1 décrit la taxonomie des évènements prise en compte dans notre approche. Cette taxonomie se base sur les modélisations existantes dans le domaine telles que celles présentées en ACE 2005 et également par l'observation des différents types d'événements rapportés dans les journaux télévisés sur plusieurs thèmes tels que politique, sport, social, etc.

Cette facette est composée essentiellement par deux étapes : élaboration de déclencheurs évènementiels et extraction des évènements

Elaboration des déclencheurs évènementiels. L'extraction des évènements est généralement assimilée à la détection de des déclencheurs (ou triggers) au sein de la phrase. Cette modélisation, introduite par la campagne d'évaluation ACE, est prédominante

Thème	Evènements
Politique	اجتماع وزاري، استقالة، انتخابات، محادثات، مصادقة، تشكيل حكومة
Sport	سباق رياضي، بطولة رياضية، مباره رياضية
Social	إضراب، مظاهرات، زو اج، الهجرة ، أعياد، موسم الحج، عودة مدرسية
Militaire	اعتقال، انفجار، انقلاب عسكري، غارة جوية، مناورات عسكرية، هجوم إرهابي، اغتيال، الحرب

Tableau 4.1: Liste des événements.

parmi les approches récentes. Les déclencheurs sont définis comme des indices linguistiques qui indiquent le déroulement de l'évènement. Ils peuvent être présentés sous la forme des verbes ou des noms. En comparaison avec les autres langues, le domaine de la recherche abordant la thématique d'extraction des évènements en langue arabe est encore très limité et même les travaux existants ne fournissent aucun corpus annoté et aucun dictionnaire de déclencheurs disponible en ligne. Afin d'accomplir ces besoins, nous avons proposé une nouvelle méthode pour l'élaboration des déclencheurs évènementiels en utilisant les dictionnaires en langue arabes comme sources des données : **الجامع المعجم الوسيط، الرائد ،لسان العرب ،المحيط**[1]

Notre proposition consiste à définir deux types de déclencheurs : déclencheurs morphologiques et déclencheurs sémantiques pour chaque évènement.

0. **Les déclencheurs morphologiques** regroupent tous les termes qui partagent la même forme morphologique de nom de l'évènement. Pour déterminer cette liste de déclencheurs, nous avons projeté le nom de l'évènement sur les dictionnaires et nous cherchons ses différentes formes morphologiques. (Figure 4.13)

1. https://www.almaany.com/ar/dict/ar-ar/

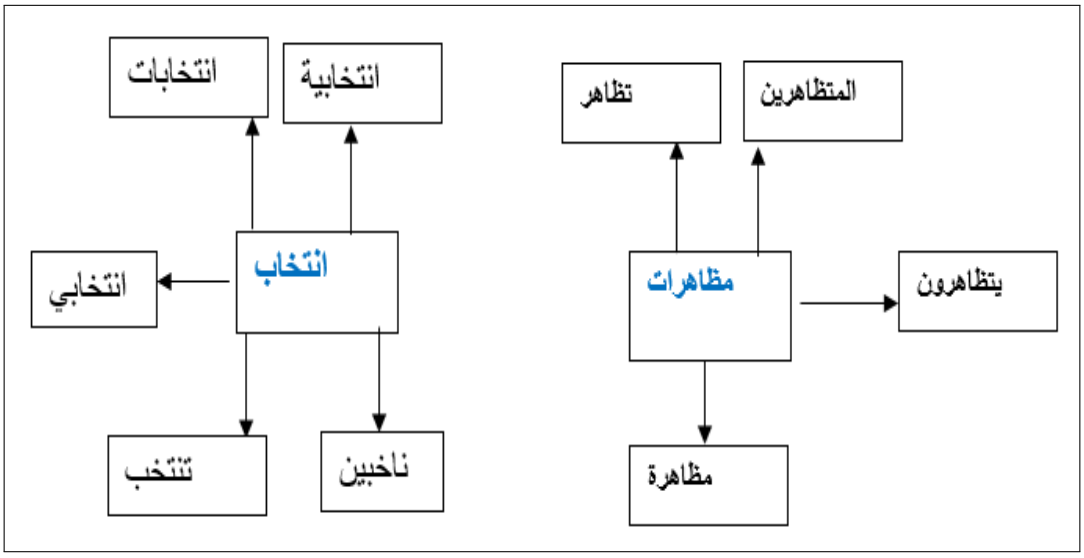

Figure 4.13: Exemples des déclencheurs morphologiques

0. **Les déclencheurs sémantiques** désignent l'ensemble des termes qui ont le même sens que le nom de l'évènement. Généralement, ce type de déclencheurs regroupe la liste des synonymes. Selon le même principe des déclencheurs morphologiques, nous avons utilisé les dictionnaires pour déterminer les termes les plus proches sémantiquement selon la thématique abordée dans la vidéo (tableau 4.2).

Evènement	Déclencheurs sémantiques
انتخاب	اقتراع , تصويت, استطلاع
مظاهرات	مسيرة احتجاج

Tableau 4.2: Exemples des déclencheurs sémantiques

Après la phase de définition des déclencheurs évènementiels, nous utilisons un analyseur morphologique de l'outil Farsasa pour attribuer à chaque déclencheur son lemme. Le processus de lemmatisation désigne l'analyse du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme (forme canonique). Dans notre cas, le but est de réduire les mots qui dérivent d'une même racine lexicale à la forme la plus réduite.

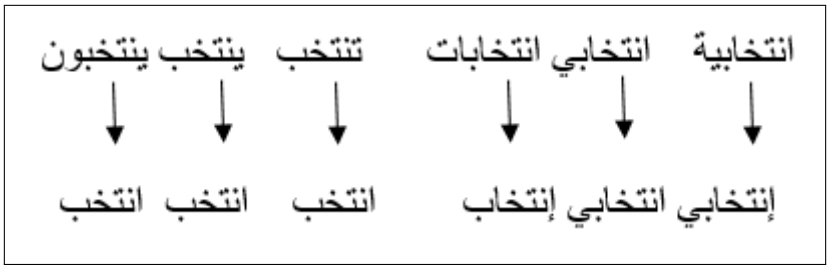

FIGURE 4.14: Processus de lemmatisation

Extraction des évènements L'extraction des évènements est basée sur une projection des textes reconnus dans la vidéo sur la liste des déclencheurs définie dans la section précédente. Les différentes étapes sont décrites dans la figure 4.15.

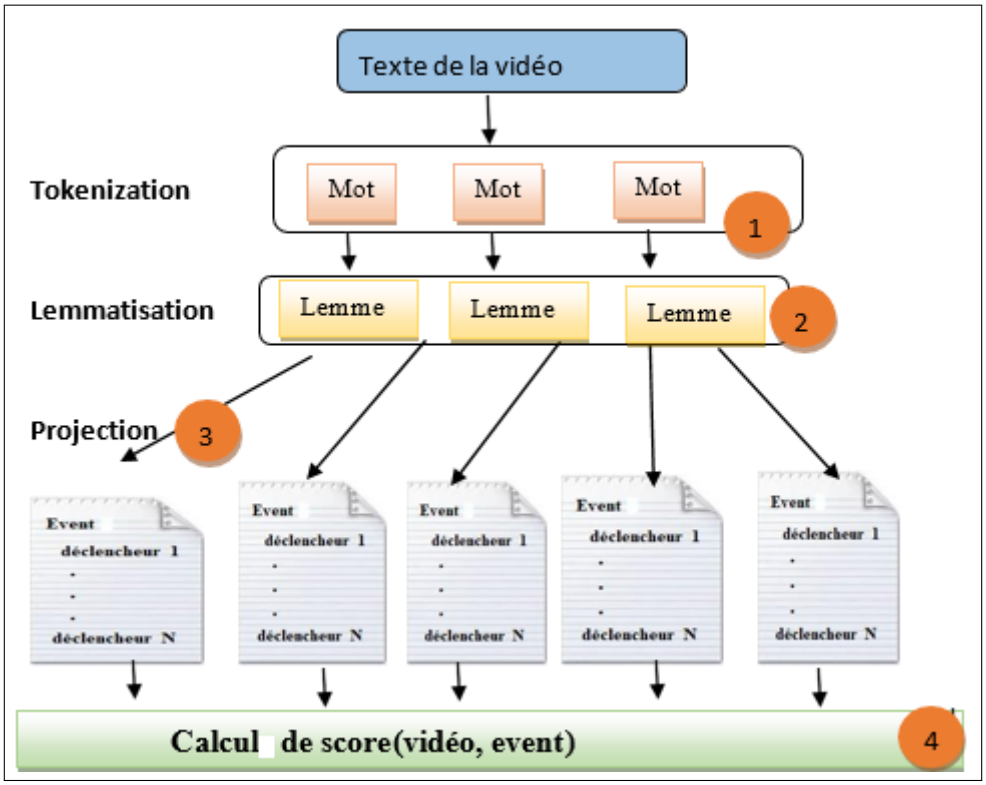

FIGURE 4.15: Processus d'extraction des évènements

En première étape, le texte issu de la vidéo est découpé en mots (étape de tokenization). Ensuite, Les mots obtenus seront lemmatisés et comparés avec les lemmes de tous les déclencheurs dans le but de déterminer l'évènement associé au contenu de la vidéo avec son score de pondération, calculé par la formule suivante.

$$Score(vidéo, event) = \frac{nombre\_des\_déclencheurs\_détectés}{nombre\_total\_des\_mots}. \tag{4.5}$$

Après le calcul de score, la vidéo est annotée par l'évènement qui possède une valeur de pondération la plus élevée.

4.3 Contribution 3 :Classfication de texte court

La facette thématique consiste à extraire les informations relatives aux thématiques de contenu vidéo. Dans le cas des journaux télévisés, les textes sont généralement courts et ne dépassent pas quinze mots (Figure 4.16).

Figure 4.16: Exemples de texte issu de la vidéo

Ceci pose un défi additionnel par rapport à la classification des textes traditionnels. En effet, à cause de la taille réduite, ces textes ne fournissent pas un nombre suffisant d'occurrences des mots, rendant ainsi l'utilisation de mesures de similarité ainsi que l'identification du contexte plus difficile. Pour résoudre à ce problème, certains travaux ont proposé d'enrichir et d'élargir la taille du texte court par des informations supplémentaires issues des ressources externes [BY07 ; SH06 ; MS07 ; GN11].
De même, les ressources lexico-sémantiques, à savoir les ontologies ont été utilisées pour améliorer la classification des textes courts. En effet, l'ontologie est une spécification formelle, explicite et sémantique de la conceptualisation d'un domaine. Elle est constituée d'un ensemble des concepts reliés entre eux par des relations sémantiques précises. L'intérêt de l'utilisation de l'ontologie réside dans son aptitude à représenter les connaissances et les interpréter sémantiquement. Les différents éléments d'une ontologie peuvent soit enrichir ou bien remplacer le texte court. Plusieurs travaux ont utilisé des ontologies universelles comme BableNet [FV15] WordNet et ConceptNet [FW14] pour améliorer la

classification des textes courts.
Cependant, l'utilisation des ontologies reste encore une problématique ouverte notamment pour la classification des journaux télévisés. Ceci est dû à la variété thématique dans ce genre de document (politique, sport, etc..) et à la couverture partielle des ontologies de domaines.
Dans le cadre de notre travail, nous avons proposé un nouveau modèle de représentation textuelle nommé, sac des marqueurs (bag of triggers). L'objectif principal est de fournir un modèle permettant de représenter et d'enrichir la description sémantique du texte court de manière que le système soit capable de déterminer efficacement la thématique du vidéo.

4.3.1 Utilisation des entités sémantiques

La particularité de ce modèle réside dans l'utilisation des entités sémantiques (les entités nommées et les évènements) comme étant des termes descriptifs de chaque thème. En effet, ces unités linguistiques jouent un rôle potentiel dans l'identification du contexte et la classification sémantique du contenu textuel. De plus, elles sont très utilisées dans les journaux télévisés.
La méthode proposée fait partie des méthodes sémantiques et utilise la même philosophie de l'extraction d'évènement. Plus particulièrement, elle s'appuie sur le repérage des marqueurs thématique existants dans le texte et marquent le type de sa thématique. Par exemple, on peut voir dans un texte que l'évènement élection (**انتخاب**) est un marqueur portant une valeur sémantique d'une action politique et permettant de déduire qu'il s'agit de la thématique "politique ". De la même façon, l'entité nommée **نادي برشلونة** réfère à une organisation sportive et permet de marquer la thématique "sport. Pour ces raisons, nous pensons que l'utilisation de ces éléments linguistiques permet de déterminer la thématique de la séquence vidéo même avec un nombre réduit des mots car ces entités sont riches en sémantique. Chaque thème est modélisé par un ensemble des marqueurs thématiques T = MA, ME

Où MA : désigne un marqueur agentif qui regroupe les noms des personnes et ses fonctions ainsi que les noms des organisations.

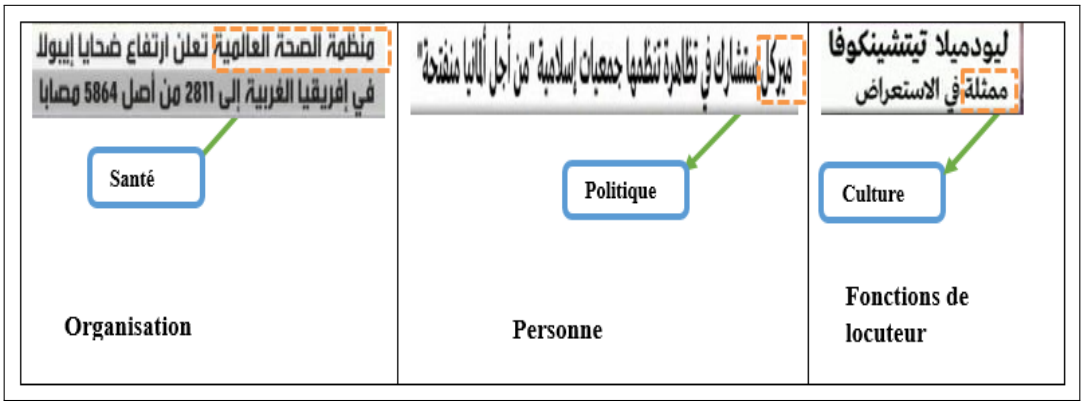

Figure 4.17: Exemples des marqueurs agentifs.

ME : Marqueur évènementiel désigne les évènements qui marquent le type de thème abordée dans la vidéo

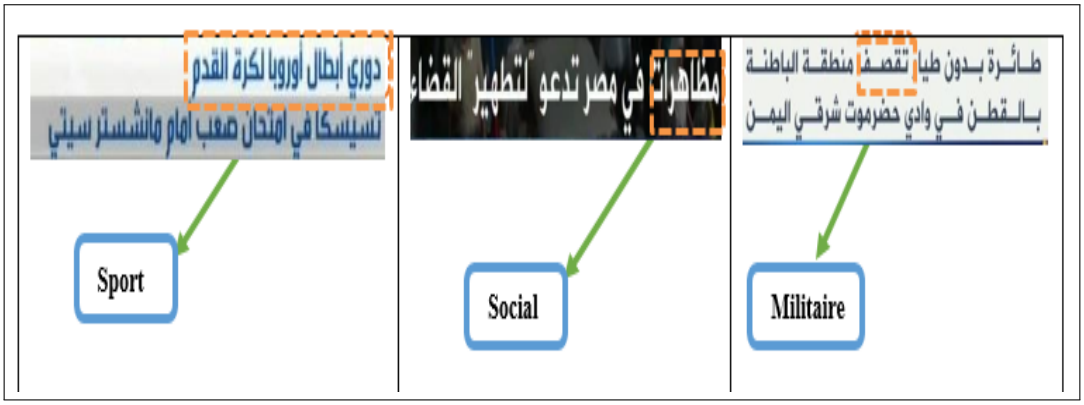

Figure 4.18: Exemples des marqueurs évènementiels.

4.3.2 Classification de texte par thématique

Nous avons proposé une méthode de classification thématique basée sur deux étapes (Figure 4.19) :

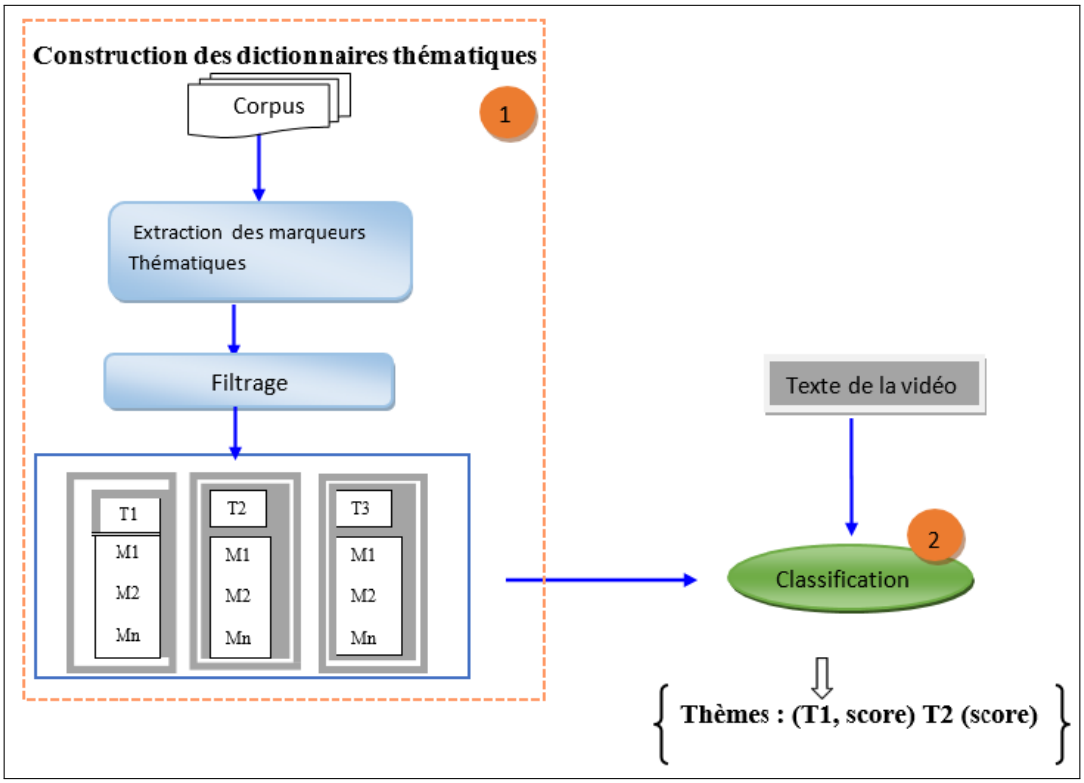

Figure 4.19: Processus de classification thématique.

Une étape d'extraction des marqueurs thématiques qui introduit plus de sémantique en utilisant un nouveau modèle de représentation nommé sac des marqueurs (Bag of triggers BOT) pour chaque thème.

Une phase de classification qui permet de déterminer les thématiques de texte issu de la vidéo.

Construction des dictionnaires thématiques L'extraction de descripteurs est une étape cruciale, elle s'effectue en amont du processus de classification. L'objectif est d'identifier l'information contenue dans le texte et de la représenter au moyen d'un ensemble des unités appelées descripteurs. Plus exactement, le processus d'extraction des descripteurs est le transfert de l'information contenue dans le texte vers un autre espace plus représentatif.

Cette étape consiste à extraire les marqueurs thématiques à partir de corpus d'entrainement d'une façon semi-automatique. Tout d'abord, nous appliquons les méthodes proposées dans la section précédente pour extraire les entités nommées et les événements. Ensuite, nous avons élaboré manuellement une étape de filtrage pour sélectionner seule-

ment les marqueurs les plus adéquates pour représenter le type de thème (supprimer les noms des présentateurs des émission TV, les noms des journalistes).
Le tableau ci-dessous décrit le corpus d'entrainement utilisé ainsi que le nombre des marqueurs de chaque thème.

classification Cette étape consiste à calculer la similarité entre le vecteur de texte et les vecteurs descriptifs des différents thèmes et associer un score qui reflète le degré de pertinence de chaque thème selon la formule suivante :

$$Similarité(Vect\_vidéo, Vect\_théme_i) = \frac{(Vect\_vidéo) \cap (Vect\_théme_i)}{(Vect\_vidéo) \cup (Vect\_théme_i)} \tag{4.6}$$

Conclusion

Ce chapitre a détaillé les méthodes que nous avons élaborées pour l'indexation sémantique des journaux télévisés en langue arabes. L'objectif de ces méthodes consiste à extraire et à analyser l'information textuelle incrustée dans la vidéo afin de générer une description symbolique de contenu. Nous avons développé en premier lieu, une méthode heuristique d'extraction de texte incrusté dans la vidéo en exploitant les caractéristiques spécifiques de texte arabe. Nous avons proposé, en deuxième lieu, un module pour la modélisation multi-facette de l'information textuelle. Ce module intègre une phase de reconnaissance des entités nommées et d'extraction des évènements. La troisième phase consiste à extraire les informations relatives à la thématique de contenu. Pour ce faire, nous avons proposé une nouvelle méthode d'extraction des descripteurs qui exploite les entités nommées et les évènements comme des termes représentatifs à fin d'améliorer le processus de classification. Le chapitre suivant détaille l'évaluation des différentes méthodes proposées. Nous allons présenter les mesures d'évaluation utilisées ainsi que les résultats obtenus.

Chapitre

5 Expérimentation et évaluation

Introduction

Ce chapitre décrit l'implémentation de notre système d'indexation SISAVIN (Semantic Indexing System for Arabic Video News) ainsi que le corpus de test et la phase d'évaluation. Nous présentons, dans la première partie les aspects ergonomiques de l'interface principale et les différentes fonctionnalités de système SISAVIN. Nous décrivons ensuite, les caractéristiques de corpus de test, notamment les types des flux vidéo et les chaînes TV utilisées. La dernière partie détaille les résultats d'évaluation obtenus pour chaque méthode proposée dans ce travail de thèse.

5.1 SISAVIN : Présentation de système

5.1.1 Architecture fonctionnelle

L'objectif de notre travail consiste à mettre en œuvre un système d'indexation qui fournit une description sémantique de contenu des journaux télévisés en se basant sur le texte incrusté comme une source d'information. L'approche proposée s'appuie sur les techniques de détection et d'extraction des informations textuelles et une modélisation multi-facette de contenu vidéo. Dans cette section, nous décrivons l'architecture globale de notre système (Figure 5.1).

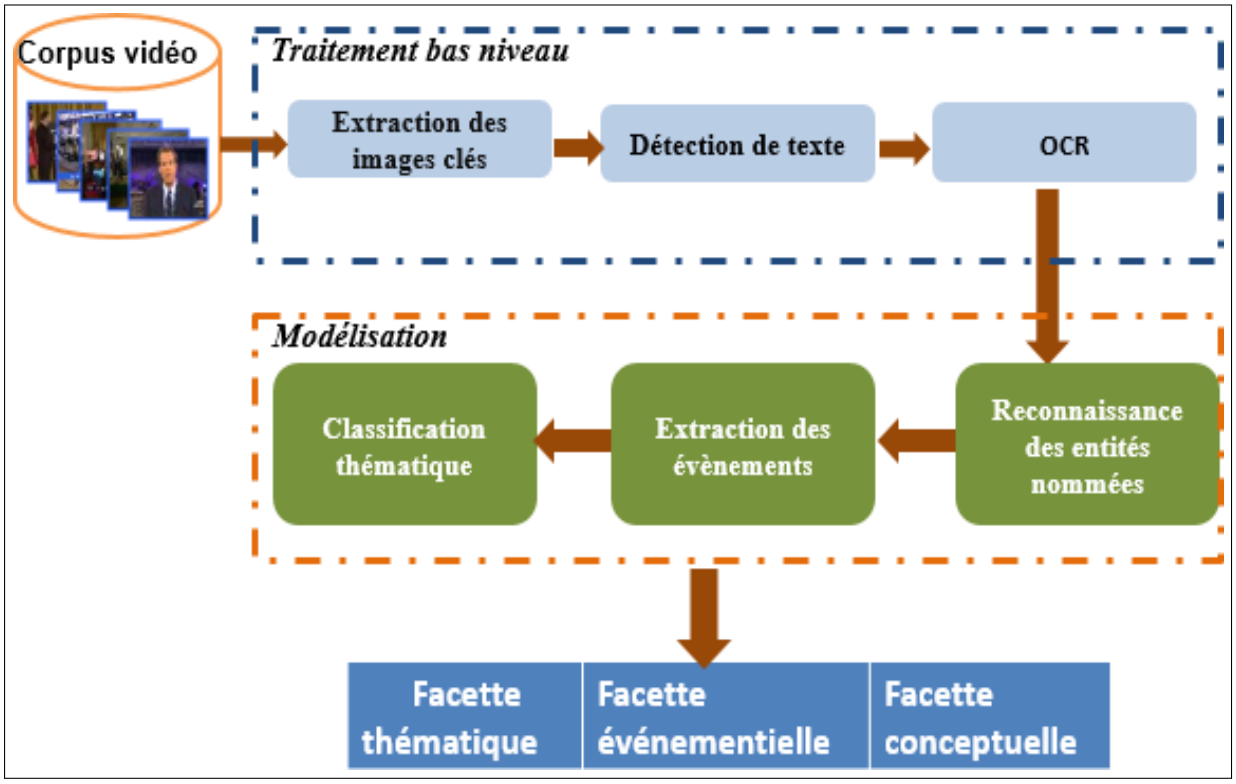

Figure 5.1: architecture fonctionelle du système SISAVIN

Cette architecture est composée de deux modules :

Un module de traitement bas niveau : introduit une phase d'extraction des images clés et une étape de détection et de reconnaissance (OCR) du texte incrusté.

1. **Extraction des images clés :** Vu le débit des flux vidéo (25 à 30 images/s) et la durée importante d'un journal télévisé qui semble être standardisée autour de 25 à 35 minutes, il est donc nécessaire d'appliquer une méthode d'extraction des images clés qui permet d'éviter la redondance des images similaires et réduire la complexité et le temps d'exécution.

   L'étude réalisée sur la structure du journal télévisé, nous permet de constater que le texte incrusté dans la vidéo est souvent apparu pendant une durée au minimum de 20 secondes pour pouvoir être lu et interprété.

   Dans ce travail, nous exploitons cette contrainte temporelle afin de sélectionner les images clés qui sont susceptibles de contenir le texte incrusté. La méthode proposée consiste à définir une fenêtre temporelle qui permet de parcourir la totalité de la vidéo et extraire une image clé après chaque durée N (1 image /20 secondes) (

2. **Détection de texte :** notre contribution, à ce niveau, réside dans l'exploitation des caractéristiques spécifiques de texte en langue arabe dans le cadre d'une approche heuristique de localisation des régions potentielles de texte.
3. **Reconnaissance du texte** Après avoir localisé la région de texte, cette dernière doit être binarisée pour séparer le texte de l'arrière-plan et faciliter la phase de la reconnaissance. Pour ce faire, nous appliquons la méthode Otsu [**103**]. Cette méthode permet de classer les pixels de l'image en deux classes (blanc pour le texte et noir pour l'arrière-plan). Après l'étape de la binarisation, nous utilisons le logiciel OCR Tesseract de Google [1] pour transformer les régions de texte vers une représentation textuelle traitable par le module de modélisation de notre système (voir Annexe 6.2).

Un module de modélisation sert à extraire les informations sémantiques permettant une description multi-facette du contenu vidéo. Il intègre une méthode de reconnaissance des entités nommées, d'extraction des événements et de classification thématique. Les informations sémantiques fournies par ce module seront structurées et répertoriées dans un référentiel (fichier xml) comme des indexes de chaque vidéo

5.1.2 Interface du système SISAVIN

Suite à notre proposition menée dans le chapitre précédant, nous avons implémenté un système d'indexation sémantique des journaux télévisés arabe nommé SISAVIN pour la réalisation des différents modules de notre approche d'indexation. La Figure 5.2 présente l'ergonomie de l'interface principale de notre système. Elle est composée essentiellement de six parties.

1. `https:https://github.com/tesseract-ocr/tesserac`

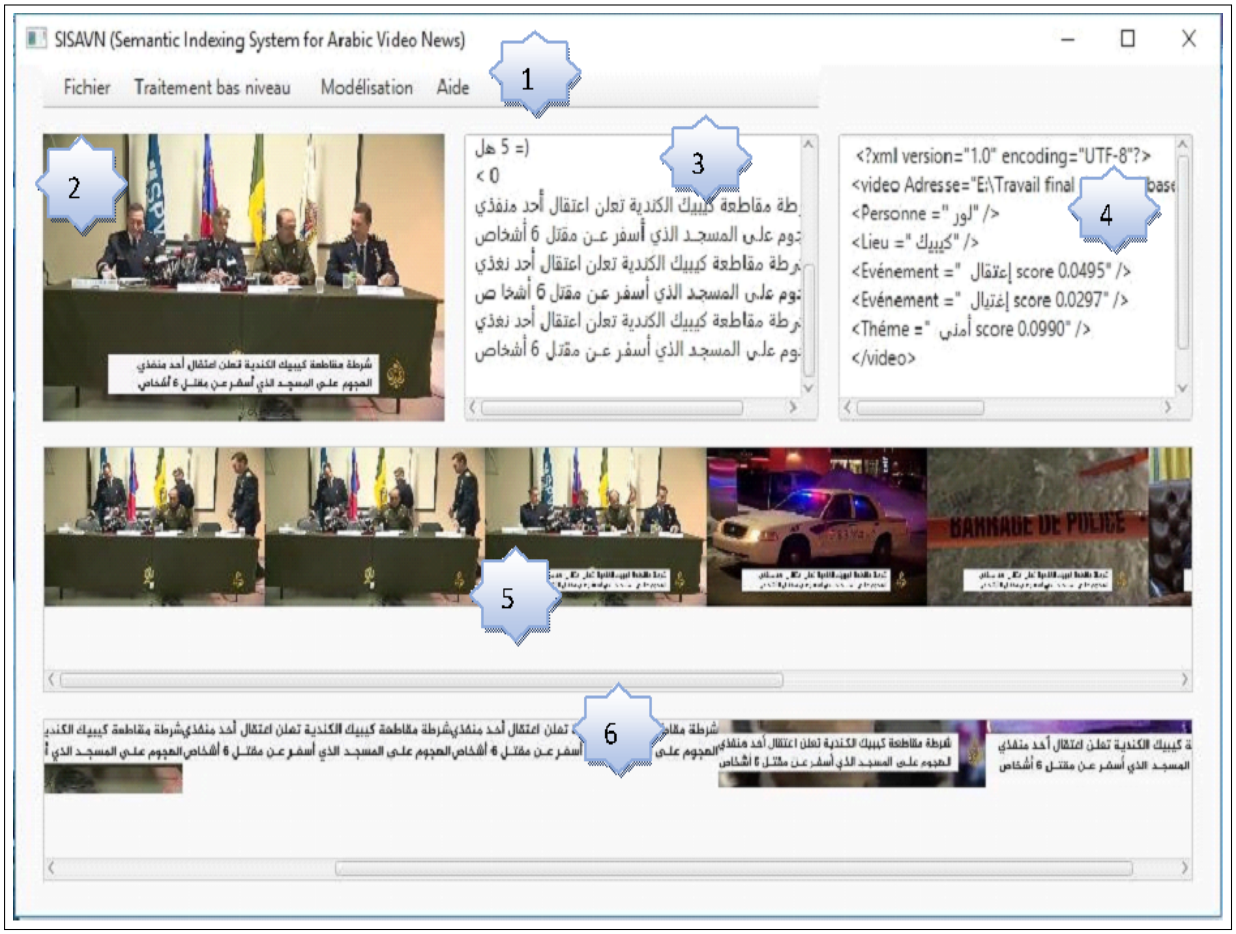

Figure 5.2: Interface principale du système SISAVIN

Le menu principal de l'interface (Figure 5.3) est composé de trois menus :

— Le menu fichier contient deux commandes : Ouvrir et quitter. La commande ouvrir permet d'importer un journal télévisé qui sera affiché en mode lecture dans la partie 2 de l'interface.

— Le menu traitement bas niveau qui fait appel aux modules d'extraction des images clés, de détection et de reconnaissance du texte.

Figure 5.3: Menu traitement bas niveau

Les résultats d'exécution seront affichés respectivement dans les parties 6, 5 et 3 de l'interface.

— Le menu modélisation permet d'exécuter et d'afficher les différentes facettes de modélisation (conceptuelle, événementielle, thématique). Le résultat est généré sous forme d'un fichier XML qui sera visualisé dans la partie 4 de l'interface.

Figure 5.4: Menu modélisation

La figure 5.5 présente un exemple d'un fichier XML généré par notre système d'indexation

```
<?xml version="1.0" encoding="UTF-8"?>
<video Adresse="E:\Arabic video news\aljazzera\video5\video5.MP4">
  <Lieu Name=" كيبيك" />
  <Evénement Name="  إعتقال score 0.495" />
  <Evénement Name="  إغتيال score 0.297" />
  <Théme Name="  أمني score 0.990" />
</video>
```

Figure 5.5: Exemple de résultat d'indexation d'un journal télévisé

5.1.3 Corpus d'évaluation

Pour tester et évaluer l'efficacité des méthodes proposées, nous avons construit un corpus de test nommé Arabic Video News et contenant plus de 126 séquences vidéo collectées à partir de six chaînes TV arabes différentes (AljazeeraHD, France24 Arabic, Russia Today Arabic TunisiaNat1 ,Almayadeen, Alarabiya). Trois types de flux vidéo ont été choisis : définition standard (SD, 720x576, 25 ips), définition standard (SD, 480x360, 25 ips) et haute définition (HD, 1920 x1080, 25 ips) et les vidéos sont repartis tel que décrit dans la table 5.5.

Chaîne TV	**Type de flux**	**Nombre des vidéos**	**Nombre des images clés**	**Durée**
Aljazeera	HD, 1920 x1080	30	720	4 heures
Alarabiya	HD, 1920 x1080	10	360	2 heures
France24 Arabic	SD, 720x576	11	360	2 heures
Russia Today Arabic	SD, 720x576	37	387	2,15 heures
TunisiaNat1	SD, 480x360	26	2961	16 ,45h
ALmayadeen	SD, 720x576	12	1800	10 heures
Total		**126**	**6588**	**36 ,6 heures**

Tableau 5.1: Corpus d'évaluation.

Les images clés sont obtenues par l'application de notre méthode d'extraction des

images clés proposée dans le chapitre 3. La Figure 5.6 présente des exemples des images extraites à partir des séquences vidéo dans le corpus d'évaluation. Nous remarquons que les caractéristiques visuelles (couleur, taille et style d'écriture) de l'information textuelle incrustée dans la vidéo varient d'une chaîne TV à une autre.

Figure 5.6: Exemples des images extraites à partir du corpus

5.1.4 Métriques d'évaluation

Une fois le SRI est construit, l'étape qui suit consiste à évaluer sa performance au niveau de satisfaction de l'utilisateur par les informations qu'il obtient. Plus les réponses du système correspondent au besoin de l'utilisateur, plus le système est meilleur. L'évaluation constitue donc une étape importante lors de la mise en œuvre d'un modèle d'information puisqu'elle permet d'estimer l'impact de chacune de ses caractéristiques pour obtenir des résultats pertinents. Plusieurs mesures ont été mises en place afin d'évaluer la pertinence du système d'indexation et de recherche d'information audiovisuelle. Parmi lesquels, nous citons :

— Rappel : elle mesure la capacité de SRI à présenter tous les documents pertinents. Mathématiquement, le rappel est défini comme le nombre de documents véritablement positifs (T_p) sur le nombre de vrais positifs plus le nombre de faux négatifs (F_n).

$$Rappel = \frac{T_p}{T_p + T_n} \tag{5.1}$$

— Précision : elle mesure la capacité de SRI à présenter que les éléments pertinents. Mathématiquement, la précision est définie comme le nombre de vrais documents positifs (T_p) par rapport au nombre de vrais positifs plus le nombre de faux positifs (F_p).

$$Précision = \frac{T_p}{T_p + F_p} \tag{5.2}$$

— F-mesure : cette mesure permet de combiner le rappel et la précision comme suit :

$$F - mesure = \frac{2 * Rappel * Précision}{Rappel + Précision} \tag{5.3}$$

Dans ce qui suit, nous allons détailler la procédure d'évaluation des méthodes proposées dans le cadre de ce travail.

5.2 SISAVIN : Evaluation du module de détection de texte

5.2.1 Mise en œuvre

5.2.1.1 Dataset

Pour évaluer la performance de la méthode proposée, nous avons utilisé deux corpus de

— ActivD [OH16] : Ce corpus est composé de 630 images collectées à partir de quatre chaînes TV (AljazeeraHD, France24 Arabic, Russia Today Arabic TunisiaNat1 pendant une période entre 2014-2015. Il est publié dans la conférence ICDAR 2015 dans le but d'évaluer les systèmes de détection de texte arabe dans les journaux télévisés. Nous avons utilisé ce corpus pour évaluer notre approche sur une base de référence et

comparer nos travaux par rapport au travail de Zayene (propriétaire de ce corpus).

— Arabic video News : Ce corpus est composé de 1000 images issues de notre collection de test. Le but consiste à évaluer et valider l'approche proposée pour la detection de texte qui sera utilisé dans d'autres modules de notre système d'indexation.

5.2.1.2 Métrique d'évaluation

Nous avons utilisé les mêmes métriques proposées dans [OH16 ; OH15]pour évaluer notre méthode. Ces métriques sont basées sur une correspondance graphique entre la région de texte G_i , avec i = 1...n dans la base de test et la région de texte $D_i$avec i = 1...m obtenue par notre système. Il existe trois cas pour l'appariement entre et :

— Appariement "one-to-one" : consiste à la mise en correspondance d'une région de la vérité terrain avec une région de texte localisée . Ce cas est présenté sur la Figure 5.7-(a).

— Appariement "one-to-many " : Une région de la vérité terrain est mise en correspondance avec un ensemble So des régions localisées . Ce cas est présenté sur la Figure 5.7-(b).

— Appariement "many-to-one " : une région localisée est mise en correspondance avec un ensemble Sm de régions de la vérité terrain. Ce cas est présenté sur la Figure 5.7-(c).

Figure 5.7: Cas de correspondance entre G_i (en vert) et D_i (en rouge) : de gauche à droite respectivement : (a) : one to one, (b) :one to many, (c) : many to one

Dans notre évaluation, nous considérons le cas de many to one vu la particularité de notre méthode qui consiste à fusionner les différents mots de texte dans une seule région. Les valeurs de rappel et de précision sont exprimées comme suites :

$$R = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (\frac{w_R(G_i, D_j)}{n}) \tag{5.4}$$

$$P = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (\frac{w_P(G_i, D_j)}{m}) \tag{5.5}$$

where

$$w_R(G_i, D_i) = \begin{cases} 1, if \quad \forall i \in G, k_R(G_i, D_j) \geq t_r \\ 0, otherwise \end{cases}$$

$$w_P(G_i, D_i) = \begin{cases} 1, if \quad \sum_{i \in G} k_p(G_i, D_j) \geq t_p \\ 0, otherwise \end{cases}$$

Notons que K $k_R(G_i, D_j) désigne la valeur de correspondance graphique entre les deux régions et est défini$

$$k_R(G_i, D_j) = \frac{Area(G_i) \cap Area(D_j)}{AreaG_i}$$

$k_P(G_i, D_j) = \frac{Area(G_i) \cap Area(D_j)}{AreaD_i}$(5.6)

t_r= seuil de rappel et t_p=seuil de précision. Durant la phase de l'expérimentation, nous avons utilisé les valeurs suivantes t_r=0.8 et t_p =0 .4.

5.2.2 Etude du choix des paramètres

5.2.2.1 Cas des paramètres de la Méthode MSER

La méthode MSER repose essentiellement sur 3 paramètres :

0. RA (RegionArea) : Ce paramètre sert à éliminer les petites et les grandes composantes connexes qui ne correspondent pas aux régions de texte. Nous choisissons RA_min=100, parce que le texte arabe contient sauvant des caractères isolés qui ne s'attachent pas aux autres caractères, commme ، ة أ ي et représentent des petites composantes connexes. RA_max=2000, car en langue arabe, les caractères s'attachent entre eux et forment des mots (grandes composantes connexes) .

0. MAV (MaxAreaVariation) :MAV est un rapport qui définit le changement relatif dans la surface d'une région sur des seuils successifs

$$MAV = \frac{R(t+1) - R(t)}{R(t)} \tag{5.7}$$

Une valeur proche de 0 signifie que la région garde la même taille alors qu'une valeur proche de 1 signifie que la région n'est pas stable. Dans note travail, nous choisissons MAV =0.25.

0. Δ : c'est le pas de seuil. Il est exprimé en pourcentage. Par exemple, dans une image 8 bits, une valeur de 4 pour cent signifie que l'incrément de seuil utilisé dans l'algorithme MSER est de 255 * 0,04 = 10.

   De ce fait, une valeur élevée retourne un minimum de composante connexes et risque de n'est pas détecter le texte et une valeur très basse engendre un nombre très élevé des fausses détections. la Figure 5.8 décrit les résultats des différentes valeurs de Δ en termes de précision et de rappel .

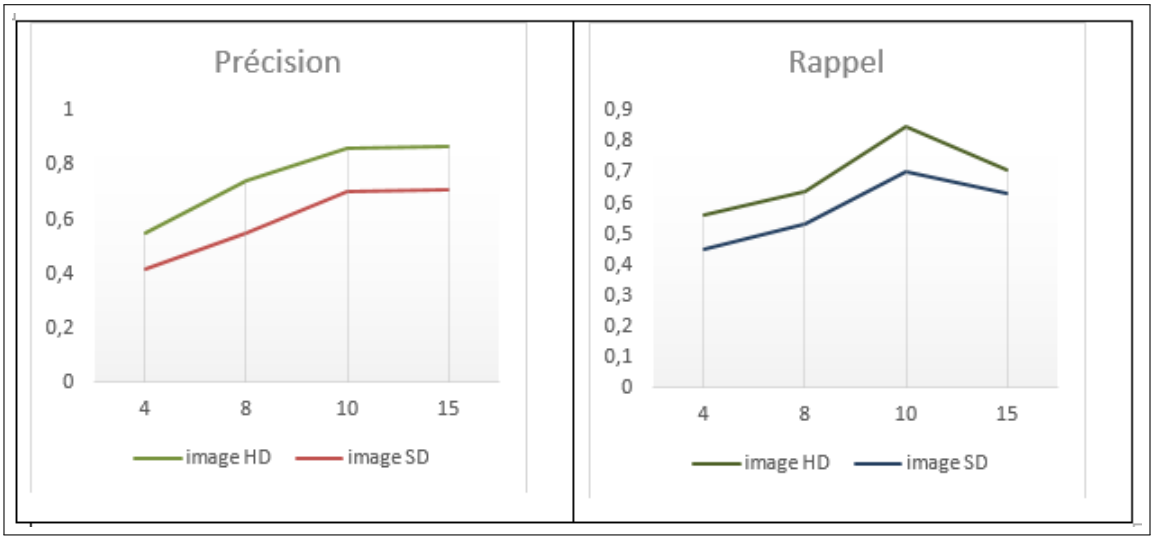

Figure 5.8: Influence de la valeur Δ

D'après les résultats obtenus, nous remarquons que la meilleure valeur de Δ est égale à 10. En effet, au delà de cette valeurs le taux de rappel commence à se dégrade. Ceci signifie que le nombre composantes connexes correspondants au texte se diminue (disparaître). De même, la meilleure valeur de précision est obtenu pur Δ=10.

5.2.2.2 Cas des paramètres des opérations morphologiques

Le choix du paramètre N à accorder au rayon de l'élément structurant B est une étape importante qui influence le résultat de regroupement des caractères en terme de précision. De ce fait, nous allons, expérimentalement, optimiser le choix de ce paramètre en le variant dans un intervalle de 10 à 30 avec un pas de 5 sur deux types des images : images HD issues de chaîne TV à haute résolution et images SD issues de chaîne TV à moyenne résolution.

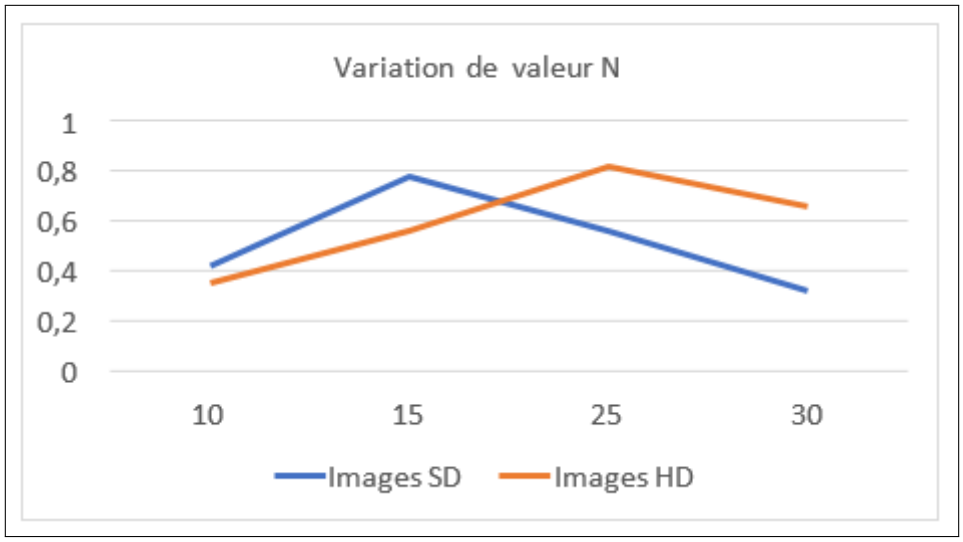

Figure 5.9: Etude de la valeur du rayon N.

En regardant le graphique de la figure 5.9, nous pouvons remarquer que la précision augmente jusqu' à la valeur 25 pour les images HD. Au-delà de cette valeur, la précision commence à se dégrader. De ce fait, nous considérons cette valeur $N = 25pixels$ comme une meilleure valeur de rayon de l'élément structurant. En effet, une accentuation de cette valeur cause des faux regroupements (très grandes régions) et une valeur inférieure de ce seuil engendre une perte des régions textuelles. Pour les images SD, la meilleure valeur est $N = 15$ pixels.

5.2.3 Etude comparative

Pendant la phase d'évaluation, nous avons comparé la méthode proposée avec les méthodes suivantes :

— **Méthode 1** : Cette méthode est proposée par Zayene [OH16] selon la méthodologie heuristique. Elle utilise la technique d'épaisseur de trait (SWT) pour identifier les caractères candidats. Ces caractères sont ensuite filtrés à l'aide des informations géométriques et pour éliminer les objets non textuels. Enfin, les régions correspondantes aux caractères sont fusionnées pour former les mots du texte.

— **Méthode 2** : C'est une méthode hybride proposée par Zayene [OH16]en utilisant l'opérateur SWT et une technique d'apprentissage non supervisé nommée, auto encoder convolutionel. Tout d'abord, les régions candidates de texte sont extraites en appliquant l'opérateur SWT et certaines règles heuristiques. Ensuite, ce système

utilise des auto- encodeurs convolutionnels pour produire automatiquement des caractéristiques textuelles qui ont été apprises à partir de régions étiquetées. Ces caractéristiques sont utilisées par la suite dans la phase de classification.

Corpus de test	**Méthode**	**Précision**	**Rappel**	**F-mesure**
ActiV-D (Chaînes TV HD)	Zayene [OH16]	0,76	0,77	0,76
	Zayene [OH15]	0 ,85	0.83	0.84
	Notre méthode	0.90	0.87	0.88
ActiV-D (Chaînes TV SD)	Zayene [OH16]	0.58	0.67	0.64
	Zayene [OH15]	0.74	0.78	0.76
	Notre méthode	0,72	0,75	0,74
Arabic Video News (Chaînes TV HD)	**Notre méthode**	0.91	0.85	0.87
Arabic Video News (Chaînes TV SD)	**Notre méthode**	0.76	0.79	0.77

Tableau 5.2: Comparaison de performances des méthodes existantes et de la méthode proposée.

Les résultats d'évaluation (tableau 5.2) prouvent que l'approche proposée fournit des excellents résultats pour les images issues des chaînes TV HD et cela en comparaison avec les deux autres travaux.

En effet, notre méthode donne un F-mesure supérieur à 0.87 pour les deux corpus de test. Ceci montre qu'elle est capable de détecter la plupart des régions de texte avec un minimum de fausses détections. De même, les résultats obtenus pour les images issues de chaînes TV SD sont globalement satisfaisants malgré la qualité et la résolution moyenne de ces images.

Quelques résultats de détection sont présentés dans la Figure 5.10. Nous remarquons que notre méthode est capable de localiser les régions de texte d'une manière précise. De plus, elle est robuste au changement de la couleur, de style de l'écriture ainsi que la taille et la position des régions textuelles.

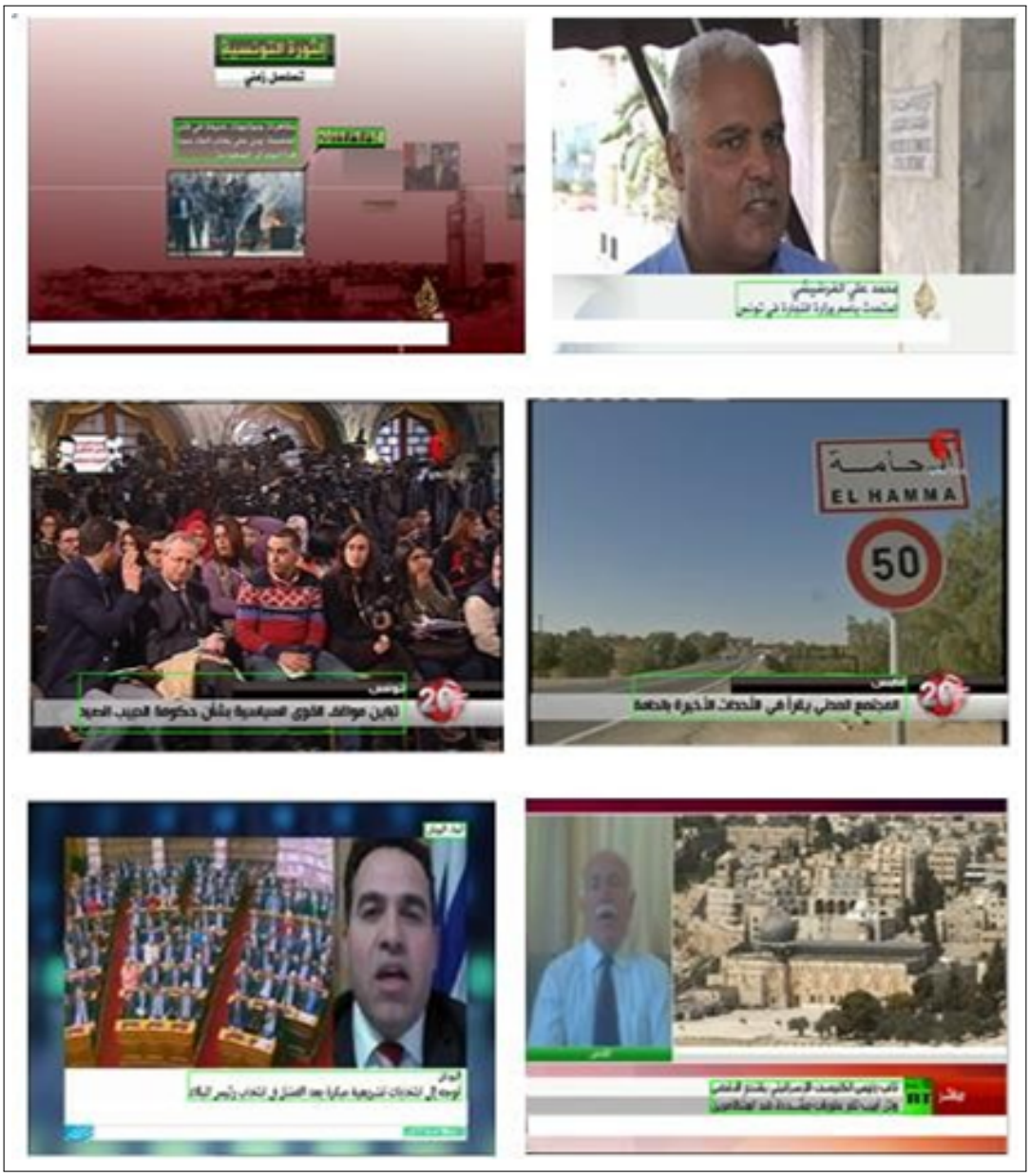

Figure 5.10: Exemples de détection du texte

5.3 SISAVIN : Évaluation du module de modélisation multi-facettes

Cette évaluation consiste à tester la performance de notre système d'indexation en termes d'exactitude des indexes sémantiques générées. Plus particulièrement, nous testons la capacité de système de fournir une description efficace par rapport aux trois facettes : conceptuelle (personne, lieu organisation), événementielle (15 évènements) et thématique (4 thèmes). Pour ce faire, nous utilisons le résultat final d'indexation généré sous forme d'un fichier xml pour l'ensemble de vidéos de notre corpus d'évaluation. Le résultat d'évaluation de différentes facettes est présenté dans la suite de cette section.

5.3.1 Cas de la facette conceptuelle

5.3.1.1 Mise en œuvre

Pour l'extraction de concepts, nous avons testé deux méthodes de reconnaissance des entités nommées :

Une méthode à base de règle implémentée sous la plateforme linguistique Nooj et inspirée du travail de Mesfar[Sli07]. Après avoir segmenté et exécuté une analyse morpho-syntaxique des textes issus des journaux télévisés, ces derniers sont annotés à travers des grammaires syntactico-sémantiques locales et des gazetteers, afin de localiser les entités pertinentes. Les gazetteers sont des dictionnaires de noms propres arabes regroupant une liste des noms et des prénoms les plus fréquents, des noms de localisations (villes, pays, fleuves, etc.) Et des noms d'organisations (organismes, compagnies, etc.). Les grammaires syntactico –sémantique locales sont représentées par des graphes sous forme de réseaux de transition augmentés.
Elles permettent d'identifier le type de chaque entité nommée par le biais des règles écrites manuellement. Ces règles sont généralement formées par des marqueurs lexicaux, nommées mots déclencheurs facilitant l'annotation des entités nommées qui n'apparaissent pas dans les dictionnaires.
Le tableau 5.3 présente les caractéristiques des données linguistiques utilisées lors l'évaluation de cette méthode.

Entité nommée	**Gazetteers**	**Marqueurs lexicaux**
Personne	27480	200
Lieu	4036	200
Organisation	17237	100

Tableau 5.3: Caractéristiques des données linguistiques.

Une deuxième méthode à base d'apprentissage implémentée à l'aide de l'outil linguistique Farasa et écrite en JAVA. Cette méthode utilise les modèles Champs Aléatoires Conditionnels (CRC) pour l'apprentissage supervisé des structures séquentielles des entités nommées. De plus, elle intègre un corpus d'apprentissage très large issu des ressources externes telles que : Arabic Wikipédia et DBpedia.

5.3.1.2 Résultats obtenus

Lors de l'expérimentation, nous avons testé la performance de ces deux méthodes dans le but de choisir la plus adéquate pour la reconnaissance des entités nommées. Les résultats de l'évaluation sont présentés dans le tableau ci-dessous.

Méthode	Entité nommée	Précision	Rappel
Méthode à base des régles	Personne	0.83	0,79
	Lieu	0,80	0.77
	Organisation	0.82	0.8
Méthode à base d'apprentissage	Personne	0.88	0.83
	Lieu	0.90	0.92
	Organisation	0,86	0.83

Tableau 5.4: Comparaison de performances des méthodes existantes et de la méthode proposée.

D'après les résultats obtenus dans le tableau 5.4, nous constatons que la méthode à base d'apprentissage est mieux adaptée pour la tâche de reconnaissance des entités nommées. Elle fournit un taux de précision et de rappel plus élevé que la méthode à base des règles pour l'ensemble des entités nommées. Ceci est dû à l'efficacité et la robustesse des techniques d'apprentissage par rapport l'hétérogénéité et la variété de contenu des journaux télévisés. Pour cette raison, nous avons utilisé cette méthode dans le cade de notre travail. Nous cherchons à automatiser la tâche de reconnaissance des entités nommées pour pouvoir l'intégrer dans l'architecture de notre système d'indexation (voir Annexe 4).

5.3.2 Cas de la facette événementielle

5.3.2.1 Mise en œuvre

la contribution principale pour cette facette consiste à construire un dictionnaire des déclencheurs morphologiques et sémantiques pour chaque évènement. Ce dictionnaire fa-

cilite et optimise le processus d'extraction des évènements.

En effet , le repérage d'évènement dans le texte de la vidéo ne se fait pas seulement sur la correspondance morphologique entre le nom de l'évènement et les autres mots de texte mais implique aussi les mots sémantiquement proches de l'évènement comme étant des déclencheurs sémantiques (tableau 5.5). La figure 5.11 décrit le nombre des déclencheurs pour chaque évènement.

Id	Evènement	Id	Evènement
1	انتخابات	9	غارة جوية
2	اجتماع	10	هجوم إرهابي
3	مصادقة	11	اغتيال
4	إضراب	12	حرب
5	مظاهرات	13	حادث مرور
6	هجرة غير شرعية	14	مواجهات عسكرية
7	اعتقال	15	بطولة رياضية
8	انفجار	16	سباق رياضي

Tableau 5.5: Liste des évènements

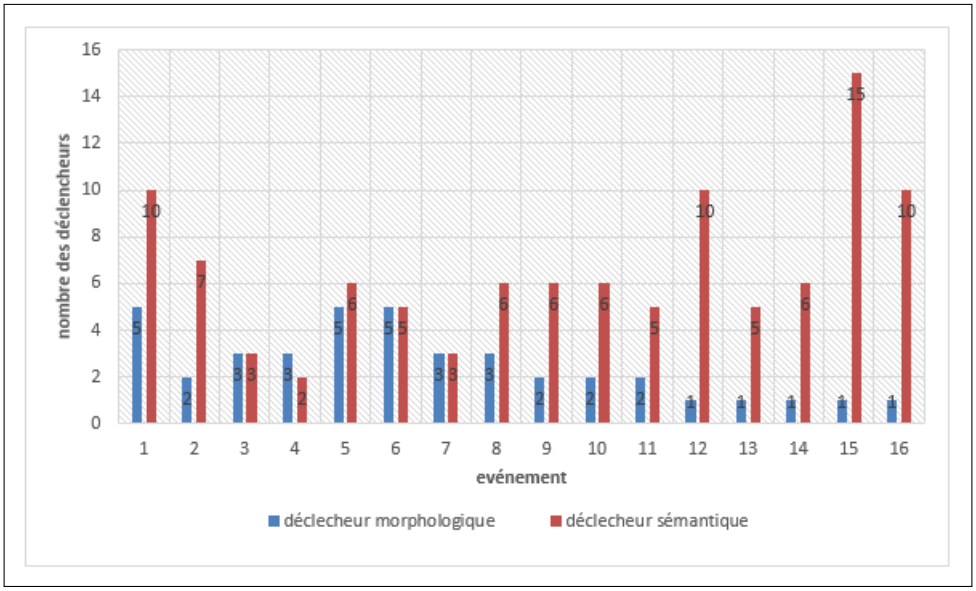

Figure 5.11: Nombre Des déclencheurs pour chaque évènements

Pour évaluer la performance de la facette événementielle , nous avons testé trois méthodes :

Méthode 1 : Cette méthode utilise uniquement le nom de l'évènement comme déclencheur principale dans la phase d'extraction. Un exemple d'application est donné

dans le Figure 5.12.

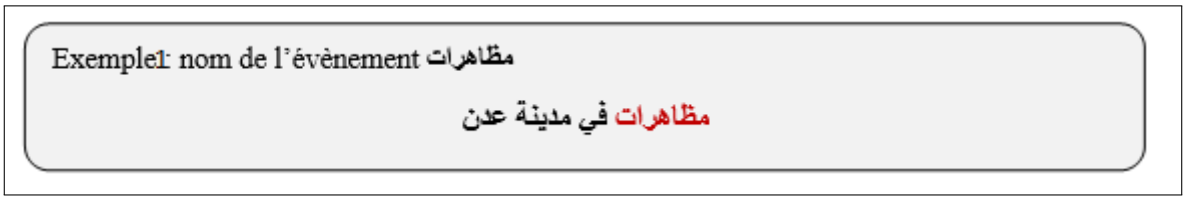

Figure 5.12: Exemple d'application de la méthode 1

Méthode 2 : Cette méthode est basée sur les déclencheurs morphologiques qui sont définit dans notre première proposition pour l'extraction des évènements. Un exemple d'application est donné dans le Figure 5.13.

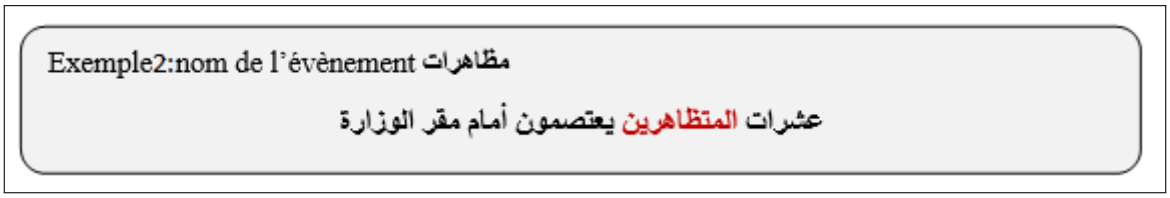

Figure 5.13: Exemple d'application de la méthode 2

Méthode 3 : Cette méthode représente notre proposition finale qui utilise les déclencheurs sémantiques et morphologiques. Un exemple d'application est donné dans le Figure 5.14.

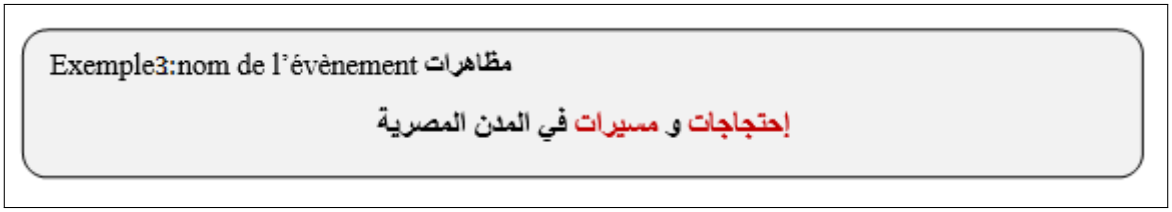

Figure 5.14: Exemple d'application de la méthode 3

5.3.2.2 Résultats obtenus

Méthode	**Précision**	**Rappel**
Méthode 1	0.75	0.5
Méthode 2	0.8	0.66
Méthode 3	0.84	0.9

Tableau 5.6: Comparaison de performance des méthodes proposées pour la facette événementielle.

Les résultats illustrés dans le tableau 5.6 montrent que le taux de rappel augmente de 0.5 vers 0.9. Ceci traduit l'efficacité de la méthode 3 qui implique plus de sémantique dans la phase d'extraction des évènements et montre bien que la méthode proposée est capable de détecter et extraire la majorité des évènements. Cependant, l'augmentation de taux de précision est assez moyenne par rapport au rappel. Ceci est dû à la présence de certaines erreurs d'extraction liées à la contexte d'apparition des évènements. La figure 5.15 montre quelques cas d'erreurs.

Figure 5.15: Exemples d'erreurs d'extraction des évènements

5.3.3 Cas de la facette thématique

5.3.3.1 Mise en œuvre

L'apport de la méthode proposée dans la facette thématique réside dans l'utilisation des marqueurs thématiques pour une classification efficace de texte court incrusté dans le journal télévise.En effet, la pondération ne dépend pas de la fréquence d'occurrence de déscripteur dans le corpus mais plutôt par sa valeur sémantique par rapport au thème. Il s'agit de construire un dictionnaire des entités sémantiques utilisés comme étant des déclencheurs thématiques(tableau 5.7).

5.3.3.2 Résultats obtenus

Afin de comparer notre model de présentation de théme vis-à-vis de l'état de l'art, nous avons testé aussi la méthode de sac de mot (bag of word BOW) qui consiste à représenter chaque thème par une liste de mots les plus fréquents sans tenir en compte les entités nommées et les événements. La Figure 5.16 décrit la performance globale de ces deux modèles en termes de rappel, précision et F-mesure.

Thème	Nombre des documents dans lescorpus d'entrainement	Nombre des marqueurs agentifs	Nombre des marqueurs évènementiels
Politique	480	350	20
Sport	362	342	15
Militaire	256	220	20
Sociale	230	180	10
Totale	1328	1092	65

Tableau 5.7: Dictionnaire des marqueurs thématiques

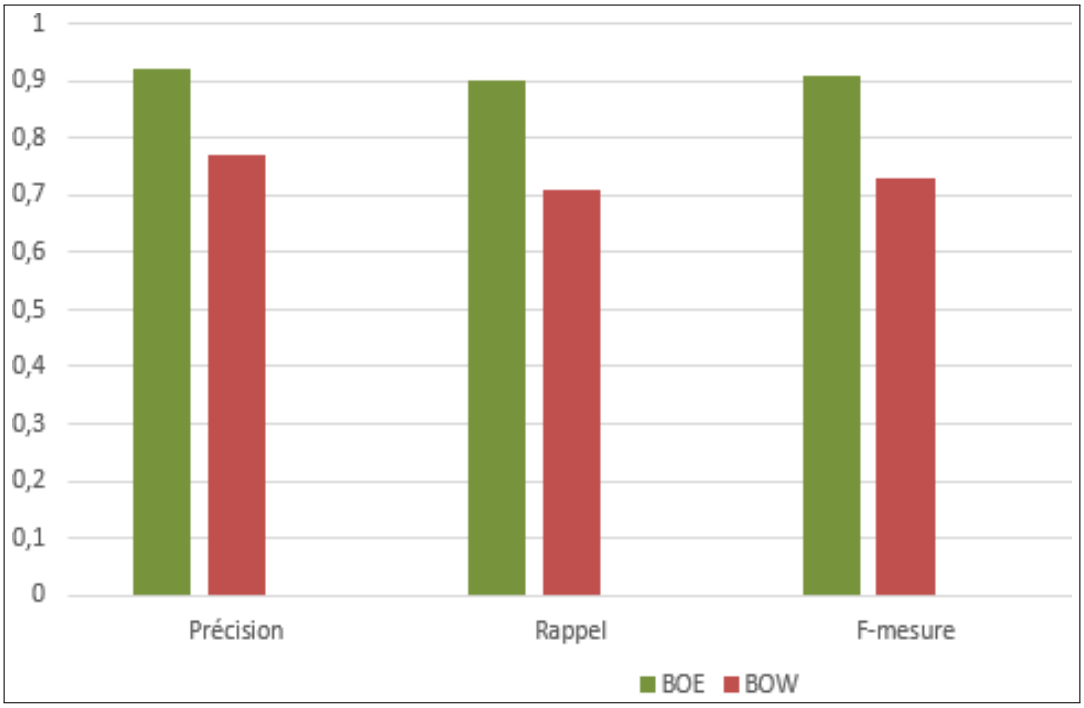

Figure 5.16: Histogramme des précisions, rappel et F-mesure globale

D'après les résultats obtenus, nous observons que la représentation des textes à l'aide des entités sémantiques permet d'augmenter nettement la précision (passant de 0,77 à 0,92) et le rappel (passant de 0,71 à 0,9) du modèle et accroît également la mesure du F-mesure (passant de 0,73 à 0 ,91). Par conséquent, nous constatons que le modèle de

représentation BOT améliore de façon significative les performances de la classification thématique de texte notamment dans le cadre des journaux télévisés.

D'autre part, ce modèle permet de ressouder le problème d'ambiguïté sémantique de termes : un terme peut être associé à plusieurs thème : termes polysémiques.

Figure 5.17: Résolution d'ambiguïté sémantique de termes.

5.3.4 Évaluation globale

Facette	Précision	Rappel	F-mesure
Facette conceptuelle	0.88	0.86	0.87
Facette événementielle	0.84	0.9	0.86
Facetté thématique	0.92	0.9	0.91
Moyenne	0.87	0.88	0.88

Tableau 5.8: Évaluation globale.

L'évaluation globale du système SISAVIN a montré qu'il est capable des fournir des annotations pertinentes en termes de trois facettes (valeurs moyenne de F-mesure dépassant 0,88 pour les trois facettes). Nous avons remarqué aussi que la valeur de F-mesure est relativement élevée notamment pour la facette thématique (0,91). Ceci prouve l'efficacité de notre modèle de représentation texte qui consiste à intégrer les entités nommées et les événements dans le vecteur descriptif de chaque thème.

Conclusion

Dans ce chapitre, nous avons décrit l'ensemble des expérimentations menées pour évaluer les différentes méthodes proposées dans le cadre d'un système d'indexation sémantique des journaux télévisés en langue arabe, nommé SISAVIN. Ce système est doté d'une architecture modulaire et repose sur les techniques de traitement bas niveau et l'extraction des informations sémantiques à travers une modélisation multifacette de contenu vidéo. Les résultats obtenus lors de l'évaluation de la partie traitement bas niveau prouvent l'efficacité et la robustesse de l'approche proposée pour la détection et la reconnaissance du texte incrusté dans la vidéo. Ensuite, nous avons effectué une évaluation globale de notre système en terme d'exactitude des indexes sémantiques générées selon les trois facettes proposées. Les résultats obtenus montrent la performance de système SISAVIN à fournir une description efficace de contenu vidéo.

Chapitre

6 Conclusion Générale

6.1 Synthèse des contributions

Les travaux menés dans cette thèse s'intègrent dans le cadre d'indexation sémantique des vidéos. Il s'agit particulièrement de proposer un nouveau modèle de représentation de contenu afin d'améliorer l'indexation des journaux télévisés en langue arabe.
Pour atteindre cet objective, nous avons commencé par la définition des concepts de base inhérents de document audiovisuel. Nous avons évoqué également les spécificités de texte arabe ainsi que les problèmes liés aux techniques d'extraction des informations. En deuxième lieu, nous avons détaillé les approches d'indexation sémantique des vidéos proposées dans la littérature, ainsi que les systèmes existants.
Nous avons ensuite présenté notre approche pour l'indexation sémantique des vidéos en langue arabe. L'originalité de cette approche consiste à proposer un modèle de représentation de contenu permettant de passer d'une représentation numérique (bas niveau) vers une description sémantique de contenu vidéo en exploitant le texte incrusté comme une source d'information. Précisément, nos contributions consistent à :

— Proposer une méthode heuristique pour la détection de texte incrusté dans la vidéo. Premièrement, cette méthode combine deux caractéristiques visuelles de texte (la couleur et le contour) pour extraire les régions candidates. La deuxième étape vise à améliorer le résultat de détection initiale à travers un processus de filtrage basé sur des règles heuristiques. Ces règles sont inspirées à partir les caractéristiques géométriques de texte arabe. En comparaison avec les autres méthodes existantes (hybrides et à base d'apprentissage), cette méthode trouve son originalité, d'une part dans le paramétrage adopté au corpus de vidéos. Ceci présente un avantage intéressant notamment lors de l'ajout des nouvelles vidéos avec des caractéristiques

différentes. D'autre part, notre méthode n'exige pas une phase d'apprentissage (rapidité d'exécution). Ceci offre la possibilité de gérer les archives des journaux télévisés qui sont en croissance rapide chaque jour.

— La deuxième contribution réside dans la proposition d'un modèle multifacette pour la représentation sémantique de contenu vidéo. Ce modèle introduit trois facettes de modélisation. La première facette permet d'extraire les concepts de type personne et organisation et lieu. La deuxième facette fournit une information sur la thématique de contenu. Alors que la dernière facette permet d'extraire la liste des événements. Ce modèle fournit une représentation abstraite et bien structurée de contenu vidéo permettant de faciliter sa compréhension et son indexation.

— Une troisième contribution concerne la proposition d'un nouveau modèle de présentation de texte court. La particularité de ce modèle réside dans l'utilisation des entités sémantique comme des éléments descriptifs de chaque catégorie. Ceci permet de fournir une représentation plus sémantique vis-à-vis aux caractéristiques spécifiques de texte court dans le but d'améliorer les processus de classification thématique de contenu vidéo.

— Au niveau expérimental, notre contribution se manifeste par le développement du système SISAVIN. Celui-ci regroupe tous les modules nécessaires pour l'indexation sémantique des journaux télévisés, allant de l'extraction des images clés jusqu'à la génération des indexes finals sous forme des fichiers xml.

6.2 Perspectives

Les résultats obtenus lors de l'évaluation sont globalement satisfaisants pour les différentes méthodes proposées. Mais, il reste encore des améliorations possibles pour ce travail de recherche. Les perspectives envisagées concernent trois niveaux :

— Niveau analyse de vidéo : vu la richesse et la variété de contenu sémantique du journal télévisé, il est intéressant de proposer une description multimodale de contenu vidéo en langue arabe. Il s'agit d'analyser et combiner les informations issues de différentes modalité (audio, visuel, texte) pour une description plus riche. Ceci permet d'interpréter le contenu en termes d'actions, des concepts visuels et des concepts

audio à fin d'améliorer la modélisation des vidéos.

— Niveau linguistique : la désambiguïsation des entités nommées présente aussi une perspective intéressante dans le cadre d'indexation sémantique. En effet, le traitement des formes polysémique d'une entité nommée permet de résoudre l'ambiguïté sémantique et de fournir une annotation fine selon le contexte d'utilisation tel est le cas du texte de la figure 6.1.

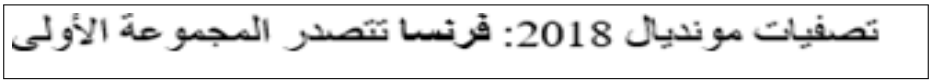

Figure 6.1: Extrait de texte

Il serait très important de distinguer entre la France tant que nom de lieu ou nom de club sportif en termes d'indexation et de recherche d'information sémantique.

— Niveau expérimental, nous envisageons de tester le système développé dans le contexte de rechercher d'information. Dans ce cadre, nous pouvons développer une interface graphique qui permet de répondre efficacement aux besoins des utilisateurs en exploitant les résultats fournis par notre système d'indexation.

Bibliographie

[Can86] J. Canny. *A computational approach to edge detection.* IEEE Trans. Patt. Anal. Mach. Intel., PAMI-8 , pp. 679-698, 1986.

[AK95] Hobbs Appelt et Kameyama. *FASTUS system : MUC-6 test results and analysis.* In Proceedings of the 6th conference on Message understanding, MUC-6, pages 237–248, Stroudsburg,PA, USA. Association for Computational Linguistics, 1995.

[KKi96] H. K.Kim. *Efficient automatic text location method and content based indexing and structuring of video database.* In Journal of visual communication et image representation, 1996.

[AR98] Halverson Aone et Ramos-Santacruz. *SRA : Description of the IE2 system used for MUC-7.* In Proceedings Seventh Message Understanding Conference (MUC-7), 1998.

[KH99] E.Kim K.Jeong K.Jung et H.Kim. *Neural network-based text location for news video indexing.* In IEEE International Conference on Image Processing (ICIP),pp :319-323, 1999.

[AR00] C. Aone et Ramos-Santacruz. *REES : A large-scale relation and event extraction system.* In Applied Natural Language Processing Conference (ANLP), pages 76–83, 2000.

[Dem00] Claire-Hélène Demarty. *Segmentation et structuration d'un document vidéo pour la caractérisation et l'indexation de son contenu sémantique.* Thèse de doctorat, 2000.

[FDi00] F.Dirfaux. *Key frame selection to represent a video.* In International Conference on Image Processing, 2000.

[HK00] D. Doermann H. Li et O. Kia. *Automatic text detection and tracking in digital video.* IEEE Transactions on Image Processing,9(1) :147–156, 2000.

[HM00] Riad Hammoud et Roger Mohr. *A probabilistic framework of selecting effective key frames for video browsing and indexing.* In International workshop on Real-Time Image Sequence Analysis(RISA '00), pages 79–88, Oulu, Finlande, 2000.

[SW00] Arnold W. M. Smeuldersz et Marcel Worring. *Content-based image retrieval at the end of the early years.* IEEE Trans. Pattern Anal. Mach. Intell.,22(12) :1349–1380, 2000.

[YL00] S.-H. Choi Y.-K. Lim et S.-W. Lee. *Text extraction in mpeg compressed video for content-based indexing.* In International Conference on Pattern Recognition, volume 4, pages 409–412, 2000.

[JP02] M. Urban J. Matas O. Chum et T. Pajdla. *Robust wide baseline stereo from maximally stable extremal regions.* In procedding of British Machine Vision Conference, pages 384–396, 2002.

[JA02] Lori Lamel Jean-Luc Gauvain et Gilles Adda. *The LIMSI Broadcast News Transcription System.* In Speech Communication, pages 89–108, 2002.

[Chi03] Chieu.H. *Closing the gap : Learning-based information extraction rivaling knowledge-engineering methods.* In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 216–223, 2003.

[DK03] S. Antani D. Crandall et R. Kasturi. *Extraction of special effects caption text events from digital video.* International Journal on Document Analysis et Recognition, 5(2-3) :138–157, 2003.

[FC03] O. Riff F. Pelisson D. Hall et J. Crowley. *Brand identification using gaussian derivative histograms.* In proceeding of International Conference on Vision Systems, Graz, Austria, 2003.

[WK03] K. Wang et J.A. Kangas. *Character location in scene images from digital camera.* Pattern recognition, 36(10) :2287–2299, 2003.

[ZKP03] Z.Cernekova, C. Kotropoulos et I. Pitas. *Video shot segmentation using singular value decomposition.* In Proc. of the IEEE International Conference on Multimedia et Expo, ICME, pages 301-304, 2003.

[Abu04] Saleem Abuleil. *Extracting names from Arabic text for question-answering systems.* In Proceedings of the 7th International Conference on Coupling Approaches, Coupling Media, et Coupling Languages for Information Retrieval (RIAO), pages 638–647, 2004.

[LG04] Lori Lamel Leonardo Canseco et Jean-Luc Gauvain. *speaker diarization from speech transcripts.* In the 5th AnnualConference of the International Speech Communication Association,INTERSPEECH, 2004.

[WC04] X. Chen W. Wu et J. Chang. *Incremental detection of text on road signs from video with application to a driving assitant system.* ACM Multimedia, 2004.

[WZ04] W.Tavanapong et J. Zhou. *Shot clustering techniques for story browsing.* IEEE Transactions on Multimedia, 517–527, 2004.

[ZX04] Z. Mingli Z. Sanyuan et Y. Xiuzi. *Car plate character extraction under complicated environment.* In proceeding of IEEE International Conference on Systems, Man et Cybernitics, pp 4722-4726., 2004.

[B B05] A. BEN Hamadou B. Bouaziz W. Mahdi. *Automatic Text Regions Location in Video Frames.* The IEEE International conference on signal image technology et internet-based system., 2005.

[Cha05] Quénot G Charhad G. *Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés.* le colloque CORESA (Compression et Représentation des Signaux Audiovisuels), 2005.

[Eri+05] Erik et al. *Learning Hierarchical Models of Scenes, Objects, and Parts.* In Proceedings of the Tenth IEEE International Conference on Computer Vision – Volume 2, pages 1331–1338, Washington, DC, USA, 2005.

[LG05] Lori Lamel Leonardo Canseco et Jean-Luc Gauvain. *A Comparative Study Using Manual and Automatic Transcriptions for Diarization.* In IEEE Workshop on Automatic Speech Recognition et Understanding, pages 415–419, 2005.

[MCh05] M.Charhad. *Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique.* Thèse de doctorat, 2005.

[RD05] Marco Cammisa Roberto Basili et Emanuale Donati. *RitroveRAI : A Web application for semantic indexing and hyperlinking of multimedia news.* In Proceedings of International Semantic Web Conference (ISWC), pages 97–111, 2005.

[SH06] M. Saharni et Heilman. *A Web-based Kernel Function for Measuring the Sirnilarity of Short Text Snippets.* In Proceedings of the 15th International Conference on World Wide Web, WWW '06 , 377- 386., 2006.

[Tra06] Sue E. Tranter. *Who really spoke when ? finding speaker turns and identities in broadcast news audio.* In the 31st IEEE International Conference on Acoustics, Speech et Signal Processing, ICASSP, pages : 1013–1016, 2006.

[BR07] Yassine Benajiba et Paolo Rosso. *ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information.* In Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007), pages 1,814–1,823, 2007.

[BY07] Matsuo Bollegala D et Y.hizuka. *Measuring Semantic Similarity Between Words Using Web Search Engines.* In Proceedings of the 16th International Conference World Wide Web, WWW '07 ' 757- 766., New York ' NY, USA. ACM, 2007.

[CM07] Patrick Nguyen Chengyuan Ma et Mahajan Milind. *Finding speaker identities with a conditional maximum entropy model.* In the 32nd IEEE International Conference on Acoustics, Speech et Signal Processing, ICASSP,pages 261–264, 2007.

[MP07] B. Gatos M. Anthimopoulos et I. Pratikakis. *Multiresolution text detection in video frames.* n The International Conference on Computer Vision Theory et Applications (VISAPP), pages 161–166, 2007.

[MS07] Durnais Metzler D. et S.Meek. *Similarity Measures for Short Segments of Text.* In Advances in Information Retrieval. ECIR . Lecture Notes in Computer Science, vol 4425, 2007.

[Sli07] Mesfar Slim. *Named entity recognition for Arabic using syntactic grammars.* In Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB), pages 305–316, 2007.

[YL07] Changsheng Xu Yifan Zhang et Hanqing Lu. *Semantic event extraction from basketball games using multimodal analysis.* In Proceedings of the IEEE Conference on Multimedia et Expo (ICME), pages 2190–2193, 2007.

[B B08] A. BEN Hamadou B. Bouaziz W. Mahdi. *Content-Based Video Browsing by Text Region Localization and Classification.* International Journal of Video & Image Processing et Network Security IJVIPNS-IJENS ., 2008.

[BPP08] Behmo, Paragios.N et Prinet.V. *Graph commute times for image representation.* In Proceedings of IEEE Conference on Computer Vision et Pattern Recognition, 2008.

[BR08] Yassine Benajiba et Paolo Rosso. *Arabic named entity recognition using optimized feature sets.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pages 284–293, 2008.

[DG08] M. Delakis et C. Garcia. *text detection with convolutional neural networks.* In the International Conference on Computer Vision Theory et Applications, pages 290–294, 2008.

[NH08] D. Nist´er et H.Stew´enius. *Linear time maximally stable extremal regions.* In Computer Vision–ECCV, pages 183–196., 2008.

[Par08] Muller P Parent G. Gagnon M. *Annotation d'expressions temporelles et d'événements en français.* In Traitement Automatique des Langues Naturelles (TALN), 2008.

[S08] Mesfar S. *Analyse morphosyntaxique automatique et reconnaissance des entités nommées arabe standard*. Thèse de doctorat, 2008.

[BR09] Mona Diab Benajiba Yassine et Paolo Rosso. *Arabic named entity recognition : A feature-driven study*. IEEE Transactions on Audio, Speech, et Language Processing, 17(5) :926–934, 2009.

[KK09] W. Kim et C. Kim. *A new approach for overlay text detection and extraction from complex video scene.* IEEE Transactions on Image Processing, 18(2) :401–411, 2009.

[MD09] Hamish Cunningham Maynard Diana et Marin Dimitrov. *A multilingual named entities corpus for Arabic, English and French*. In Proceedings of the Second International Conference on Arabic Language Resources et Tools, pages 213–216, 2009.

[MC09] Stéphane Chaudiron Mostefa Djamel et Gael de Chalendar. *A multilingual named entities corpus for Arabic, English and French*. In Proceedings of the Second International Conference on Arabic Language Resources et Tools, pages 213–216, 2009.

[SR09a] Khaled Shaalan et Hafsa Raza. *NERA : Named entity recognition for Arabic.* Journal of the American Society for Information Science et Technology, 60(8) :1,652–1,663, 2009.

[SR09b] Khaled Shaalan et Hafsa Raza. *NERA : Named entity recognition for Arabic.* Journal of the American Society for Information Science et Technology, 60(8) :1,652–1,663, 2009.

[AH10] Ben Hamadou Abdelmajid et Fehri Hela. *Recognition and Arabic-French translation of named entities : Case of the sport places.* In arXiv, pages 1–10, 2010.

[AD10] Ahmed Abdul-Hamid et Kareem Darwish. *Simplified feature set for Arabic named entity recognition.* In Proceedings of the Named Entities Workshop (NEWS 2010), pages 110–115, 2010.

[BW10] E. Ofek B. Epshtein et Y. Wexler. *Detecting text in natural scenes with stroke width transform.* In International Conference on Computer Vision et Pattern Recognition (CVPR), pages 2963–2970, 2010.

[HN10] Estela Saquete Hector Llorens et Borja Navarro. *TIPSem (English and Spanish) : Evaluating CRFs and Semantic Roles in TempEval-2.* In Proceedings of the 5th International Workshop on Semantic Evaluation, ACL, pages 284–291, Uppsala, Sweden, 15-16, 2010.

[MP10] B. Gatos M. Anthimopoulos et I. Pratikakis. *A two-stage scheme for text detection in video images.* Image et Vision Computing, 28(9) :1413–1426, 2010.

[MO10] S. Mozaffari M. Moradi et A.A. Orouji. *Farsi/arabic text extraction from video images by corner detection.* In Iranian Conference on Machine Vision et Image Processing, pages 1–6, 2010.

[ZS10] Bruno Pouliquen Zaghouani Wajdi et Ralf Steinberger. *Adapting a resource-light highly multilingual named entity recognition system to Arabic.* In Proceedings of the Seventh International Conference on Language Resources et Evaluation (LREC), pages 563–567, 2010.

[GN11] Sakamoto Gene et Nickerson. *Discovering Context : Classifying Tweets through a Semantic Transform Based on Wikipedia.* In International Conference on Foundations of Augmented Cognition, pp 484- 492, 2011.

[YT11] C. Yi et Y. Tian. *Text string detection from natural scenes by structure-based partition and grouping.* IEEE Transactions on Image Processing, 20(9) :2594–2605, 2011.

[Ali12] H. Karra Alimi. M. Ben Halima. *Nfsavo :Neuro-fuzzy system for arabic video ocr.* International Journal of Advanced Computer Science et Applications, 3(10) :128–136, 2012.

[BM12] Denise DiPersio Bies Ann et Mohamed Maamouri. *Linguistic resources for Arabic machine translation : The Linguistic Data Consortium (LDC).* Challenges for Arabic Machine Translation of Natural Language Processing., 2012.

[AG12] Paloma Al-Jumaily Harith et Erik Goo. *A real time named entity recognition system for Arabic text mining*. Language Resources et Evaluation Journal, 46(4) :543–563, 2012.

[Cha13] Mounir Zrigui Chahira Lhioui Anis Zouaghi. *A combined method based on stochastic and linguistic paradigm for the understanding of arabic spontaneous utterances*. International Conference on Intelligent Text Processing et Computational Linguistics, 2013.

[J13] Poignant J. *Identification non-supervisée de personnes dans les fluxtélévisés.* Thèse de doctorat, 2013.

[KDY13] Küçük, D. et Yazici. *A Semi-Automatic Text-Based Semantic Video Annotation System for Turkish Facilitating Multilingual Retrieval*. Expert Systems with Applications. Volume 40, Issue 9, pp. 3398–3411, 2013.

[OS13] Mai Oudah et Khaled Shaalan. *Person name recognition using the hybrid approach*. In Natural Language Processing et Information Systems, volume 7934 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pages 237–248, 2013.

[Poi13] Bredin H. et Poignant J. *Integer linear programming for speaker diarization and cross-modal identification in tv broadcast*. Proceedings of InterSpeech, 2013.

[SZ13] Anis Zouaghi Souheyl Mallat et Mounir Zrigui. *Method of lexical enrichment in information retrieval system in Arabic*. In International Journal of Information Retrieval Research (IJIRR). Vol (3), N°4, pp : 35-51, 2013.

[BT14] Vincent Claveau Béatrice Arnulphy et Xavier Tannier. *Supervised Machine Learning Techniques to Detect TimeML Events in French and English.* In Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB),pages : 19-32, 2014.

[XH14] Kaizhu Huang Xu-Cheng Yin Xuwang Yin et Hong-Wei Hao. *Robust Text Detection in Natural Scene Images.* IEEE Transactions on Pattern Analysis et Machine Intelligence (Volume : 36, Issue : 5), 2014.

[Dri14] Houssem Eddine Dridi. *Detection d'évènements à partir de Twitter*. Thèse de doctorat, 2014.

[FW14] Zhoujun Li Fang Wang Zhongyuan Wang et Ji-Rong Wen. *Concept-based Short Text Classification and Ranking*. CIKM, 2014.

[SG14] S.-A. Berrani S. Yousfi et C. Garcia. *Arabic text detection in videos using neural and boosting-based approaches : Application to video indexing*. In International Conference on Image Processing (ICIP), pages 3028–3032, 2014.

[WT14] Y. Qiao W. Huang et X. Tang : *Robust scene text detection with convolution neural network induced mser trees*. In Computer Vision–ECCV 2014, pages 497–511, 2014.

[FV15] Stilo Faralli et Velmdi. *What women like : A gendered analysis of twitter uers ' interests based on a twixonomy*. In Ninth International AAAI Conference on Web et Social Media, 2015.

[Gay15] Paul Gay. *Segmentation et identification audiovisuelle de personnes dans des journaux télévisés. (Audiovisual segmentation and identification of persons in broadcast news)*. Thèse de doctorat, 2015.

[Nai15] Mounir Zrigui Naim Terbeh Mohsen Maraoui. *Probabilistic approach for detection of vocal pathologies in the arabic speech*. International Conference on Intelligent Text Processing et Computational Linguistics, 2015.

[OH15] Sameh Masmoudi Touj Oussama Zayene Mathias Seuret et Jean Hennebert. *Data, protocol and algorithms for performance evaluation of text detection in Arabic news video*. In the International Conference on Advanced Technologies for Signal et Image Processing, pp : 258-263, 2015.

[RZ15] M. Maraoui R. Ayadi et M. Zrigui. *Arabic Text Representation and Classification Methods : Current State of the Art*. In the International Conference on Artificial Intelligence (ICAI), Las Vegas, Nevada, USA, 2015.

[Sou15] Anis Zouaghiand Mounir Zrigui Souheyl Mallat. *Use of external resources in a information retrieval system in Arabic*. In 17th International Conference on Artificial Intelligence (ICAI), pages 27-30, 2015.

[TZ15] Fériel Trabelsi et Chiraz Ben Othmane Zribi. *Combined Classification for Extracting Named Entities from Arabic Texts.* First International Conference on Arabic Computational Linguistics (ACLing), 2015.

[Emn16] Mounir Zrigui Emna Hkiri Souheyl Mallat. *Events Automatic Extraction from Arabic Texts.* Zrigui : Events Automatic Extraction from Arabic Texts. In International Journal of Information Retrieval Research, IJIRR 6(1) : 36-51, 2016.

[EZ16] Souheyl Mallat Emna Hkiri et Mounir Zrigui. *Events Automatic Extraction from Arabic Texts.* IJIRR 6(1) : 36-51, 2016.

[Lhi16] Zrigui M Lhioui C Zouaghi A. *Knowledge Extraction with NooJ Using a syntactico-Semantic Approach for the Arabic Utterances Understanding.* In Computational Linguistics et Intelligent Text Processing, 2016.

[Moh16] Mounir Zrigui Mohamed Labidi Mohsen Maraoui. *New birth of the Arabic phonetic dictionary.* 2016 International Conference on Engineering MIS (ICEMIS), 2016.

[OH16] Sameh Masmoudi Touj Oussama Zayene Mathias Seuret et Jean Hennebert. *Text Detection in Arabic News Video Based on SWT Operator and Convolutional Auto-Encoders,* International Workshop on Document Analysis Systems : 13-18, 2016.

[R16] Ayadi R. *Conception et réalisation d'une plate-forme d'indexation et de classification de documents : cas de l'arabe.* Thèse de doctorat, 2016.

[Ram16] Mounir Zrigui Rami Ayadi Mohsen Maraoui. *A Survey of Arabic Text Representation and Classification Methods.* Res. Comput. Sci., 2016.

[Adn17] Mounir Zrigui Adnen Mahmoud Ahmed Zrigui. *A text semantic similarity approach for Arabic paraphrase detection.* International conference on computational linguistics et intelligent text processing, 2017.

[Cha17] Mounir Zrigui Chahira Lhioui Anis Zouaghi. *A rule-based semantic frame annotation of arabic speech turns for automatic dialogue analysis.* Procedia Computer Science, 2017.

[Emn17] Mounir Zrigui Emna Hkiri Souheyl Mallat. *Arabic-English text translation leveraging hybrid NER*. Proceedings of the 31st Pacific Asia Conference on Language, Information et Computation, 2017.

[EZ17] Souheyl Mallat Emna Hkiri et Mounir Zrigui. *Integrating Bilingual Named Entities Lexicon with Conditional Random Fields Model for Arabic Named Entities Recognition.* In International Conference on Document Analysis et Recognition (ICDAR) pages : 609-614, 2017.

[Moh17] Mounir Zrigui Mohamed Ali Batita. *The enrichment of arabic wordnet antonym relations*. International Conference on Computational Linguistics et Intelligent Text Processing, 2017.

[Adn18] Mounir Zrigui Adnen Mahmoud. *Artificial method for building monolingual plagiarized Arabic corpus*. Computación y Sistemas, 2018.

[Emn18] Mounir Zrigui Emna Hkiri Souheyl Mallat. *Enhancing deep learning gender identification with gated recurrent units architecture in social text*. Comp. y Sist.[online]. 2018, vol. 22, n. 3, pp. 757-766. ISSN 1405-5546. https ://doi. org/10.13053/cys-22-3-3036, 2018.

[Moh18] Mounir Zrigui Mohsen Maraoui Naim Terbeh. *Arabic discourse analysis based on acoustic, prosodic and phonetic modeling : elocution evaluation, speech classification and pathological speech correction*. International journal of speech technology, 2018.

[Adn19a] Mounir Zrigui Adnen Mahmoud. *Deep neural network models for paraphrased text classification in the Arabic language*. International Conference on Applications of Natural Language to Information Systems, 2019.

[Adn19b] Mounir Zrigui Adnen Mahmoud. *Sentence embedding and convolutional neural network for semantic textual similarity detection in arabic language.* Arabian Journal for Science et Engineering, 2019.

[Emn20] Mounir Zrigui Emna Hkiri Souheyl Mallat. *Events automatic extraction from Arabic texts*. Natural Language Processing : Concepts, Methodologies, Tools, et Applications, 2020.

[Moh20] Mounir Zrigui Mohamed Ali Sghaier. *Rule-Based Machine Translation from Tunisian Dialect to Modern Standard Arabic*. Procedia Computer Science, 2020.

[Naf20a] Mounir Zrigui Nafaa Haffar Emna Hkiri. *Enrichment of Arabic TimeML Corpus*. International Conference on Computational Collective Intelligence, 2020.

[Naf20b] Mounir Zrigui Nafaa Haffar Emna Hkiri. *Using Bidirectional LSTM and Shortest Dependency Path for Classifying Arabic Temporal Relations*. Procedia Computer Science, 2020.

Annexe

A Alphabet Arabe

L'alphabet arabe comporte 28 lettres qui s'écrivent de droite à gauche. Certaines lettres n'ont pas la même graphie lorsqu'elles sont placées au début, au milieu ou à la fin d'un mot.

De plus, les lettres s'attachent toutes entre elles sauf 6 qui ne s'attachent jamais à gauche. Ces dernières sont précédées d'une petite étoile rouge comme illustré dans la figure 1.1.

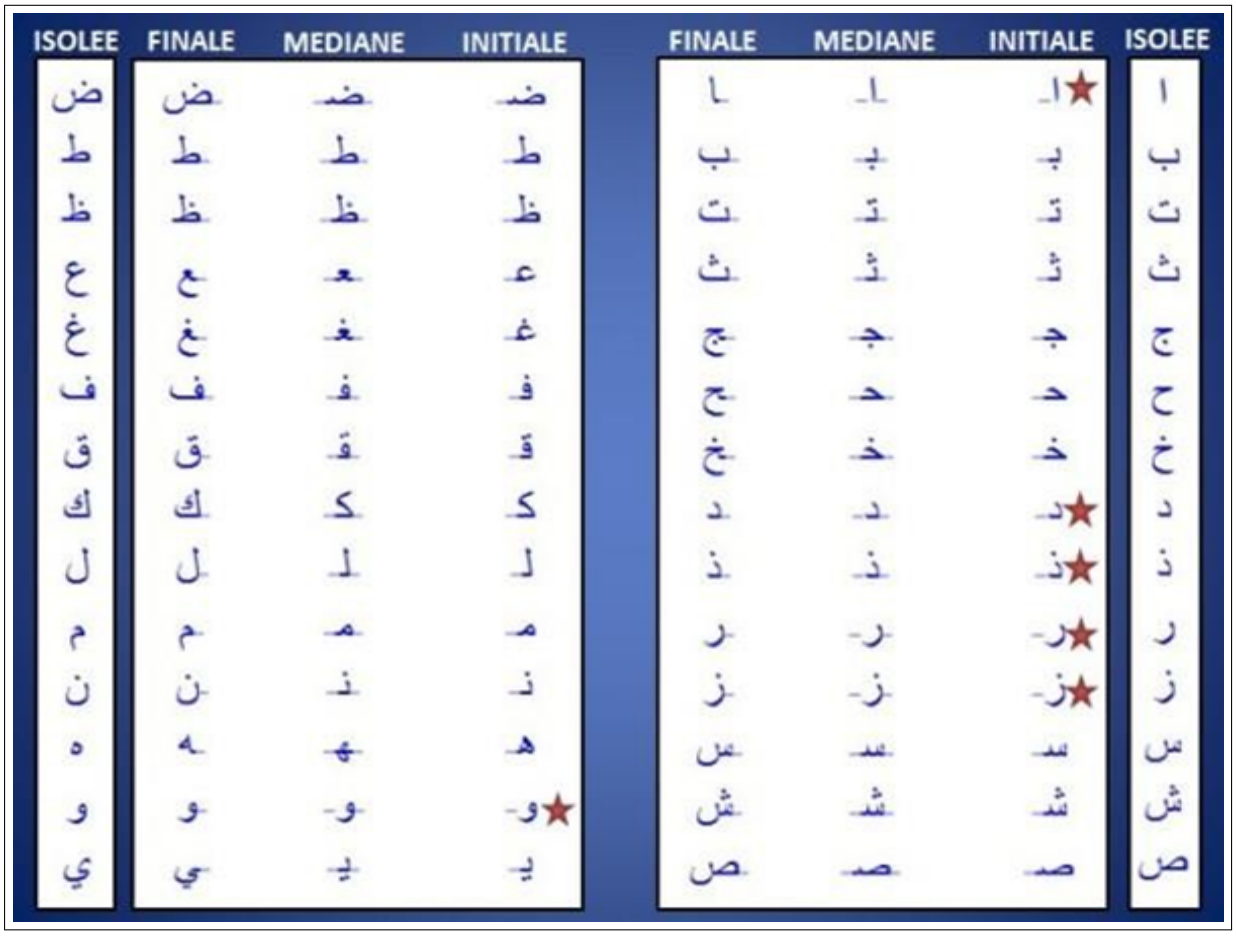

Figure 1.1: Alphabet Arabe

Annexe

B Opérateurs morphologiques

L'opérateur morphologique fermeture est défini comme suit :

$$\text{FermB(X)} = \text{ErosB(DilB(X))}$$

La fermeture de l'image X par l'élément structurant B est la composition de la dilatation par B suivie de l'érosion par B. Soit une image X et un élément structurant B, la dilatation de X par B est l'ensemble obtenu en remplaçant chaque pixel p de X par sa fenêtre Bp. L'effet de la dilatation est d'élargir l'image. En effet, la hauteur et largeur de la figure dilatée seront les sommes respectivement des hauteurs et largeurs de la figure originelle et de l'élément structurant. Si l'élément structurant est décentré, la dilatation décalera la figure dans le même sens. Enfin, les coins convexes de la figure seront déformés en fonction de l'élément structurant (par exemple si celui-ci est un disque, les coins convexes seront arrondis).

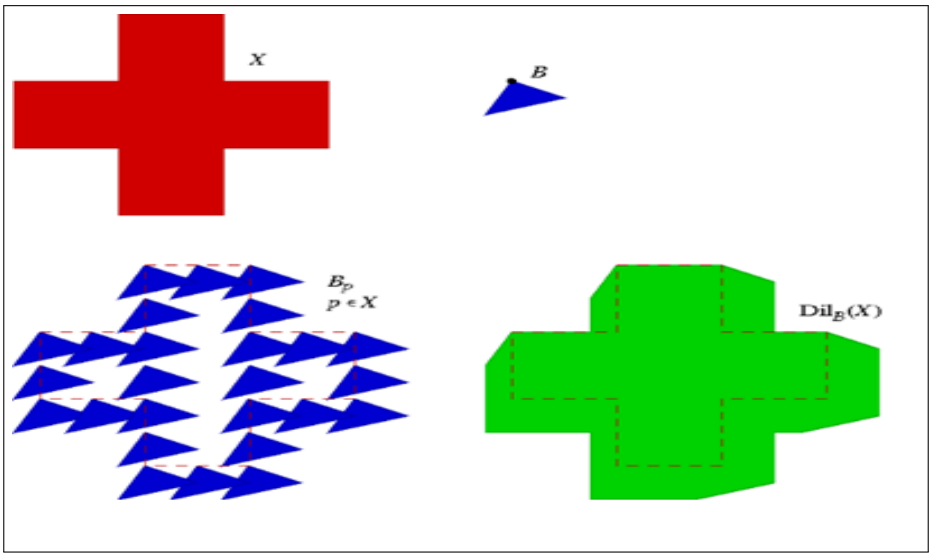

Figure 2.1: Un exemple illustratif de la dilatation

En haut à gauche, la figure X, et à droite, l'élément structurant B (la boule noire

indiquant la position du point de référence). En bas à gauche, les fenêtres Bp pour un certain nombre de pixels $p \in X$. En bas à droite, la dilatation de X par B ; les coins convexes de la figure deviennent biseautés par les côtés du triangle B.

L'effet de l'érosion est d'abord de rétrécir l'image, la hauteur et largeur de la figure érodée seront les différences respectivement des hauteurs et largeurs de la figure originelle et de l'élément structurant (en particulier si l'élément structurant est plus large ou plus haut que la figure, l'érosion de celle-ci sera vide). Si l'élément structurant est décentré, l'érosion décalera la figure en sens inverse. Enfin les coins concaves de la figure seront déformés en fonction de l'élément structurant (par exemple si celui-ci est un disque, les coins concaves seront arrondis).

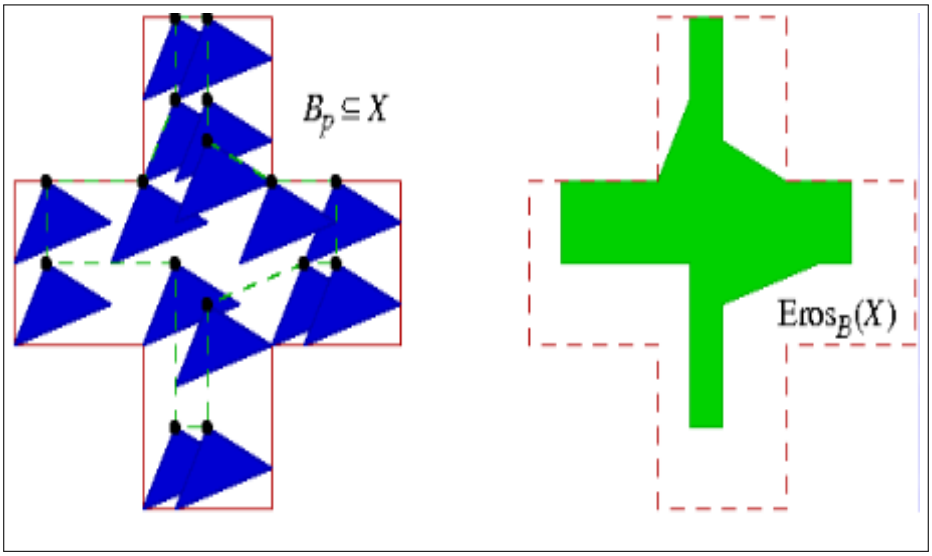

Figure 2.2: Un exemple illustratif de l'érosion

On considère la même figure X et élément structurant B que plus haut. À gauche, plusieurs positions (boules noires) de pixels p dont les fenêtres Bp sont incluses dans X. À droite, l'érosion de X par B ; les coins concaves de la figure deviennent biseautés par les côtés du triangle B. L'ouverture est définie comme une composition de l'érosion suivie par une dilatation :

$$\mathrm{OuvB(X)} = \mathrm{DilB(ErosB(X))}.$$

Annexe

C Intégration de Tesseract

```
package ocrf;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.io.InputStream;
import java.io.PrintStream;
import java.io.PrintWriter;
import java.util.logging.Level;
import java.util.logging.Logger;
public class OCRF {
public static void main(String[] args) throws IOException {
   File dir = new File("C:\\regions de texte");
   FileWriter ecrivain;
   FileWriter ocr= new FileWriter("E:\\Travail de thése\\base de video\\Arabic video news\\aljazzera\\video2\\Result-OCR.txt");
   String output_file="test1";
   String tesseract_install_path="C:\\Program Files (x86)\\Tesseract-OCR\\tesseract";
   String s[] = dir.list();

   for (int i=0; i<s.length; i++){
      System.out.println("fichier : " + s[i]);
      String[] command =
      {
         "cmd",
      };
      Process p;
```

```
        try {
            p = Runtime.getRuntime().exec(command);
            new Thread(new SyncPipe(p.getErrorStream(), System.err)).start();
            new Thread(new SyncPipe(p.getInputStream(), System.out)).start();
            String h="";
            PrintWriter stdin = new PrintWriter(p.getOutputStream());
            stdin.println("\""+tesseract_install_path+"\"  \""+dir+"\\"+s[i]+"\"  \""+output_file+"\"  -l
ara");
            stdin.close();
            p.waitFor();
            //System.out.println();
            // System.out.println();
            // System.out.println();
            // System.out.println();
            h=Read_File.read_a_file(output_file+".txt");
            System.out.println(Read_File.read_a_file(output_file+".txt"));
            String recup = s[i].substring(0, s[i].length()-4);
            //ecrivain = new FileWriter("E:\\Travail de thése\\base de video\\Arabic video
news\\aljazzera\\video14\\texte\\"+recup+".txt");
           // ecrivain.write(h);
            ocr.write(h+"\n");
           //  ecrivain.close();
            // System.out.println(h);
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
    ocr.close();
}
```

Annexe

D Intégration de FARASA

```
package ner;
import com.qcri.farasa.segmenter.Farasa;
import com.qcri.farasa.pos.FarasaPOSTagger;
import com.qcri.farasa.ner.ArabicNER;
import com.qcri.farasa.pos.Sentence;
import com.qcri.farasa.pos.Clitic;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
 class NER {
  public static void main (String[] args) throws IOException, FileNotFoundException,
ClassNotFoundException, UnsupportedEncodingException, InterruptedException, Exception {
      Farasa segmenter = new Farasa ();
      FarasaPOSTagger tagger = new FarasaPOSTagger(segmenter);
      ArabicNER ner = new ArabicNER(segmenter, tagger);
    BufferedReader lecteurAvecBuffer=null;
    FileWriter ner1=null;
```

```
package ner;
import com.qcri.farasa.segmenter.Farasa;
import com.qcri.farasa.pos.FarasaPOSTagger;
import com.qcri.farasa.ner.ArabicNER;
import com.qcri.farasa.pos.Sentence;
import com.qcri.farasa.pos.Clitic;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
 class NER {
  public static void main (String[] args) throws IOException, FileNotFoundException, ClassNotFoundException, UnsupportedEncodingException, InterruptedException, Exception {
       Farasa segmenter = new Farasa ();
       FarasaPOSTagger tagger = new FarasaPOSTagger(segmenter);
       ArabicNER ner = new ArabicNER(segmenter, tagger);
     BufferedReader lecteurAvecBuffer=null;
     FileWriter ner1=null;
```

```
File dir = new File("E:\\Travail de thése\\base de video\\Arabic video news\\Wataniya1");
String s[] = dir.list();
int k ;
for (int j=0; j<s.length; j++){
      String ligne;
k=j+1;
   try
    {


       lecteurAvecBuffer = new BufferedReader(new FileReader("E:\\Travail de thése\\base de video\\Arabic video news\\Wataniya1\\video"+k+"\\Result-OCR.txt"));
    ner1= new FileWriter("E:\\Travail de thése\\base de video\\Arabic video news\\Wataniya1\\video"+k+"\\Result-NER.txt");
    }
   catch(FileNotFoundException exc)
    {
        System.out.println("Erreur d'ouverture");
    }
   while ((ligne = lecteurAvecBuffer.readLine()) != null)
   {
    System.out.println(ligne);
    ArrayList output = ner.tagLine(ligne);
          int loc = 0;
        //for (String s : output)
     for(int i=0; i<output.size(); i++)
          {
               String plusSign = " ";
               if (loc == 0)
               {
```

```
                    plusSign = "";
                }
            if(((output.get(i).toString().indexOf("B"))!=-
1)||((output.get(i).toString().indexOf("I"))!=-1))
            {  System.out.println(output.get(i));

             ner1.write(output.get(i).toString()+"\n");}
                loc++;
            }
}
  lecteurAvecBuffer.close();
  ner1.close();
  }
 }
}
```

Annexe

E Liste des publications

5.1 Publications dans des revues internationales

1. Sadek Mansouri, Mbarek Charhad, Mounir Zrigui : A Heuristic Approach to Detect and Localize Text on Arabic News Video. Computación y Sistemas 22(1) (2018). (Indéxée : Scopus, Dblp, Classe : Q3)
2. Sadek Mansouri, Mbarek Charhad, Mounir Zrigui : Arabic Text Detection in News Video Based on Line Segment Detector. Research in Computing Science 132 : 97-106 (2017) (Indéxée : Dblp).
3. Sadek Mansouri, Mbarek Charhad, Mounir Zrigui : A new Approach for Automatic Arabic-Text Detection and Localization in video frames. International Journal of Advanced Intelligence Paradigm (2017). (Indéxée : Scopus, Dblp, Classe : Q4)
   `http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijaip`
4. Sadek Mansouri, Chahira Lhioui, Mbarek Charhad, Mounir Zrigui : A Rule-Based Approach for Arabic Named Entity Metonymy Resolution. Research in Computing Science (2018) (Indéxée : Dblp).
5. Lamia Bouchriha , Sadek Mansouri, Mounir Zrigui : A Deep Learning approach for automatic handwritten Arabic characters recognition. Chapitre de livre : Large Scale Neuroevolution – on Swarm, and Evolutionary Computing Approaches to Deep Learning (2021).(Indexée : Web of Science, EICompendex, DBLP, SCOPUS, Google Scholar, and Springerlink)

5.2 Publications dans des conférences internationales

1. Sadek Mansouri, Chahira Lhioui, Mbarek Charhad, Mounir Zrigui :Text-to-Concept : A Semantic Indexing Framework for Arabic News Videos. CICLing (2) 2017 : 575-584 (Indéxée : Dblp, Scopus, Springer, Classe : B).
2. Sadek Mansouri, Mbarek Charhad, Mounir Zrigui : A Framework for Semantic Video Content Indexing Using Textual Information. IEEE Second International Conference on Data Stream Mining & Processing 2018. (Indéxée : Scopus, IEEE).
3. Sameh Manita, Sadek Mansouri, Mounir Zrigui, Salma Berchech : Arabic text detection in news video using RetinaNet. KES 2021 : 796-803. (Indexée : Dblp, Scopus, Elsevier, Classe : B).
4. Sadek Mansouri, Saleh Zrigui, Mounir Zrigui, Dhaou Berchech : Text detection in Arabic news video based on MSER and RetinaNet, 18th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2021. (Indexée : Dblp, Scopus, IEEE, Classe : C).

5.3 Soumissions dans des revues internationales

1. Sadek Mansouri, Mbarek Charhad, and Mounir Zrigui : SISAVIN : A Semantic Indexing System for Arabic Videos News.International Journal of Multimedia Information Retrieval (IJMIR) 2021 (Q1).