République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

N° d'ordre : 2016/IMI13/89/385.

UNIVERSITE MOHAMED BOUDIAF-M'SILA

FACULTEDETECHNOLOGIE

DEPARTEMENTD'ELECTRONIQUE

MEMOIRE DE MASTER

DOMAINE : SCIENCES ET TECHNOLOGIE

FILIERE : GENIE ELECTRIQUE

OPTION : INSTRUMENTATION ET MAINTENANCE INDUSTRIELLE

Thème

ETUDE ET ANALYSE DE L'EFFET D'ACQUISITION OPTIQUE DES DOCUMENTS TEXTUELS SUR L'ATTRIBUTION D'AUTEURS.

Présenté par :

BENZERROUG Hocine

Proposé et dirigé par :

KHENNOUF Salah

SAYOUD Halim

Promotion : JUIN 2016

ملخص:

إيجاد الكاتب (قياس الأسلوب) هو عبارة عن التطبيق الذي يعتمد على دراسة الأسلوب اللغوي لنص مجهول المؤلف بغرض إيجاد مؤلفه الحقيقي الذي كتبه. من أجل ذلك، عدة خصائص (مميزات) تم استخدامها مثل: Character N-grams، Word N-gram.

في هذا العمل البحثي، نهتم بتحديد هوية كاتب نص مستخرج من صورة ماسح ضوئي، تم تشويشها بدرجات مختلفة من التشويش، باستعمال برنامج (OCR). تكمن أصالة هذا العمل في قدرة النظام المصمّم على تحديد هوية المؤلف مع وجود التشويش، وكذا في طريقة إدخال النص (بواسطة ماسح ضوئي) و استخراجه (بواسطة برنامج (OCR)).

بالإضافة إلى ذلك، تم استخدام عدة مصنفات مثل: SMO، MLP، و كذا بعض المسافات الإحصائية.

الكلمات المفتاحية : إيجاد المؤلف، برنامج (OCR)، تشويش الصور.

Résumé :

L'attribution d'auteur (Stylometrie) est l'application qui consiste à étudier le style linguistique d'un texte anonyme pour connaître l'auteur réel qui l'a écrit. Pour cela, plusieurs caractéristiques ont été utilisé telles que : Character N-grams, Word N-grams.

Dans ce travail de recherche, on s'intéresse à identifier l'auteur d'un texte, extrait à partir d'une image scannée et bruitée à des degrés de bruitage différents, à l'aide d'un OCR. L'originalité de ce travail réside dans la robustesse du système de reconnaissance d'auteur en présence de bruitage, ainsi que la méthode acquisition (scanner) et d'extraction (OCR) du document texte.

Par ailleurs, plusieurs classifieurs ont été utilisé tels que : SMO, MLP et les distances statistiques.

Mots-clés : Attribution d'auteur, logiciel OCR, bruitage d'image.

Abstract

Authorship attribution (Stylometrie) is the application that studies the linguistic style of an anonymous text to recognize the real author who wrote it. For this, several characteristics have been used such as: Character N-grams, Word N-grams.

In this research work, we are interested to identify the author of a text extracted from a scanned image, noised at different degrees of noise, using an OCR. The originality of this work lies in the robustness of the author identification system in the presence of noise.

In addition, several Classifiers have been used such as: SMO, MLP and statistical distances.

Keywords: Author attribution, OCR Software, image noise.

REMERCIEMENTS

Nous remercions avant tout Allah le tout puissant pour son aide, sa bénédiction et pour tout ce qu'il nous a donné.

Comme nous tenons tant à remercier l'ensemble des enseignants ayant contribué à notre formation durant les différents cycles primaire, moyen, secondaire et universitaire.

Un grand merci à nos encadreurs ***Pr. SAYOUD Halim*** *et* ***Mr. KHENNOUF Salah*** *à qui nous devons beaucoup, pour leurs attention, leurs disponibilité, leurs conseils et leurs sympathie que nous avons trouvés en eux, nous sommes très reconnaissants.*

Nous remercions messieurs le chef du département d'électronique et le doyen de la faculté de technologie et tous les enseignants du département d'électronique qui ont contribué à notre formation, ainsi que tous les membres du cadre administratif.

Nous tenons à remercier, enfin, tous ceux qui ont aidés de près ou de loin lors de ce projet de fin d'études.

BENZERROUG HOCINE

DEDICACES

Après un travail ardu, il faut s'arrêter pour dédier :

Je dédie ce modeste travail à la plus chère personne à moi :

A mes très chers parents pour tous leurs amours, qui ont sacrifié les plus belles années de leur vie pour me voir un jour réussir et pour leur soutien moral et matériel et leurs encouragement durant tout ma vie et ou moment particulier du projet.

A mes grandes mères et mes grands pères.

A mes chers frères : Alla Eddine et Mouhamed Badar.

A mes chères sœurs : Amina et son époux Sadouki Akram et Maria.

A toute la famille Benzerroug.

A mes fideles amies que je les considère comme mes frères et sœurs.

Et toutes mes tantes, oncles cousins et cousines.

A toute la promotion d'électronique et toutes les personnes qui je porte dans mon cœur.

BENZERROUG HOCINE

Sommaire

CHAPITRE 2 : Exploration et exploitation des documents textuels

CHAPITRE 3 : Méthodes proposées pour l'attribution d'auteur

CHAPITRE 4 : Résultats et discussions

❖ **Liste des figures :**

❖ **Liste des tableaux :**

LES ABRÉVIATIONS

TAL: Traitement automatique de la langue.

GN: groupe nominal.

GV: Groupe verbal.

GP: Groupe préposionnel.

PNG: Portable Network Graphics.

TIFF: Tagged Image File Format.

JPEG: Joint Photographic Experts Group.

GED: Gestion électronique de documents.

DCT: Découverte des Connaissances à partir des Textes.

ECBD: Extraction de Connaissances dans des Bases de Données.

CAT: Categorization automatique des textes.

SMO: Sequential Minimal Optimization.

MCD: Manhattan Centroid Distance.

MLP: Multi Layer Perceptron.

OCR: Reconnaissance Optique de Caractères.

Introduction générale

INTRODUCTION GENERALE

❶ Notre motivation

Les techniques liées aux traitements de l'information connaissent actuellement un développement très actif en liaison avec l'information et présentent un potentiel de plus en plus important dans le domaine de l'interaction Homme-Machine. L'homme veut communiquer avec l'ordinateur avec la façon la plus simple, la plus naturelle et la plus facile. Pour accélérer l'échange d'informations, l'homme cherche à rendre ces machines accessibles par la voix, capables de lire des textes, de traiter et d'analyser rapidement les informations reçues.

Ecrire pour communiquer a été depuis tous les temps une préoccupation première de l'homme. L'écrit a été, et restera, l'un des grands fondements des civilisations et le mode par excellence de conservation et de la transmission du savoir. En effet, beaucoup d'objets qui nous entourent comportent des traces écrites : les panneaux indicateurs, les notices d'emploi des produits, les journaux, les livres, …etc.

La reconnaissance de l'auteur d'un texte inconnu ou douteux est l'un des plus anciens problèmes de la statistique appliquée à la littérature. Il s'agit toujours de joindre le texte en question à d'autres dont les auteurs sont certains et dont on soupçonne qu'ils ont pu participer à sa rédaction.

❷ Nos Objectifs

Le présent travail s'intègre dans le cadre de l'attribution d'auteurs des documents textes. L'originalité de ce travail réside dans la manière dont on introduit ces documents textes à l'ordinateur. Cette opération consiste à scanner les pages contenant des textes, ensuite convertir les textes scannés ont format (.txt) à l'aide d'un logiciel appelé OCR (Optical Character Recognition).

Dans cette étude, on s'est fixé trois objectifs principaux :

- Faire le point sur la numérisation d'un document texte et son acquisition à l'ordinateur à l'aide d'un scanner et l'extraction d'un texte à partir d'une image numérisée à l'aide d'un OCR,

- Réalisation d'un système de reconnaissance d'auteur basé sur Character N-grams et Word N-grams comme caractéristiques et MLP et SMO comme classifieurs.
- Faire le point sur l'addition d'un bruit (bruitage) aux images numérisées pour voir l'influence de ce dernier sur la robustesse de notre système.
- Conception d'une base de données textuelle (Corpus) qu'on a appelée OCR5P pour valider les techniques proposées.

❸ Structure de la thèse

Notre thèse est structurée en quatre chapitres, de la manière suivante : Dans le premier chapitre on donnera une brève présentation de l'acquisition et l'extraction des documents textes à travers une initiation à la numérisation et les logiciels OCR. Le deuxième chapitre exposera des généralités sur l'exploration et l'exploitation des documents textuels. Les méthodes proposées pour l'attribution automatique des auteurs des textes ainsi que les algorithmes implantés seront abordés dans le troisième chapitre. Les résultats expérimentaux que nous avons obtenus sont présentés dans le dernier chapitre suivi de quelques discussions.

Enfin, nous achèverons ce mémoire par une conclusion générale, donnant les explications possibles aux résultats obtenus et proposant des perspectives pour les futures recherches.

CHAPITRE 1

ACQUISITION ET EXTRACTION DES DOCUMENTS TEXTUELS

CHAPITRE 1
ACQUISITION ET EXTRACTION DES DOCUMENTS TEXTES

Introduction :

Aujourd'hui, nous vivons dans un monde où l'information est disponible en grande quantité tout en étant de qualité très diverse. Internet s'enrichit continuellement de nouveaux contenus. Par exemple, les entreprises emmagasinent de plus en plus de données, le courriel devient un moyen de communication extrêmement populaire, des documents autrefois manuscrits sont aujourd'hui disponibles sous format numérique. Mais toute cette information complexe serait sans intérêt si notre capacité à y accéder efficacement n'augmentait pas elle aussi. Pour cela, nous avons besoin d'outils permettant de chercher, classer, conserver, mettre à jour et analyser les données accessibles. Il est ainsi nécessaire de proposer des systèmes afin d'accéder rapidement à l'information désirée, réduisant ainsi l'implication humaine. Un des domaines qui tente d'apporter des améliorations et de réduire la tâche de l'humain est la classification automatique de documents. Celle-ci consiste à associer une catégorie à un Classification de documents OCR (Reconnaissance Optique de Caractères).

Document pouvant être une phrase, un paragraphe, un texte, etc. Généralement, une classification de documents complexes est effectuée manuellement et sa réalisation est donc coûteuse en termes de temps. En effet, chaque texte (ou une partie) doit être manuellement lu pour attribuer une catégorie adaptée (classe). C'est la raison pour laquelle le domaine de la classification automatique de documents est en perpétuel développement.

Alors, nous exposons dans ce chapitre une brève définition sur les documents textuels et leurs typologies et technologies linguistiques, aussi l'utilisation d'un scanner pour acquérir ces documents puis l'étape d'extraction des documents textuels à partir d'un système de reconnaissance d'OCR.

1.1 Les documents textuels :

1.1.1 Définition de document :

Le terme document peut être défini de plusieurs façons. Le nom document vient du verbe latin docere qui signifie « instruire ». Par voie de conséquence, on peut considérer qu'un document est une « chose » qui peut servir à renseigner, à prouver. On utilise le but pour construire la définition.

On peut aussi définir la notion de document en s'appuyant sur ses composantes. À ce moment-là, un document est un ensemble d'informations porteur de sens pour un auditoire ciblé. D'une manière générale, le document est envisagé comme un ensemble formé par un support et une information qui peut être lue par l'homme ou la machine.

1.1.2 Types de document :

Il y a deux types de document :

- **Les documents textuels :** tous les documents où prédominent le texte, l'écrit (livres, périodiques,...etc.).
- **Les documents non-textuels :** tout document où prédominent l'image, le son ou la combinaison des deux ou des trois, image + son + écrit (documents iconographiques, audiovisuels, multimédia, sonores...).

1.1.3 Typologie des documents :

Etablir une typologie consiste à regrouper des objets selon un même critère. Dans le cas d'un document, ceux-ci sont multiples, et peuvent se décliner selon :

- **La forme des données enregistrées :** texte écrit, image, son, données…
- **La nature des données :** statistiques, références bibliographiques.
- **le support d'enregistrement :** papier, audio-visuel (magnétique, argentique), numérique.
- **La forme éditoriale, liée au mode de production du document :** livre, périodique, rapport, thèse, brevet, carte, annuaire, répertoire …
- **Le contenu documentaire :** document primaire (document originel/source), secondaire (références bibliographiques, métadonnées), tertiaire (synthèses, bilans réalisés à partir de plusieurs documents ou données…).

- **Le champ disciplinaire :** juridique, scientifique, ce qui indique les usages/utilisateurs possibles.
- **La source :** document interne (compte-rendu de réunion, rapport d'activités) administratif (courrier, mél), document externe (littérature éditée, périodiques.)…
- **Le statut :** confidentiel, (accessible en lecture seule ou aussi en écriture) ou du moment de son cycle de vie (en relecture, validé).

1.2 Les producteurs d'information textuelle :

1.2.1 Les auteurs :

Bien qu'il soit un utilisateur averti de l'information textuelle, cette typologie présente l'auteur comme le producteur exclusif du contenu intellectuel du document. Pour déterminer une typologie des auteurs, on admettrait 1'existence d'un premier niveau de distinction qui consiste à traiter chaque auteur par rapport a un secteur disciplinaire bien défini.

Actuellement il est "convenu" que les connaissances humaines sont reparties en trois secteurs disciplinaires:[SAV 98]

- **Le secteur "sciences" :** renferme: la chimie, le génie, l'informatique, etc. Ce secteur est sans doute celui qui est le plus confronte a l'explosion documentaire. De ce fait, il est la cible, de presque tous les travaux en matière de recherche d'information.
- **Le secteur "sciences sociales" :** qui inclut: anthropologie, géographie, gestion, linguistique, psychologie, science politique, sociologie. Il est beaucoup plus difficile a cerner que le milieu scientifique.
- **Le secteur "science humaine" :** il regroupe les disciplines tel que les arts, l'histoire, langues, la littérature, sciences religieuses, etc.

1.2.2 Les éditeurs :

Bien étant un objet intellectuel le document textuel (livres, périodiques,...) est aussi un phénomène économique qui reste dépendant de la politique et de l'économie générale du pays. La production du document textuel devait se conformer aux normes du marche notamment la fameuse relation qui lie 1'offre à la demande. Cette notion de marche aurait besoin naturellement d'être approfondie. Il existe deux types d'éditeurs [MAR 90] :

- Editeur non commercial, il peut être un organisme national ou international, il contribue à la diffusion de certains travaux des institutions de recherche. La règle del a gestion financière se base sur le principe de recouvrement du cout de la publication. II s'agit bien de la définition d'une institution à but non lucratif.
- Editeur commercial est un véritable entrepreneur d'une entreprise à but lucratif. Den ce fait, son premier souci reste le bénéfice. A nos jours ce type d'éditeurs détient la grande partie de la marche de "savoir».

Les deux types d'éditeurs manifestent parfois les mêmes besoins:

- Habileté a attiré les bons auteurs.
- Contrôler l'usage de leurs productions.

Ce dernier besoin touche les notions de la propriété patrimoniale et morale, qui découlent du droit d'auteur, et des lois régissant ce secteur d'activité humaine. L'approfondissement de 1'étude de ce sujet pourrait aboutir à une typologie plus affinée des éditeurs.

1.3 Technologies linguistiques de documents textuels :

La manipulation des documents textuels pour l'extraction de connaissances, pour l'indexation automatique ou pour le résumé est une pratique dont l'importance est reconnue depuis longtemps. Ces systèmes de traitement automatique prennent en entrée des textes ou ensembles de textes qu'ils transforment pour obtenir en sortie une ou plusieurs représentations du sens. La tâche essentielle de l'opération de transformation consiste à traduire des documents potentiellement ambigus en représentations non ambiguës (à l'exception des ambiguïtés structurelles initiales).

La question de la « compréhension » d'un document textuel, qui est au cœur de toute tâche du traitement automatique de la langue (TAL), renvoie donc à deux problèmes majeurs : le premier concerne la représentation du sens du texte et le second la prise en compte du monde de connaissance de référence.

Un système de TAL peut donc commencer l'analyse au niveau du mot pour en déterminer la nature et la structure morphologique, continuer au niveau de la phrase pour déterminer l'ordre des mots, la structure syntaxique et le sens de la phrase entière, avant de s'intéresser enfin au contexte et à l'environnement ou au domaine de référence.

Un mot ou une phrase peut avoir un sens spécifique ou une connotation particulière en fonction d'un contexte ou d'un domaine et peut être en résonance avec d'autres mots ou d'autres phrases dans un contexte donné ou en fonction d'un usage particulier. Pour effectuer une tâche de TAL (traitement automatique de la langue), on distingue classiquement (pour la langue écrite) six niveaux de traitement :

1. Le niveau de la segmentation en mots et en phrases.
2. Le niveau morphologique qui traite de la manière dont sont constituées les unités lexicales (flexion, dérivation, composition, etc.) et vise à déterminer la catégorie de discours de l'unité considérée.
3. Le niveau syntaxique qui détermine la structure des phrases en fonction de la grammaire de référence.
4. Le niveau sémantique qui traite du sens des mots et des phrases.
5. Le niveau du discours qui vise à identifier la structure discursive et argumentative du document.
6. Le niveau pragmatique qui traite du monde de connaissance de référence, c'est-à-dire qui prend en compte les informations extra-linguistiques qui peuvent contribuer à la compréhension du texte.

Cette décomposition en six niveaux est bien sûr toute théorique. Elle ne correspond pas nécessairement au mode de fonctionnement réel de tous les logiciels de TAL. Certains groupent les niveaux 2, 3 et 4 en une seule étape du traitement, alors que d'autres ne prennent pas en compte certaines des étapes mentionnées (par exemple, le niveau pragmatique est rarement pris en compte en tant que tel mais des connaissances de nature pragmatique peuvent être intégrées dans les dictionnaires de référence, en particulier les connaissances métiers). Enfin, les algorithmes utilisés pour les différents niveaux d'analyse ne procèdent pas tous de la même manière (analyse descendante ou montante, avec ou sans retour arrière, etc.).

1.3.1 La segmentation en mots et en phrases :

La première tâche du système consiste à identifier les mots puis les phrases constitutifs du texte. La phrase est en effet, dans la très grande majorité des cas, l'unité linguistique de référence pour l'analyse. Cela n'est pas sans poser de problème dans la mesure où un texte n'est pas une suite d'énoncés isolés les uns des autres mais une suite d'énoncés co-référencés, c'est-à-dire qui s'articulent et « font sens » les uns par rapport aux autres.

De ce point de vue, la résolution des problèmes posés par les relations anaphoriques (par exemple, entre un nom de personne et le pronom qui le désigne dans les phrases suivantes) est loin d'être évidente.

La segmentation en mots (tokenisation en anglais) vise tout d'abord à reconnaître puis regrouper les chaînes de caractères alphabétiques [a…z], [A…Z] ainsi que les différents caractères avec leurs signes diacritiques comme les lettres accentuées, numériques [0…9] et typographiques [?, ;. etc.] pour former des unités lexicales. Le principe consiste donc à identifier préalablement les signes qui vont jouer le rôle de séparateurs entre les unités lexicales.

Ainsi, si on considère que les quatre caractères apostrophent, espace, tiret et point d'interrogation sont des séparateurs, l'énoncé L'entends-tu ? Est constitué de trois mots. Cette liste de séparateurs pose néanmoins un problème avec l'énoncé Que fais-tu aujourd'hui ? Qui serait segmenté en cinq mots avec aujourd'hui considéré comme deux mots. Inversement, l'énoncé Que mange-t-il ? Est constitué de trois mots et non de quatre.

Pour éviter ce genre de problème, il convient de distinguer les contextes dans lesquels un caractère jour le rôle de séparateur. On obtient une liste de séparateurs sans condition (virgule, point-virgule, points d'exclamation et d'interrogation, etc.) et une liste de caractères dont le rôle varie en fonction du contexte (apostrophe, point, tiret, etc.). Une autre solution consiste également à fournir la liste des formes pour lesquelles le caractère ne joue pas le rôle de séparateur (comme dans aujourd'hui).

La segmentation en phrases obéit au même principe mais en considérant comme séparateurs les ponctuations dites « fortes », à savoir le point, les points d'exclamation et d'interrogation et de suspension. Comme pour la segmentation en mots, le rôle du point est ambigu puisqu'il peut être utilisé dans les abréviations : ainsi S.N.C.F. ne correspond pas à quatre phrases mais au sigle correspondant du transporteur ferroviaire national.[CHA 04].

1.3.2 L'analyse morphologique :

L'analyse morphologique consiste à reconnaître la structure des formes de surface telles qu'elles ont été segmentées précédemment puis à leur affecter une catégorie grammaticale.[NOR 07]

La première tâche de l'analyseur morphologique est donc de procéder à la lemmatisation des formes de surface appelées « formes fléchies » en référence aux flexions

qui sont utilisées pour conjuguer les verbes et accorder les adjectifs en genre et en nombre. Une forme fléchie (par exemple chantais) correspond à la concaténation de sa forme de base (chant-) et de la flexion indiquant la première personne du singulier à l'indicatif imparfait (-ais). La morphologie flexionnelle donne l'ensemble des règles permettant d'associer les formes de base avec les flexions, pour les verbes, les noms et les adjectifs.

Ainsi, à partir des formes fléchies du texte, le lemmatiseur va identifier la forme de base et le lemme de référence (par exemple, la forme infinitive des verbes ou l'adjectif au masculin singulier par convention) et la flexion qui lui est associée. L'analyse morphologique du français nécessite de connaître les formes de base constituant les formes fléchies ainsi que les modèles flexionnels. Pour cela, un dictionnaire flexionnel peut être utilisé, qui associe la forme de base, le lemme de référence et le modèle flexionnel.

La seconde tâche de l'analyseur morphologique :Est d'attribuer une catégorie ou étiquette syntaxique à chacune des formes fléchies identifiées (nom, verbe, adjectif, etc.). Le choix des catégories syntaxiques (on parle également de partie du discours, ou de part of speech en anglais) est un problème extrêmement délicat. Même s'il existe un accord de fait concernant l'emploi des catégories principales (comme nom, verbe, adjectif, etc.), il n'existe néanmoins pas de norme ni de standard concernant le nombre, la nature ou l'intitulé de ces catégories. De plus, la finesse des catégories dépend des objectifs poursuivis. Ainsi, dans certains cas, il sera nécessaire de différencier les types de pronoms au sein de la catégorie générale des pronoms personnels alors que, dans d'autres cas, ce ne sera pas utile. Une autre question concerne la nécessité ou non de segmenter en composants élémentaires certaines expressions (par exemple, faire marche arrière ou machine à vapeur). Dans certaines situations (l'indexation d'un texte, par exemple), il peut même être utile de considérer comme expression figée ou semi-figée un multi-terme (par exemple, crise économique ou encéphalopathie spongiforme bovine).

Par ailleurs, le français, comme d'autres langues, possède également une morphologie dérivationnelle. Celle-ci définit les règles permettant d'associer un affixe (suffixe ou préfixe) à une forme de base. Par exemple, le préfixe 're' peut être utilisé avec de nombreux verbes comme refaire ou rejouer ; le préfixe in est quant à lui utilisé pour les adjectifs, comme dans injuste ou insatisfait. De même, un grand nombre de suffixes existent en français, comme isme, ité ou iste.

Les règles de morphologie dérivationnelle sont alors utilisées pour retrouver une forme de base et son lemme à partir d'une forme de surface correspondant, par exemple, à un néologisme rencontré dans un texte, et pour aider ainsi à son analyse (attribution d'une catégorie syntaxique, par exemple).

D'un point de vue logiciel, les analyseurs morphologiques peuvent relever de deux grands types de méthodes. Le premier type concerne les méthodes à base de règles linguistiques qui utilisent les connaissances linguistiques propres à la langue considérée pour déterminer l'attribution des catégories. Un travail important de description linguistique est donc nécessaire en amont pour formaliser les connaissances dans les règles d'attribution. Le second type correspond aux méthodes par apprentissage où les analyseurs sont entraînés sur des corpus traités manuellement. Ainsi, lorsque le logiciel a identifié une catégorie pour une forme donnée, il disposera de la probabilité la plus forte pour trouver la catégorie suivante. Pour ces méthodes, plus les corpus d'entraînement ne sont importants et diversifiés, meilleurs sont les résultats.

À la sortie du module d'analyse morphologique, le texte apparaît sous la forme d'une liste de lemmes avec leur catégorie syntaxique et les informations morphologiques nécessaires.

1.3.3 L'analyse syntaxique :

Le rôle de l'analyse syntaxique est d'abord d'identifier les différents éléments constitutifs de la phrase (appelés syntagmes ou constituants), puis de construire la structure globale de l'énoncé. Pour ce faire, l'analyse est régie par une grammaire de la langue qui est utilisée au niveau local pour la construction des syntagmes et au niveau global pour l'attribution des rôles syntaxiques à chacun des syntagmes (groupe sujet, groupe verbal, groupe complément, etc.).[LAN 91]

L'identification des syntagmes, et particulièrement des syntagmes nominaux, correspond à un enjeu important dans la mesure où de nombreux systèmes utilisent ces derniers comme candidats-descripteurs pour représenter le contenu informationnel d'un texte. Dans cette liste de candidats, le système détermine ensuite ceux qui possèdent les propriétés pour devenir les descripteurs. Pour ce faire, le système peut recourir à un calcul de fréquence ou à une comparaison avec un vocabulaire contrôlé, une liste d'autorités ou un thésaurus.

Cette extraction de syntagmes ou groupes nominaux plus ou moins complexes est utilisée bien sûr par les systèmes d'indexation automatique, mais également par certains types de systèmes de « résumé » automatique. Le « résumé » obtenu est alors moins une véritable « condensation » du texte source qu'une suite d'extraits jugés suffisamment significatifs pour constituer ce que certains appellent une « signature » du texte. Pour constituer ces syntagmes, deux grandes familles de méthodes peuvent être utilisées :

- **D'une part, les méthodes fondées sur l'utilisation de « patrons » (patterns en anglais) où la structure syntaxique :**Est définie à l'avance (par exemple, les groupes nominaux constitués de la suite <Nom Adjectif Adjectif> comme encéphalopathie spongiforme bovine). La méthode des patrons peut également être utilisée pour détecter dans un document ou un flux informationnel comme une dépêche de presse des événements à surveiller. Dans le cadre d'une veille économique, on peut, par exemple, construire des patrons syntaxiques permettant de repérer les opérations de rachat d'entreprises (un exemple très simplifié peut être : Entreprise1 a racheté Entreprise2). La méthode des patrons, qu'ils soient génériques ou spécifiques, est efficace car le traitement effectué prend en compte le contexte immédiat. Inversement, elle risque d'exclure des informations qui pourraient être importantes et qui se trouvent, par exemple, dans un constituant non identifié par le patron, ou dans le verbe de la phrase s'il s'agit d'un patron uniquement destiné à extraire les syntagmes nominaux.
- **Le second type de méthodes repose sur des grammaires à base de règles de réécriture :** Elles permettent à la fois de rendre compte de manière souple des différentes manières de composer un même syntagme et d'exprimer les diverses structures de constituants qui sont acceptables pour une phrase. Le pouvoir d'expression de ces grammaires est beaucoup plus important que la méthode des patrons. En effet, ces grammaires de constituants (dont il existe de très nombreuses versions) permettent de dériver plusieurs constituants à partir d'une seule règle.

Ces règles de réécriture sont constituées de deux parties : une partie gauche qui correspond à l'un des symboles utilisés pour désigner les constituants et une partie droite qui indique la suite de constituants ou de catégories syntaxiques attendus. Par exemple, GN (qui signifie groupe nominal) pourra se réécrire par la suite Déterminant Nom Adjectif ou Déterminant Adjectif Nom ou Nom propre ; GV (groupe verbal) se réécrit Verbe suivi de GN.

Un aspect important de ce formalisme est la possibilité d'utiliser les symboles non terminaux dans la partie droite de la règle, permettant ainsi d'exprimer la récursivité. Cette fonction augmente la puissance d'expression des grammaires en autorisant l'analyse de syntagmes de longueur variable. Ainsi, l'exemple un logiciel de traitement automatique des langues naturelles est reconnu par la grammaire suivante où l'on constate la récursivité par la présence des symboles GN et GP (pour groupe préposionnel) à gauche et des règles à droite :

GN ? Déterminant + Nom + GP

GN ? Nom + Adjectif

GP ? Préposition + GN

L'intérêt de ces grammaires réside à la fois dans leur grande souplesse d'écriture et dans leur pouvoir d'expression. Inversement, elles ont tendance à proposer de nombreuses analyses pour les phrases complexes (suggérant en particulier différentes solutions pour le rattachement des GP en cascade). L'ajout d'une nouvelle règle impose de procéder à des tests de non-régression afin de vérifier que la règle n'a pas d'effet de bord sur l'ensemble de la grammaire.

1.3.4 L'analyse sémantique :

Le quatrième niveau de l'analyse linguistique concerne le traitement sémantique du document et vise à en identifier le sens intrinsèque. Alors que l'analyse syntaxique définit l'acceptabilité grammaticale des phrases, l'analyse sémantique permet de « calculer » leur sens en utilisant soit un système de relations (graphe conceptuel, réseau sémantique), soit un système de traits sémantiques, soit une représentation conceptuelle pivot.[LALL 05].

❖ **La première approche :** Consiste à établir des relations de significations entre les lemmes. C'est donc la place du lemme dans le réseau qui détermine son sens et non pas une description sémantique fine de chacun des lemmes.

Un exemple bien connu est celui du thésaurus, utilisé depuis longtemps dans le monde documentaire, et qui décrit les relations existant entre les termes (relations de synonymie, d'hyponymie, d'hyperonymie, etc.).

Autant cette approche est très efficace pour décrire des mondes conceptuels fermés (domaines de spécialités), autant sa généralisation à la langue générale pose de nombreux et sérieux problèmes.

On peut tout d'abord observer que la polysémie de la plupart des termes, les glissements de sens, les nouvelles acceptions rendent difficilement « maintenable » un réseau de cette taille, sauf à simplifier les relations. Mais, surtout, se pose la question de l'universalité de la représentation du monde qui est sous-jacente au réseau, dans le choix et la nature des relations que dans la place des lemmes les uns par rapport aux autres.

❖ **La deuxième approche :** Consiste à décrire les lemmes au moyen de traits sémantiques (ou sèmes) qui correspondent à des étiquettes. De même que le lemme est décrit, sur le plan syntaxique, par sa catégorie morphologique et le modèle flexionnel qui lui est associé, il est décrit, sur le plan sémantique, par les sèmes qui le caractérisent. Ainsi, le terme avocat sera affecté des traits sémantiques indiquant qu'il peut s'agir d'un fruit ou d'un homme de loi. Si, dans la même phrase, on rencontre le lemme plaider affecté des traits sémantiques indiquant qu'il s'agit d'une prise de parole pour défendre un accusé, seule l'acception homme de loi sera retenue.

La compatibilité des traits sémantiques entre les lemmes d'une même phrase est vérifiée dans un processus d'unification. L'unification vérifie qu'il existe un même trait (ou ensemble de traits) commun aux différents lemmes de la phrase pour conclure à la validité de celle-ci. Par exemple, le syntagme l'avocat marron est accepté car marron comporte le sème malhonnête alors que le syntagme l'avocat bleu ne sera pas accepté.

Si le principe de fonctionnement de l'approche par traits sémantiques est simple, sa mise en œuvre s'avère délicate. D'une part, il est impossible de déterminer a priori tous les sèmes qui seront nécessaires pour les différentes applications. Par exemple, définir le terme caviar uniquement avec les sèmes indiquant qu'il s'agit d'œufs d'esturgeon salés est insuffisant et il conviendrait. Mais, avec cet exemple, on voit bien que les sèmes sont dépendants du type de représentation que l'on donne du monde de référence et du contexte d'usage de l'application qui va manipuler ces connaissances. On retrouve donc d'une certaine manière les objections adressées à l'approche par relations sémantiques pour la question de l'universalité des sèmes. Enfin, se posent également la question de l'adaptation du système de traits à des domaines de spécialités nouveaux ainsi que celle de la maintenance du dictionnaire comportant la description sémantique des termes.

❖ **La troisième approche :** Consiste à adopter une représentation conceptuelle pivot. Elle est le plus souvent utilisée dans des applications multilingues comme les systèmes de traduction automatique ou les systèmes de recherche d'information inter ligne, Cette approche repose sur l'hypothèse que le sens d'une phrase peut être représenté au moyen d'un langage non spécifique entièrement indépendant des langues. Ce formalisme peut être composé de symboles, de codes ou, fréquemment, de termes empruntés à l'une des langues considérées. Ainsi, chaque terme d'une langue est associé à un concept pivot qui permet de générer les termes équivalents dans d'autres langues. Par exemple, le terme neige en français réfère au concept de /neige/ (peu importe le label qui code le concept) et permet de générer le terme snow en anglais, neve en italien, Schnee en allemand, etc.

Si la mise en œuvre de ce type de système pivot est relativement simple, elle pose néanmoins un sérieux problème linguistique car il n'y a pas de rapport bi-univoque entre une langue source et une langue cible (par exemple, le mot neige en français se traduit par de nombreux termes différents en finnois, selon sa qualité, sa température, etc.).

La description sémantique des lemmes s'avère donc une tâche extrêmement difficile et coûteuse. Mais, même si les nombreuses questions théoriques rapidement évoquées ci-dessus n'ont toujours pas trouvé de réponses évidentes, des systèmes linguistiques intégrant le niveau d'analyse sémantique sont désormais opérationnels. D'un point de vue fonctionnel, l'apport de la sémantique permet de désambiguïser les textes qui sont analysés. Du point de vue de l'usager, la décision de recourir à ces approches dépend de plusieurs critères :

- La délimitation conceptuelle du domaine : plus le domaine n'est spécialisé, bien délimité, meilleurs sont les résultats.
- L'évolutivité du domaine : plus le domaine est stable, moins le système de représentation sémantique devra évoluer, moins la maintenance sera fastidieuse.
- Le volume des données à traiter : on ne peut guère envisager un traitement sémantique complexe pour l'indexation du Web, mais traiter un intranet d'entreprise ne pose aucun problème.

1.4 Chaîne de numérisation des documents textuels :

Les principales étapes d'une chaîne de numérisation sont :

1.4.1 L'acquisition des documents textuels :

Permettant la conversion du document papier sous la forme d'une image numérique (bitmap). Cette étape est importante car elle se préoccupe de la préparation des documents à saisir, du choix et du paramétrage du matériel desnaisie (scanner), ainsi que du format de stockage des images.

1.4.2 Extraction des documents textuels :

Cette phase nécessite généralement trois étapes importantes :

- **Le prétraitement** dont le rôle est de préparer l'image du document au traitement. Les opérations de prétraitement sont relatives au redressement de l'image, à la suppression du bruit et de l'information redondante, et enfin à la sélection des zones de traitement utiles.
- **La reconnaissance** du contenu qui conduit le plus souvent à la reconnaissance du texte et à l'extraction de la structure logique. Ces traitements s'accompagnent le plus souvent d'opérations préparatoires de segmentation en blocs et de classification des médias (graphiques, tableaux, images, etc.).
- **La correction des résultats** de la reconnaissance en vue de valider l'opération de numérisation. Cette opération peut se faire soit automatiquement par l'utilisation de dictionnaires et de méthodes de correction linguistiques, ou manuellement au travers d'interfaces dédiées. Seules l'acquisition et la reconnaissance du contenu seront détaillées dans la suite. Le prétraitement est aujourd'hui intégré dans les OCR et est considéré comme suffisant en première approximation.

1.5 Acquisition des documents textuels :

1.5.1 Principe :

La technicité des matériels d'acquisition (scanner) a fait un bond considérable ces dernières années. On trouve aujourd'hui des scanners pour des documents de différents types (feuilles, revues, livres, photos, etc.). Leur champ d'application vadu "scan" de textes au "scan" de photos en 16 millions de couleurs (et même plus pour certains).

1.5.2 Le scanner :

Un scanner, francisé en scanneur, numériseur de document ou numériseur à balayage, est un périphérique informatique qui permet de transformer un document réel(photo, lettre, ...) ou une partie de document en une image numérique.

Un papier **Un scanner** **Image numérique**

Figure 1.1 **:** Acquisition des documents textuels (texte) à travers d'un scanner.

1.5.3 Types de scanners :

On distingue les principaux types de scanner suivants :

- Les scanners à plats : c'est le plus pratique et le plus facile à utiliser. Il suffit de poser son document (même un livre jusqu'à 2 kg en moyenne) sur la vitre comme pour une photocopieuse. Le balayage de la lumière se fait automatiquement à vitesse constante. Le scanner est de grande taille (30 x 50 cm environ).
- Les scanners à main, de taille réduite (taille d'une main d'adulte), doivent être déplacés manuellement sur le document, par bandes successives, afin de le numériser. Dans les années 1990, les scanneurs à main ont été les premiers scanneurs grand public, mais ils sont beaucoup moins utilisés aujourd'hui.
- Les scanners par défilement font défiler le document devant une fente lumineuse fixe pour le numériser, à la manière des télécopieurs (fax). Ce type de numériseur est fréquemment intégré dans les Imprimantes multifonctions.

Il existe d'autres types de scanners : Scanners film, Scanneurs à tambour, Scanneurs automatiques de livre, Scanners à Diapos.

1.5.4 Fonctionnement de scanner :

Dans un scanner à plat, le document à numériser est posé contre une vitre, sous laquelle un miroir et une source de lumière de grande intensité effectuent ensemble un passage. Le document est alors soumis au balayage d'un rayon lumineux.

La lumière réfléchie par le document est renvoyée par le miroir mobile à un système optique qui le transmet à une série de capteurs. Les capteurs convertissent les intensités lumineuses reçues en signaux électriques transmis à l'ordinateur. Chaque ligne du document est décomposée en "points" correspondant à des pixels et les capteurs analysent la couleur de chacun des pixels selon 3 composantes (rouge, vert, bleu).

Les signaux électriques sont à leur tour convertis en données numériques par un convertisseur analogique-numérique afin de recomposer l'image.

1.5.5 Formats des fichiers images :

Il existe différents formats de représentation des fichiers images : TIFF, JPEG, GIF, PNG, etc. dépendant des propriétés de l'image, comme le chromatisme et l'animation, et de la tolérance à la perte de l'information dans le processus de compression. La structuration des données est différente pour chaque format d'image. Un format d'image comprend habituellement un en-tête, contenant des informations générales sur l'ensemble du fichier (par ex. n° de version, ordre des octets, etc.), un ou plusieurs répertoires de paramètres, caractérisant la représentation bitmap de l'image, suivis par les données de l'image qui seront lues et interprétées suivant la valeur des paramètres. Un paramètre particulier est dédié au type de compression autorisé sur le format, avec ou sans perte d'information. Il est important, dans la mesure du possible, d'écarter les formats propriétaires (comme GIF, sous Licence Unisys) et de leur préférer des formats libres de tous droits.

D'autres précautions sont également à prendre en compte concernant les changements de version que peut recouvrir un format, comme c'est le cas du format TIFF dont certaines versions peuvent ne pas être reconnues par certains logiciels. En attendant la généralisation du PNG, qui est l'émanation de recommandations du consortium W3C (1996), le format TIFF est pour l'instant le format le plus répandu pour les documents textuels.

1.6 Extraction des documents textuels :

1.6.1 Principe :

La reconnaissance de caractères est réalisée à l'aide de systèmes dédiés appelés OCR. Son but est de convertir l'image du texte en un texte lisible par ordinateur, en faisant le moins de fautes possibles sur la conversion des caractères. L'existence aujourd'hui de plusieurs outils de ce type a conduit peu à peu à définir des critères de choix pour sélectionner l'OCR le plus efficace et surtout le mieux adapté à son application. Longtemps, le critère d'efficacité était lié à un taux de reconnaissance élevé, pensant qu'une technologie efficace est une technologie sans défaut. En effet, il faut admettre que le taux de 100 reste un objectif à atteindre. Mais réussir une opération de numérisation exploitant la technologie d'OCR nécessite un certain nombre de règles dans la mise en œuvre de ces applications. Il est confirmé que le taux de reconnaissance ne dépend pas du seul moteur de reconnaissance mais d'un ensemble de précautions à prendre lors de la mise en œuvre de l'application.

1.6.2 Qu'est –ce que L'OCR ?

Scanner un document revient à dire qu'on prend une photo de celui-ci. Le résultat d'un scan est donc soit une image (jpeg, tiff, ...) soit un simple PDF (qui est une image en soi également). Le texte qui se trouve sur le document scanné est donc considéré comme de simples pixels noirs disposés sur un fond blanc. Il n'existe donc pas en tant que texte à proprement parlé. Impossible alors de l'utiliser.

L'OCR « Reconnaissance Optique de Caractères » permet d'interpréter des textes inclus sur un document scanné pour le récupérer dans un traitement de textes. Le taux d'erreur est actuellement de moins de 1 % des caractères scannés, même si cette fonction est liée à la qualité du papier ou même à la police utilisée.

L'OCR est donc un processus qui va traiter une photo de texte pour transformer les pixels qui la compose en caractères alphanumériques. Le résultat de cette reconnaissance de caractères peut être un PDF recherchable, un document Word ou encore un tableau Excel, par exemple. Voici un modèle générale pour les systèmes OCR :

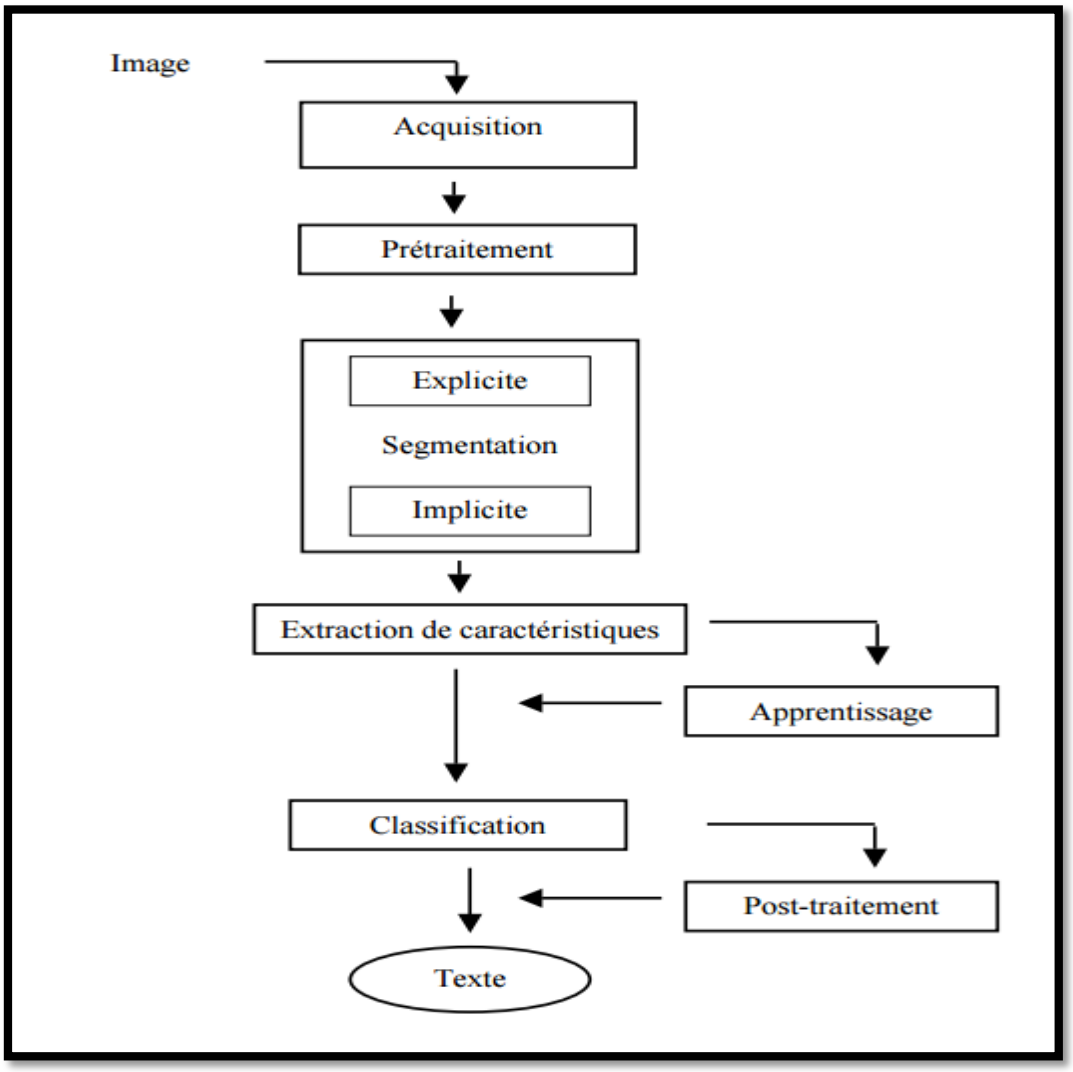


Figure 1.2 : un modèle générale pour les systèmes OCR.

1.6.3 Fonctionnement :

Un texte est une association de caractères appartenant à un alphabet, réunis dans des mots d'un vocabulaire donné. L'OCR doit retrouver ces caractères, les reconnaître d'abord individuellement, puis les valider par reconnaissance lexicale des mots qui les contiennent.

Cette tâche n'est pas triviale car si l'OCR doit apprendre à distinguer la forme de chaque caractère dans un vocabulaire de taille souvent importante, il doit en plus être capable de la distinguer dans chacun des styles typographiques (polices), chaque corps et chaque langue, proposés dans le même document. Cette généralisation omni fonte et multilingue n'est pas toujours facile à cerner par les OCRs et reste génératrice de leurs principales erreurs. Un système de reconnaissance de textes est composé de plusieurs modules : segmentation, apprentissage, reconnaissance et vérification lexicale.[NAGY 92].

A. La segmentation :

Permet d'isoler les éléments textuels, mots et caractères, pour la reconnaissance. Elle se base sur des mesures de plages blanches (interlignes et inter caractères) pour faire la séparation. La multiplicité des polices et la variation des justifications empêchent de stabiliser les seuils de séparation, conduisant à la génération de blancs inexistants ou au contraire à l'ignorance de blancs séparateurs de mots. Ce type d'erreur est très fréquent.

B. La reconnaissance de caractères :

Permet de se prononcer sur l'identité d'un caractère à partir d'un apprentissage de sa forme. Cette étape nécessite une étape préalable de paramétrisation de la forme, définissant des données, des mesures, ou des indices visuels sur lesquels s'appuie la méthode de reconnaissance. Suivant la nature de ces informations, il existe plusieurs catégories de méthodes : syntaxique (description par une grammaire), structurelle (description par un graphe), ou statistique (description par partionnement de l'espace). Ces dernières ont de loin le plus grand intérêt avec les méthodes à base de réseaux de neurones, ou de modèles stochastiques. La complexité de la tâche vient de l'apprentissage qui nécessite, pour sa stabilité, d'un très grand nombre d'échantillons par classe, et de la recherche d'indices visuels discriminants, ce qui n'est pas aisé dans un contexte omni fonte comme celui concerné par la numérisation automatique.

C. .Le post-traitement :

Est effectué quand le processus de reconnaissance aboutit àla génération d'une liste de lettres ou de mots possibles, éventuellement classés par ordre décroissant de vraisemblance. Le but principal est d'améliorer le taux de reconnaissance en faisant des corrections orthographiques ou morphologiques à l'aide de dictionnaires de digrammes, trigrammes ou n-grammes.

1.7 Evaluation des performances des OCR :

L'emploi d'un OCR dans une chaîne de numérisation nécessite une évaluation de ses performances par rapport au jeu de documents à reconnaître. Une approche simpliste d'un système d'OCR serait d'imaginer que sa tâche se limite à localiser l'image de chaque caractère et à le reconnaître indépendamment de toute autre information.

Les logiciels d'OCR sont devenus de plus en plus des systèmes experts à bases de règles intégrant différentes techniques de reconnaissance et entraînés pour maximiser la performance globale. Parce que ces systèmes sont à bases de règles, leur performance n'est pas prévisible de manière théorique. Ce type de système à un comportement qui peut changer de manière imprévisible voyant la performance parfois chuter sur des cas limites. Il suffit parfois d'une petite variation pour que deux caractères, en général bien différenciés, soient confondus. Ces systèmes utilisent des sources d'informations différentes et étagées (taille des caractères, inclinaison des caractères, etc.) pour reconnaître. A cause de la multiplicité de la typographie, ces sources d'information ne peuvent pas être fiables, conduisant à des défaillances dans les cas limites. Ces défaillances sont en plus amplifiées d'étage en étage car chaque étage prend pour acquis les résultats de l'étage précédent.

Conclusion :

Nous avons présenté dans ce chapitre le système d'acquisition et extraction des documents textuels. La numérisation de documents est une étape importante dans la mise en place d'un système d'attribution d'auteurs. Le choix de la solution de numérisation doit prendre en compte toutes les étapes de traitement des documents depuis l'acquisition, la conversion du contenu jusqu'à la correction et la mise en exploitation du document final.

L'extraction des documents textuels est effectuée grâce à les outils informatiques de Reconnaissance Optique de Caractères tel que l'OCR.

CHAPITRE 2

EXPLORATION ET EXPLOITATION DES DOCUMENTS TEXTUELS

CHAPITRE 2
EXPLORATION ET EXPLOITATION DES DOCUMENTS TEXTUELS

Introduction

Ce chapitre présente d'abord la notion d'exploration de données textuelles, ces différentes tâches et domaines d'applications, ensuite son exploitation dans le domaine de classification de textes. Il expose ensuite la notion de représentation numérique des documents textuels. Enfin nous présentons les techniques de classification utilisées pour classer les textes par auteur.

2.1 Exploration de données textuelles :

L'exploration de données textuelles connue aussi sous l'expression de la fouille de texte (Text Mining en anglais) est l'héritière directe de la fouille de données (Data Mining), née dans les années 90. Elle vise à extraire des connaissances dans les textes à l'aide d'un ensemble de traitements informatiques selon des critères de nouveauté ou de similarité.

Plusieurs tâches de cette nouvelle discipline ont étés proposées dans la littérature telles que la classification automatique de textes qui s'impose comme une technologie clé dans la recherche et l'extraction da l'information.

Le développement d'Internet a rendu accessible une énorme quantité de textes, souvent mal rédigés, potentiellement riches d'informations utiles. Ceci a conduit à un intérêt progressif de la catégorisation automatique de textes afin de réduire le travail humain de façon significative, et de résoudre des problèmes d'accès à l'information demandée [HAL 06].

2.1.1 Définitions de l'exploration de données textuelles :

La première définition de l'exploration de données textuelles, présentée en tant que telle, revient à Feldman et al. [FEL98] : « L'exploration de données textuelles est la science qui extrait des motifs cachés à partir de grandes collections de textes ». Une autre définition, celle de Sebastini [F.Y.Y 0, LIO 02] la définie : « L'exploration de données textuelles est de plus en plus employée pour désigner toutes les tâches qui, en analysant de grandes quantités de textes et en détectant des motifs, essaie d'extraire des informations probablement utiles.

Kodratoff [FEL 98], définit la DCT (Découverte des Connaissances à partir des Textes) comme étant : « La science qui découvre les connaissances dans les textes ». Cette définition est assortie des mêmes exigences qu'en ECBD (Extraction de Connaissances dans des Bases

de Données), à savoir que les connaissances découvertes doivent être ancrées dans le monde réel et doivent modifier le comportement d'un agent humain ou mécanique.

2.1.2 Taches de l'exploration de données textuelles :

Quand on parle de l'exploration de données textuelles (ou fouille de textes), on utilise un terme générique, traduction approximative de l'anglais « Text Mining », et l'interprétation la plus immédiate pourrait se référer à la recherche d'information, ou à l'extraction de connaissances. C'est en effet dans ces thématiques que la fouille de texte a pris naissance. Cependant, avec le succès grandissant de cette discipline, et surtout ses possibilités d'applications, d'autres thématiques sont venues compléter ce premier ensemble. [HAC 04]

En fouille de textes comme en fouille de données, tout est quantifiable, les différentes solutions envisagées peuvent être évaluées et comparées. La qualité d'un programme se mesurera à sa capacité à s'approcher le plus possible d'une solution de référence validée par un humain. La fouille de textes peut, parfois, exploiter les statistiques, mais la caractérisation des propriétés d'un texte ou d'un corpus n'est pas sa finalité dernière. Elle a toujours en vue un autre but, formulé dans cette notion de tâche. Certaines tâches élémentaires joueront en outre le rôle d'unités de base en fouille de textes. Dans ce qui suit on donne les principales tâches de l'exploration de données textuelles.

A. Recherche d'Information (RI) :

La tâche de Recherche d'Information (RI), ou Information Retrieval (IR) en anglais, a pour but de retrouver un ou plusieurs document(s) pertinent(s) dans un corpus, à l'aide d'une requête plus ou moins informelle.

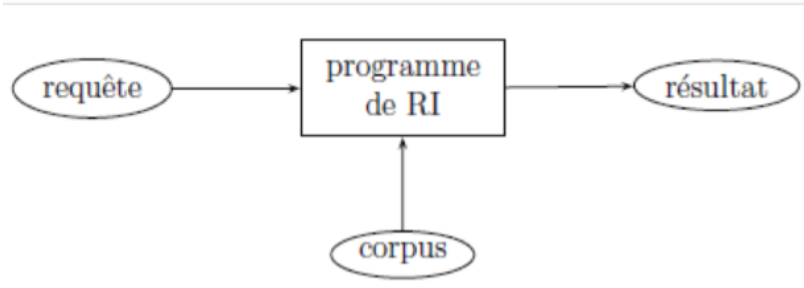


Figure 2.1 **:** Schéma général de la tâche de Recherche d'Information.

La RI est une tâche très utilisée par tous les usagers d'Internet quotidiennement dès qu'ils utilisent un moteur de recherche. Des systèmes de recherche sont aussi intégrés au cœur même de chaque ordinateur, pour aider l'utilisateur à fouiller dans son disque dur afin de trouver un fichier ou un mail.

B. Classification :

La classification consiste à associer une « classe » à chaque donnée d'entrée, comme l'illustre la figure 2.2 suivante :

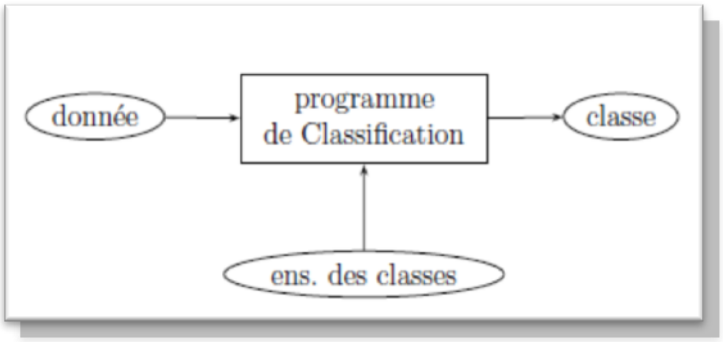


Figure 2.2 : Schéma général de la tâche de Classification.

Les données à classer sont des textes bruts ou des documents semi structurés, et l'ensemble de classes possibles est fini et connu au moment où le programme de classification est sollicité. Dans le cas où seulement deux classes sont possibles, on parle d'une classification binaire.

C. Extraction d'Information (EI) :

Le but de l'Extraction d'Information (ou Information Extraction en anglais), est d'extraire automatiquement des documents textuels des informations servant à remplir les champs d'un formulaire prédéfini come le montre le schéma de la figure 2.4 :

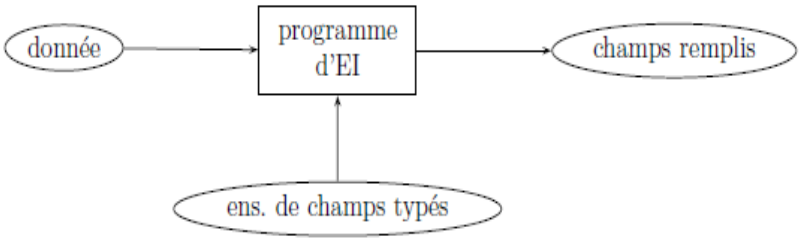


Figure 2.3 : Schéma général de la tâche d'Extraction d'Information.

L'extraction d'information est née d'un challenge organisé lors des conférences MUC ("Message Understanding Conference") qui se sont déroulées entre 1987 et 1998 aux Etats Unis, sous l'impulsion de la Darpa (l'agence de recherche du département de la Défense américain). Les participants se voyaient confier des corpus et leurs programmes étaient comparés en fonction de leur capacité à remplir à partir de chaque texte les champs d'un formulaire prédéfini. Par exemple, en 1992, il s'agissait d'extraire de dépêches d'agences de

presse décrivant des attentats des informations telles que : date, lieu, auteur présumé ou revendiqué, nombre de victimes, etc. On mesure aisément l'intérêt stratégique de ce genre d'applications...

Une de ses applications phare actuelle est la reconnaissance des *entités nommées*, ces mots ou groupes de mots qui identifient soit des noms propres (désignant des personnes, des lieux ou des organisations) soit des quantités mesurables (exprimant notamment des dates, des valeurs numériques ou monétaires). Les fameux "cinq W" du journalisme anglo-saxon ("who did what, where and when, and why", c'est à dire "qui a fait quoi, où, quand et pourquoi" en français) attendent, pour la plupart, une réponse en forme d'entité nommée, la reconnaissance des noms propres et des dates présentes dans les textes ou les pages HTML ou XML. L'analyse automatique de CV, ou de sites marchands pour faire de la comparaison de prix, sont encore d'autres applications potentiellement très utiles de l'extraction d'information.1

D. La segmentation de textes :

La segmentation de texte **[**HAC 04**]** est une tâche de reconnaissance thématique.

Ce que pourraient avoir en commun toutes les recherches sur la segmentation de texte ce sont les caractéristiques suivantes :

- La détection de la cohésion (thématique, lexicale) dans un texte

- La définition de la limite de segment lorsqu'il y a rupture de cohésion : soit par changement lexical **[**S.J 00**]**.

- La capacité à présumer de l'unité du segment par rapport à une unité connexe et cohérente : ainsi, des segments seront constitués de plusieurs phrases adjacentes si toutefois ces dernières maintiennent la cohésion choisie. Des phrases séparées par plusieurs autres phrases ne pourront pas relever d'un même segment, sauf à considérer les phrases intermédiaires comme une forme remplissage (filler) qui ne rompt pas la chaîne (thématique, lexicale) ainsi créée. Toujours est-il que les tâches de segmentation dépendent fortement de ce pourquoi elles sont réalisées. Elles peuvent être par exemple associées :

- A une tâche de recherche d'information, dans laquelle on cherchera à fournir en réponse non seulement un texte (issu d'une URL par exemple) mais plutôt, dans ce texte, le ou les fragments les plus véritablement compatibles avec la question posée **[**KON 04**].**

- A une tâche d'indexation d'un texte pour des buts de création de méta-données à usage pédagogique ou documentaire **[**R.JAL**]**.

- A une tâche de résumé automatique ou semi-automatique, dirigé par le thème, où le résumé se fait par extraction des segments les plus appropriés à un thème donné et création d'un nouveau document

- A des travaux d'extraction du plan ou de la structure du document pour diverses fonctions ultérieures.

E. La reconnaissance d'auteurs :

La détermination du véritable auteur d'un écrit (œuvre littéraire, article de presse, lettre, courriel) a donné lieu à de nombreuses études au cours de ces deux dernières décennies **[**SEB 05**].** La question de l'attribution d'auteur connaît de nouveaux prolongements comme la vérification d'auteur. Dans ce cas, on souhaite savoir si un texte a été écrit ou non par un auteur donné. De plus, au lieu d'obtenir le nom probable de l'auteur, on peut se limiter à déterminer des informations socio-économiques le concernant (profilage) comme le sexe, l'âge, la nationalité, le niveau d'éducation, etc. **(9-Argamon et al. 2009**). Enfin, l'attribution d'auteur fait également partie des sciences forensiques, de débats légaux, mais surtout d'un intérêt grandissant sur le Web avec, comme variantes, l'analyse de la crédibilité des auteurs (blogs, Twitter), voire la détection de plagiat.

2.1.3 Les étapes de la fouille de textes :

La figure 2.4, montre les principales étapes de la fouille de textes, depuis la collecte des textes jusqu'à la découverte de nouvelles informations.

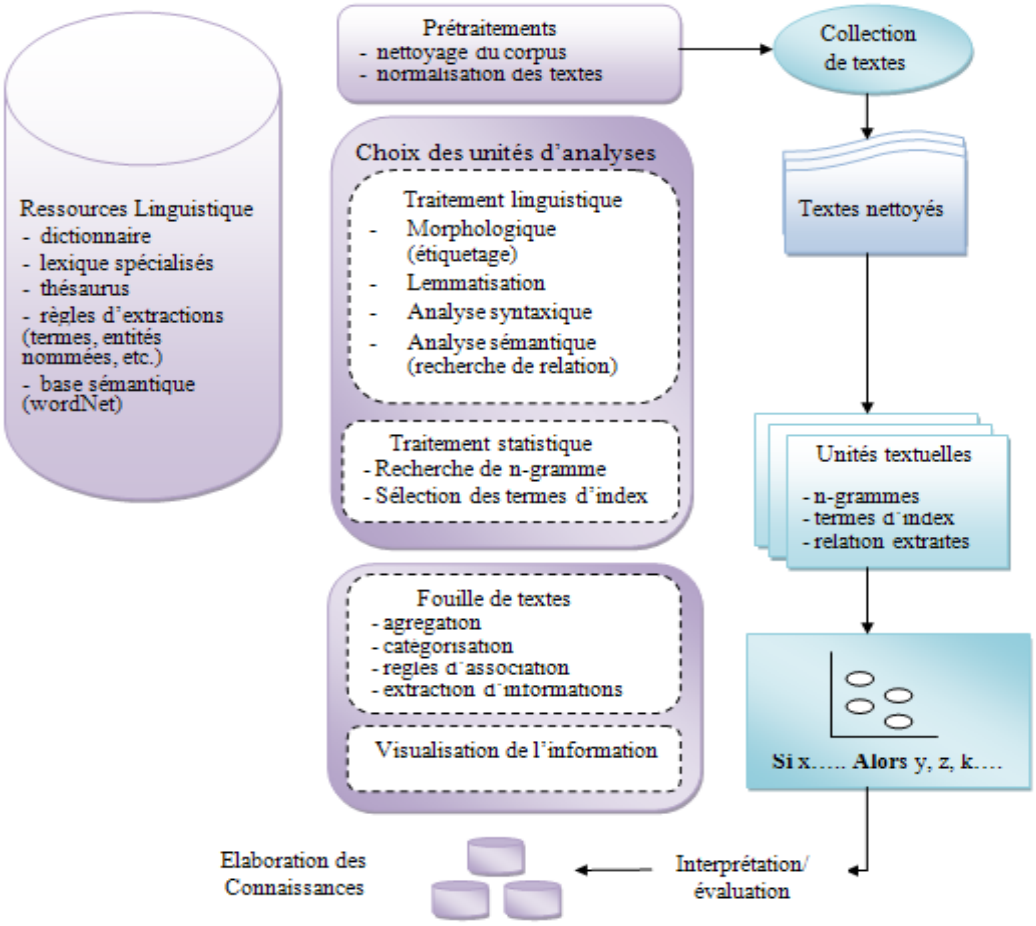


Figure 2.4 : Vue schématique des étapes de la FT [10].

Dans ce schéma proposé dans **[10]**, les traitements effectués à chaque stade varient en fonction de l'application et du type de connaissances utilisé. La phase de prétraitements peut être plus ou moins élaborée. Elle peut inclure l'élimination de mots vides (mots grammaticaux) ou une normalisation plus poussée des textes dans le cas d'un corpus très technique (corpus médical, par exemple). La phase du choix des unités d'analyse peut faire appel aux connaissances linguistiques (extraction des termes, des relations sémantiques entre eux) ou simplement statistique, avec la recherche des n-grammes dans les textes (séquences de mots adjacents qui se répètent dans le corpus). Les deux techniques peuvent enfin être combinées lorsqu'il s'agit de choisir, parmi les unités extraites, celles qui ont un poids discriminant (indexation automatique). Certaines applications de fouille de textes peuvent nécessiter des traitements de nature sémantique avec la recherche de relations entre les unités extraites. La phase de fouille à proprement parlé va puiser dans un panel de techniques pour choisir celle(s) adaptée(s) à l'application et aux types de résultats attendus.

2.1.4 Text Mining :

Le Text Mining est une technique permettant le traitement de gros volumes de textes pour en extraire les principales tendances et répertorier de manière statistique les différents sujets évoqués ainsi que découvrir des connaissances et des relations à partir des documents disponibles.

L'outil de Text Mining va générer de l'information sur le contenu du document. Cette information n'était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document. Les applications en Text Mining peuvent être :

- Recherche d'information
- Correction orthographique/grammaticale
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- La veille technologique
- **Attribution d'auteurs.**

2.2 L'attribution d'auteurs :

Toute solution en attribution d'auteur repose sur une représentation des documents et un modèle ou règle de catégorisation (Juola, 2006). Afin de représenter un texte, les études quantitatives précédentes ont cherché à définir une mesure stylométrique unique devant être constante pour un auteur donné et différente d'un écrivain à l'autre (Holmes, 1998).

2.2.1 Modèles d'attribution d'auteur :

A. La règle Delta :

Afin de déterminer l'auteur probable d'un écrit, Burrows (2002) propose de tenir compte des vocables les plus fréquents et, en particulier, des mots fonctionnels, tout en ignorant les signes de ponctuation. Afin de représenter chaque document ou profil d'auteur, Burrows (2002) tient compte de 40 à 150 termes les plus fréquents, cette dernière valeur donnant souvent les meilleures performances. Pour être précis, Burrows distingue entre les homographes comme, par exemple, *to* comme préposition (e.g., *to you*) ou partie de l'infinitif (e.g., *to be*). Une telle distinction complexifie le traitement des documents sans générer une amélioration significative des résultats (Hoover, 2004).

B. La distance du chi-carré :

Comme deuxième modèle d'attribution d'auteur, nous avons repris l'une des meilleures solutions empiriques testées par Grieve (2007). Dans ce cas, la meilleure représentation des documents se basait sur la fréquence relative des vocables comprenant également huit signes de ponctuation (. , : ; - ? ('). Afin de réduire l'univers lexical, Grieve (2007) impose comme critère de sélection une *k*-limite dans laquelle l'entier *k* indique le nombre minimum d'articles par auteur dans lesquels le terme doit apparaître. Ainsi, le critère 5-limite implique que les termes retenus doivent apparaître dans au moins cinq articles écrits par chacun des auteurs.

C. La divergence Kullbach-Leibler (KLD) :

Zhao & Zobel (2007) proposent de définir a priori une liste restreinte de vocables permettant de discriminer le style des divers auteurs. Pour la langue anglaise, leur liste comprend 363 termes comprenant principalement des mots fonctionnels.

2.3 Démarche à Suivre Pour la Catégorisation de Textes :

Pour réaliser l'opération de catégorisation automatique de textes, la démarche commune est la suivante : la première phase consiste à formaliser les textes afin qu'ils soient compréhensibles par la machine et utilisables par les algorithmes d'apprentissage. La catégorisation des documents est la deuxième phase, cette étape est bien entendu décisive car c'est elle qui va permettre ou non aux techniques d'apprentissage de produire une bonne généralisation à partir des couples (Document, Classe).

La démarche d'une approche standard de catégorisation automatique de textes peut être résumée de la manière suivante :

- Eliminer les caractères de séparation, les signes de ponctuations, les mots vides, etc.
- Les termes restants sont tous des attributs
- Un document devient un vecteur <terme, fréquence>
- Entrainer le modèle de catégorisation à partir des couples (Document, Classe).
- Évaluer les résultats du classifieur.

La figure 2.5 illustre la démarche de catégorisation de textes.

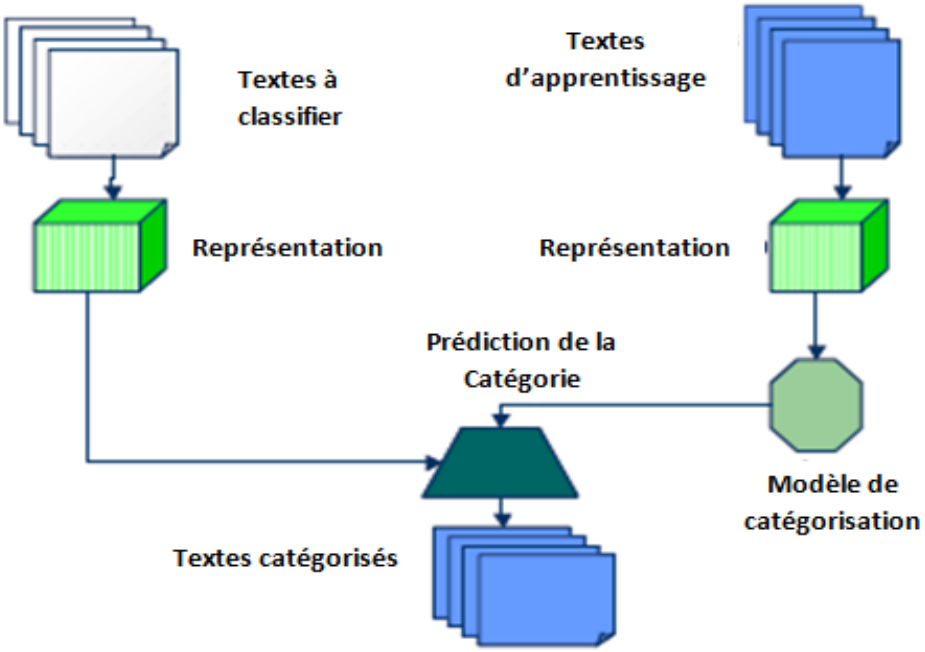


Figure 2.5 : Démarche de la catégorisation de textes.

Conclusion :

Nous avons passé en revu, dans ce chapitre l'exploration et l'exploitation de données textuelles (ou fouille de texte). Cette dernière est actuellement définie par ces objectifs qui sont divers et visent à extraire des informations utiles et exploitables des documents textes. Ensuite, nous avons abordé l'une des principales tâches de cette nouvelle discipline qui est la classification des documents textuels et l'attribution d'auteur. Les méthodes et les techniques utilisées pour effectuer cette tâche feront l'objet du prochain chapitre.

CHAPITRE 3

METHODES PROPOSEES POUR L'ATTRIBUTION d'AUTEUR

CHAPITRE 3
METHODES PROPOSEES POUR L'ATTRIBUTION D'AUTEUR

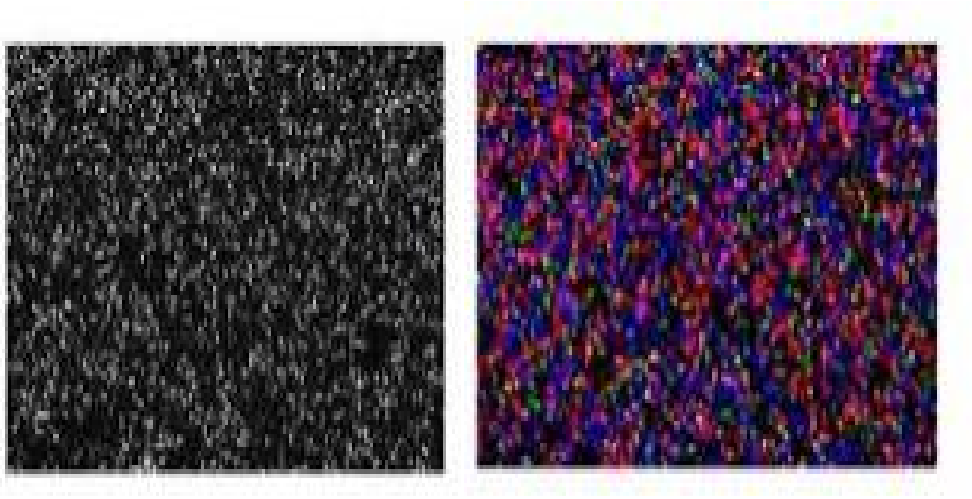
Introduction

Dans le présent chapitre, nous allons parler sur les méthodes proposées pour l'attribution d'auteur des textes déjà scanner et bruités sous matlab. Les textes d'images bruités sont convertis et les enregistrés en format (.txt) avec logiciel OCR.Les tests d'évaluation des performances de notre système de reconnaissance d'auteur à partir des documents textuels ont été effectués par plusieurs méthodes d'analyse avec différents classifieurs tel que : MLP, SMO et Manhattan Centroid Distance (MCD), etc.…

3.1 Bruitage des documents :

Dans toute image numérique, les valeurs de gris ou de couleur observées présentent une incertitude. Cette incertitude est due aux aléas du comptage des photons arrivant sur chaque capteur. Les valeurs de couleur mesurées sont perturbés car les capteurs reçoivent des photons parasites et subissent des fluctuations électrostatiques lors de leurs charges et d´échanges. Quand un capteur reçoit beaucoup de photons venant d'une scène bien éclairée, les parasites sont négligeables par rapport au flux de vrais photons. Mais, même dans une photo d'exposition suffisante, les pixels sombres reçoivent très peu de photons et sont donc !bruités ". Visuellement, on distingue en général deux types de bruit d'image qui s'accumulent :

- Le bruit de chrominance, qui est la composante colorée des pixels bruités : il est visible sous la forme de taches de couleurs aléatoires.
- Le bruit de luminance, qui est la composante lumineuse des pixels bruités : il est visible sous la forme de taches plus foncées ou plus claires donnant un aspect granuleux à l'image.



Bruit de luminance **Bruit chrominance**

Figure 3.1 : Les types de bruit d'image.

Le bruit d'image est la présence d'informations parasites qui s'ajoutent de façon aléatoire aux détails de la scène photographiée numériquement. Il est plus particulièrement visible dans les zones peu éclairées, où le rapport signal/bruit est faible, mais aussi dans les parties uniformes telles qu'un ciel bleu. Il a pour conséquence la perte de netteté dans les détails.

3.1.1 Bruit Salt &Pepper noise (Poivre et Sel) :

Le bruit Poivre et Sel " ou bruit impulsionnel est un bruit qui assigne à un certain nombre de pixels de l'image une valeur 0 ou 255 (noirs et blancs) aléatoirement. Ce bruit est dû soit à des erreurs de transmission de données, soit à la défaillance d'éléments de capteur CCD, soit à la présence de particules fines sur le capteur d'images.

La fonction de Salt & Pepper sous Matlab est comme suit :

```
I = imread ('Corliss lamont1.jpg');
J = imnoise (I,'salt& pepper',0.06);
figure, imshow(I)
figure, imshow(J)
```

3.1.2 Exemple d'une image bruitée de type bruit de luminance utilisé :

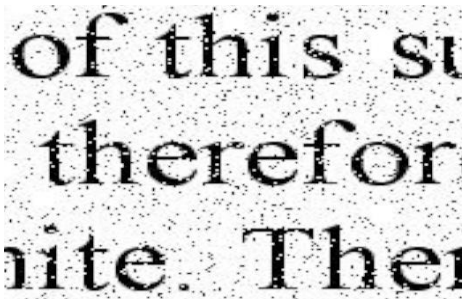
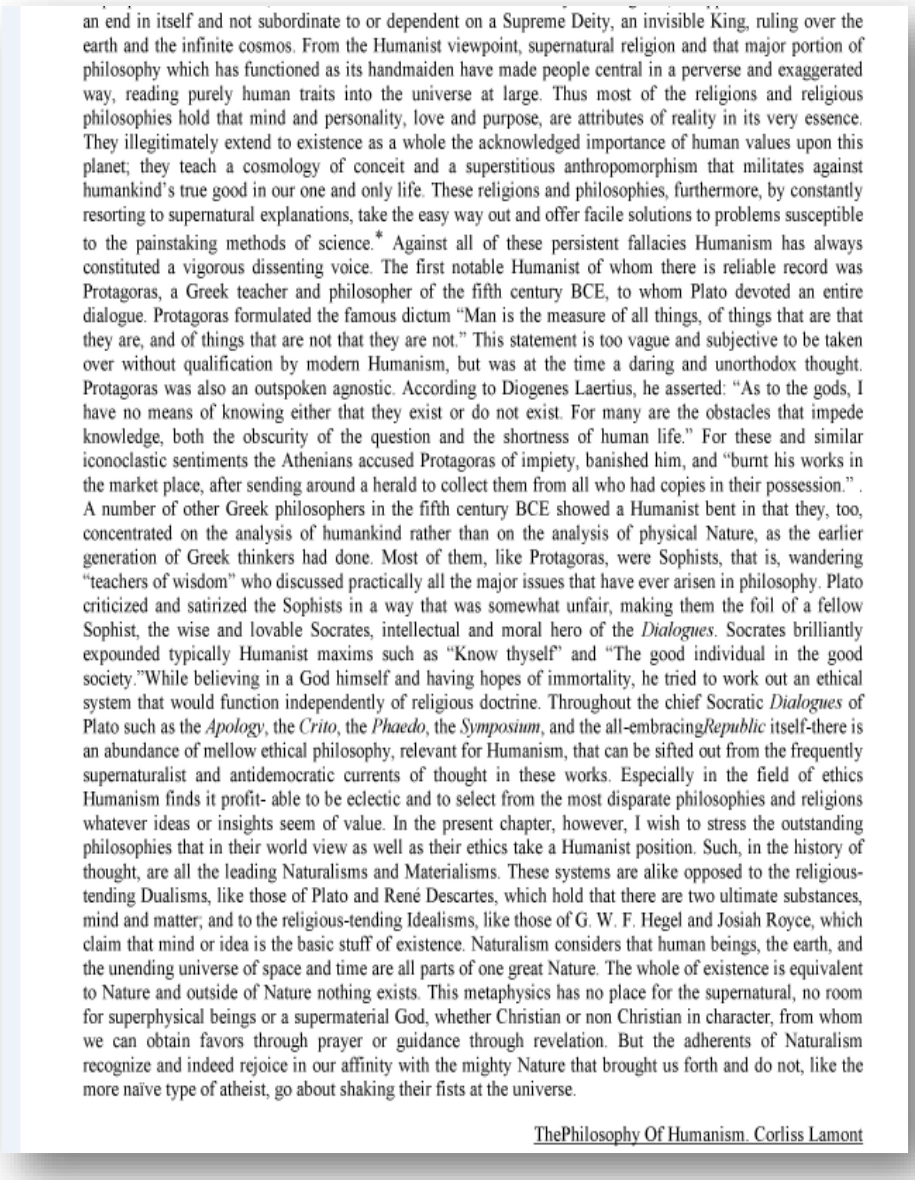


Figure 3.2 : Le type de bruit.

3.1.3 Exemple de bruit Salt & Pepper d'un texte d'image :

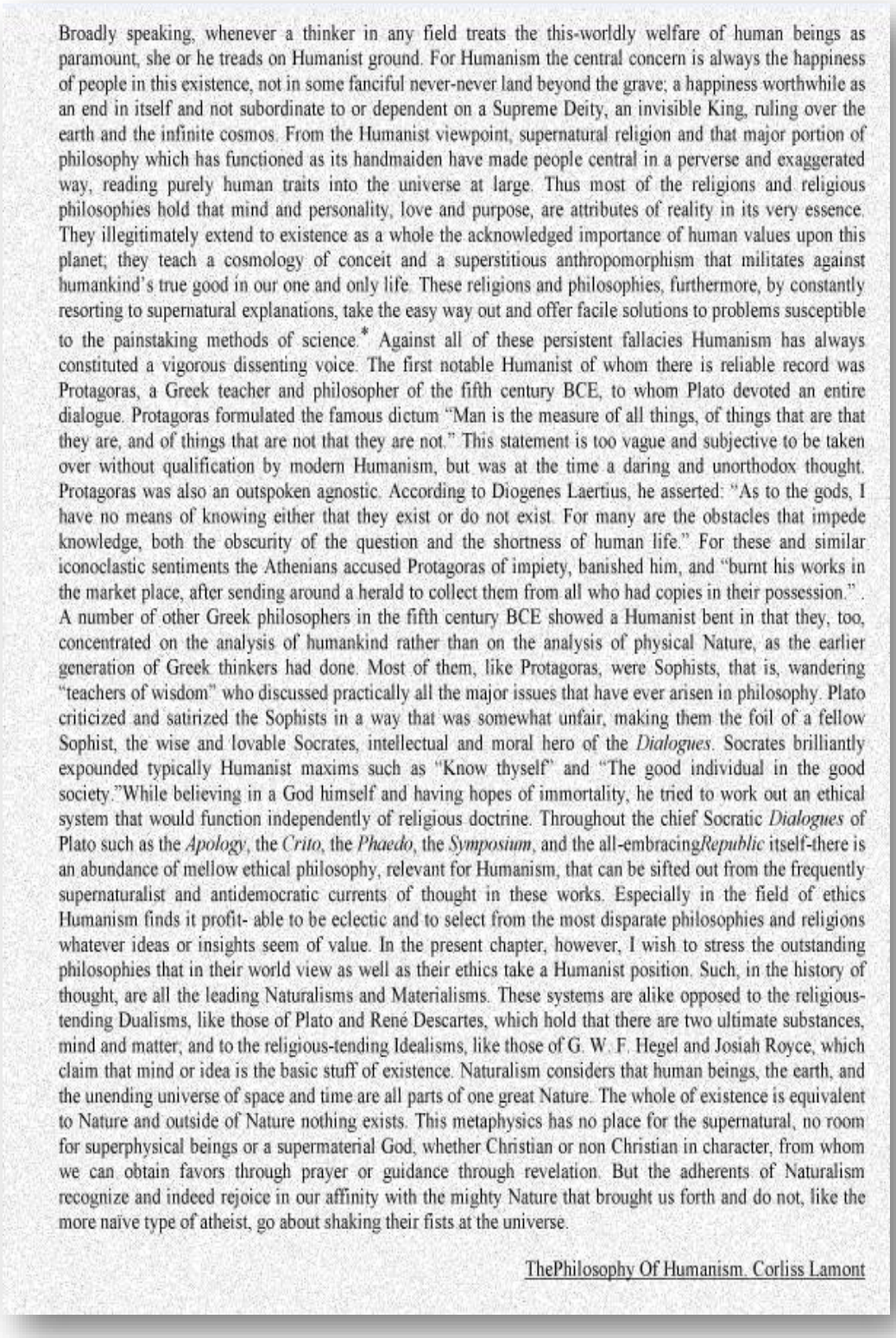


an end in itself and not subordinate to or dependent on a Supreme Deity, an invisible King, ruling over the earth and the infinite cosmos. From the Humanist viewpoint, supernatural religion and that major portion of philosophy which has functioned as its handmaiden have made people central in a perverse and exaggerated way, reading purely human traits into the universe at large. Thus most of the religions and religious philosophies hold that mind and personality, love and purpose, are attributes of reality in its very essence. They illegitimately extend to existence as a whole the acknowledged importance of human values upon this planet; they teach a cosmology of conceit and a superstitious anthropomorphism that militates against humankind's true good in our one and only life. These religions and philosophies, furthermore, by constantly resorting to supernatural explanations, take the easy way out and offer facile solutions to problems susceptible to the painstaking methods of science.* Against all of these persistent fallacies Humanism has always constituted a vigorous dissenting voice. The first notable Humanist of whom there is reliable record was Protagoras, a Greek teacher and philosopher of the fifth century BCE, to whom Plato devoted an entire dialogue. Protagoras formulated the famous dictum "Man is the measure of all things, of things that are that they are, and of things that are not that they are not." This statement is too vague and subjective to be taken over without qualification by modern Humanism, but was at the time a daring and unorthodox thought. Protagoras was also an outspoken agnostic. According to Diogenes Laertius, he asserted: "As to the gods, I have no means of knowing either that they exist or do not exist. For many are the obstacles that impede knowledge, both the obscurity of the question and the shortness of human life." For these and similar iconoclastic sentiments the Athenians accused Protagoras of impiety, banished him, and "burnt his works in the market place, after sending around a herald to collect them from all who had copies in their possession." .
A number of other Greek philosophers in the fifth century BCE showed a Humanist bent in that they, too, concentrated on the analysis of humankind rather than on the analysis of physical Nature, as the earlier generation of Greek thinkers had done. Most of them, like Protagoras, were Sophists, that is, wandering "teachers of wisdom" who discussed practically all the major issues that have ever arisen in philosophy. Plato criticized and satirized the Sophists in a way that was somewhat unfair, making them the foil of a fellow Sophist, the wise and lovable Socrates, intellectual and moral hero of the *Dialogues*. Socrates brilliantly expounded typically Humanist maxims such as "Know thyself" and "The good individual in the good society."While believing in a God himself and having hopes of immortality, he tried to work out an ethical system that would function independently of religious doctrine. Throughout the chief Socratic *Dialogues* of Plato such as the *Apology*, the *Crito*, the *Phaedo*, the *Symposium*, and the all-embracing*Republic* itself-there is an abundance of mellow ethical philosophy, relevant for Humanism, that can be sifted out from the frequently supernaturalist and antidemocratic currents of thought in these works. Especially in the field of ethics Humanism finds it profit- able to be eclectic and to select from the most disparate philosophies and religions whatever ideas or insights seem of value. In the present chapter, however, I wish to stress the outstanding philosophies that in their world view as well as their ethics take a Humanist position. Such, in the history of thought, are all the leading Naturalisms and Materialisms. These systems are alike opposed to the religious-tending Dualisms, like those of Plato and René Descartes, which hold that there are two ultimate substances, mind and matter; and to the religious-tending Idealisms, like those of G. W. F. Hegel and Josiah Royce, which claim that mind or idea is the basic stuff of existence. Naturalism considers that human beings, the earth, and the unending universe of space and time are all parts of one great Nature. The whole of existence is equivalent to Nature and outside of Nature nothing exists. This metaphysics has no place for the supernatural, no room for superphysical beings or a supermaterial God, whether Christian or non Christian in character, from whom we can obtain favors through prayer or guidance through revelation. But the adherents of Naturalism recognize and indeed rejoice in our affinity with the mighty Nature that brought us forth and do not, like the more naïve type of atheist, go about shaking their fists at the universe.

ThePhilosophy Of Humanism. Corliss Lamont

Figure 3.3 : L'image numérique 'Corliss lamont1.jpg' (Image d'un texte scanné).

L'image numérique après le bruitage de degré 6%qui fait sous Matlab :



Broadly speaking, whenever a thinker in any field treats the this-worldly welfare of human beings as paramount, she or he treads on Humanist ground. For Humanism the central concern is always the happiness of people in this existence, not in some fanciful never-never land beyond the grave; a happiness worthwhile as an end in itself and not subordinate to or dependent on a Supreme Deity, an invisible King, ruling over the earth and the infinite cosmos. From the Humanist viewpoint, supernatural religion and that major portion of philosophy which has functioned as its handmaiden have made people central in a perverse and exaggerated way, reading purely human traits into the universe at large. Thus most of the religions and religious philosophies hold that mind and personality, love and purpose, are attributes of reality in its very essence. They illegitimately extend to existence as a whole the acknowledged importance of human values upon this planet; they teach a cosmology of conceit and a superstitious anthropomorphism that militates against humankind's true good in our one and only life. These religions and philosophies, furthermore, by constantly resorting to supernatural explanations, take the easy way out and offer facile solutions to problems susceptible to the painstaking methods of science.* Against all of these persistent fallacies Humanism has always constituted a vigorous dissenting voice. The first notable Humanist of whom there is reliable record was Protagoras, a Greek teacher and philosopher of the fifth century BCE, to whom Plato devoted an entire dialogue. Protagoras formulated the famous dictum "Man is the measure of all things, of things that are that they are, and of things that are not that they are not." This statement is too vague and subjective to be taken over without qualification by modern Humanism, but was at the time a daring and unorthodox thought. Protagoras was also an outspoken agnostic. According to Diogenes Laertius, he asserted: "As to the gods, I have no means of knowing either that they exist or do not exist. For many are the obstacles that impede knowledge, both the obscurity of the question and the shortness of human life." For these and similar iconoclastic sentiments the Athenians accused Protagoras of impiety, banished him, and "burnt his works in the market place, after sending around a herald to collect them from all who had copies in their possession." . A number of other Greek philosophers in the fifth century BCE showed a Humanist bent in that they, too, concentrated on the analysis of humankind rather than on the analysis of physical Nature, as the earlier generation of Greek thinkers had done. Most of them, like Protagoras, were Sophists, that is, wandering "teachers of wisdom" who discussed practically all the major issues that have ever arisen in philosophy. Plato criticized and satirized the Sophists in a way that was somewhat unfair, making them the foil of a fellow Sophist, the wise and lovable Socrates, intellectual and moral hero of the *Dialogues*. Socrates brilliantly expounded typically Humanist maxims such as "Know thyself" and "The good individual in the good society."While believing in a God himself and having hopes of immortality, he tried to work out an ethical system that would function independently of religious doctrine. Throughout the chief Socratic *Dialogues* of Plato such as the *Apology*, the *Crito*, the *Phaedo*, the *Symposium*, and the all-embracing*Republic* itself-there is an abundance of mellow ethical philosophy, relevant for Humanism, that can be sifted out from the frequently supernaturalist and antidemocratic currents of thought in these works. Especially in the field of ethics Humanism finds it profit- able to be eclectic and to select from the most disparate philosophies and religions whatever ideas or insights seem of value. In the present chapter, however, I wish to stress the outstanding philosophies that in their world view as well as their ethics take a Humanist position. Such, in the history of thought, are all the leading Naturalisms and Materialisms. These systems are alike opposed to the religious-tending Dualisms, like those of Plato and René Descartes, which hold that there are two ultimate substances, mind and matter, and to the religious-tending Idealisms, like those of G. W. F. Hegel and Josiah Royce, which claim that mind or idea is the basic stuff of existence. Naturalism considers that human beings, the earth, and the unending universe of space and time are all parts of one great Nature. The whole of existence is equivalent to Nature and outside of Nature nothing exists. This metaphysics has no place for the supernatural, no room for superphysical beings or a supermaterial God, whether Christian or non Christian in character, from whom we can obtain favors through prayer or guidance through revelation. But the adherents of Naturalism recognize and indeed rejoice in our affinity with the mighty Nature that brought us forth and do not, like the more naïve type of atheist, go about shaking their fists at the universe.

ThePhilosophy Of Humanism. Corliss Lamont

Figure 3.4 : l'image bruitée avec un bruit de poivre et sel de degré 6%.

3.2 Extraction de texte d'une image numérisée (OCR vers word «Txt ») :

Le logiciel OCR sera l'un des meilleurs choix. L'OCR, sigles d'Optical Character Recognition (en français on dit Reconnaissance Optique de Caractères), est une fonction pour extraire le texte d'une image ou des documents imprimés qu'on a scanné. Après la conversion du fichier avec OCR, vous aurez un fichier Texte que vous pourrez utiliser, modifier et retravailler à loisir. Essayez Renée PDF Aide pour convertir un document PDF scanné ou une image en texte éditable.

L'OCR capable de lire :

- Les fichiers images.
- Les fichiers PDF.
- Et gérant les scanner compatibles twain.

Et de convertir leur contenu en :

- texte qu'on peut copier.
- en fichier texte.
- ou en document Word.

3.2.1 Exemples des images bruitées convertis en Txt à l'aide de l'OCR :

A. Exemple 01 :

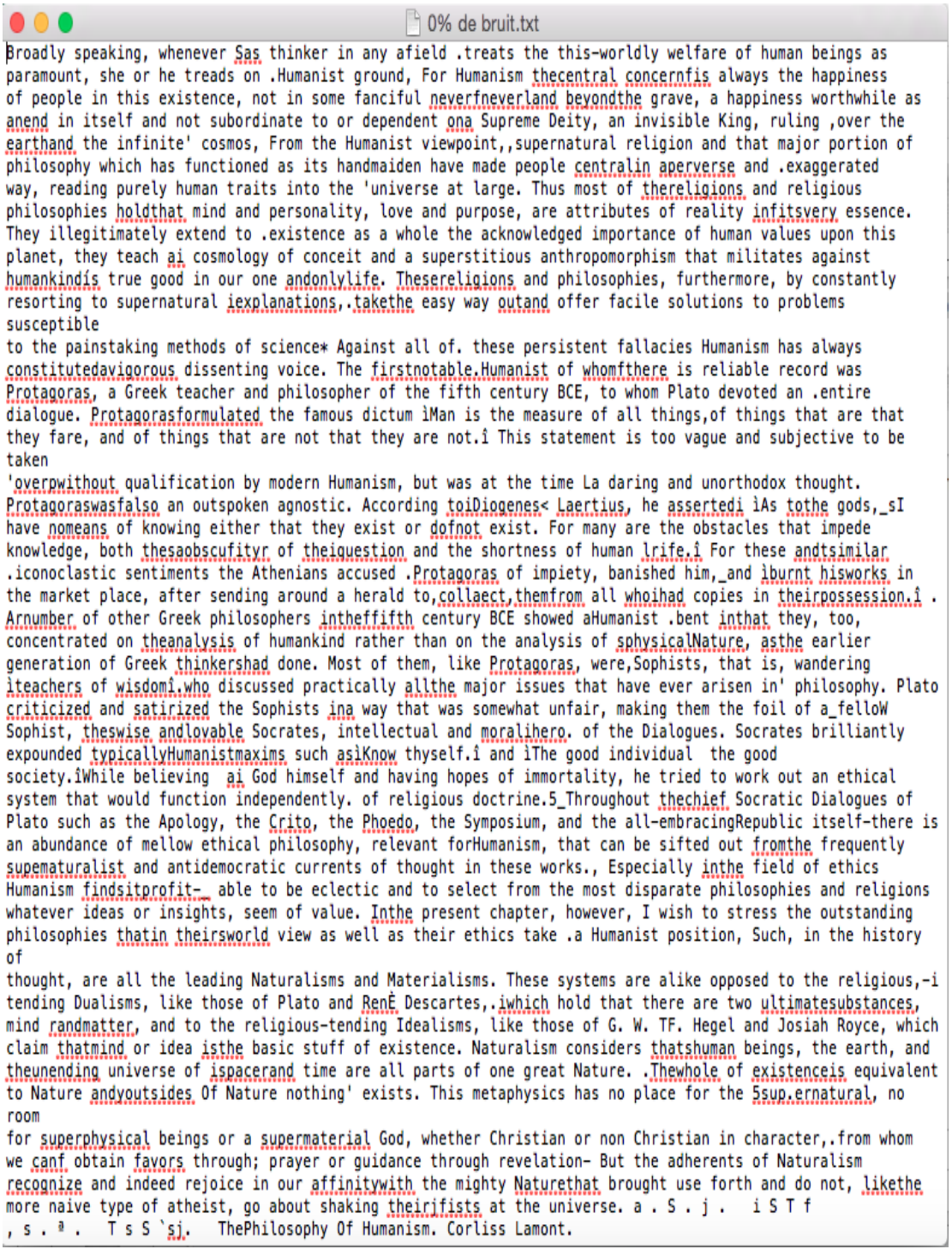


Figure 3.5 : L'image d'un exemple de texte saint (0% de bruit) converti à l'aide de l'OCR.

B. Exemple 02 :

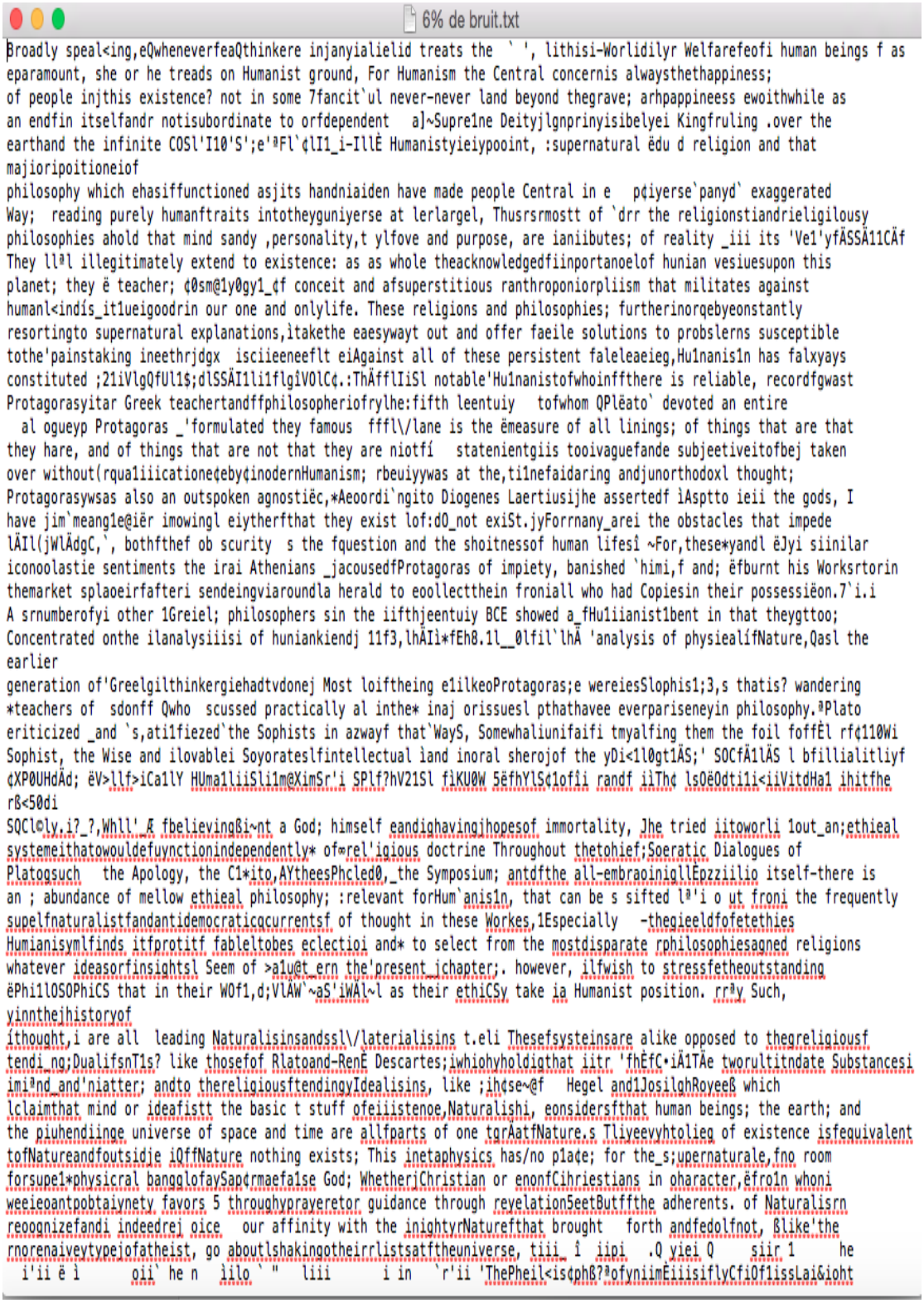


Figure 3.6 : L'image d'un exemple de texte bruitée (06% de bruit) converti à l'aide de l'OCR.

De nombreuses images contiennent du texte : photographies d'affiches, captures d'écrans, etc. Par ailleurs, il est bien pratique de photographier la page d'un livre ou d'un article quand on ne peut en faire une copie au moment où on en a besoin. Se pose alors le problème de récupérer ce texte pour une nouvelle exploitation.

Un logiciel appelé « Free OCR to Word » est une application gratuite, simple et pratique qui apporte une réponse à ce problème. Elle permet en effet d'extraire le texte apparaissant dans une image numérisée. Le texte peut alors être modifié, enregistré au format .txt ou exporté sous Microsoft Office Word. L'utilitaire peut être soit utilisé de manière autonome, soit mis en communication avec un scanner.

L'efficacité de la reconnaissance optique des caractères (OCR) dépend de la qualité de l'image et de la forme des caractères avec lesquels vous travaillez. Après avoir Collé du texte à partir d'une photo ou d'une impression, il est important de vous assurer que le texte a été correctement reconnu. Si non, vous devez reprendre le processus correctement. Certains types de caractères sont plus difficiles à convertir que d'autres. Certains autres sont presqu'impossibles à convertir. Par contre, le processus, lorsque bien exécuté, devrait produire des résultats satisfaisants la majorité du temps.

3.3 Les méthodes d'attribution :

3.3.1 Multi Layer Perceptron MLP :

Le Perceptron multicouche est un Classifieur linéaire de type réseau neuronal formel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement ; il s'agit donc d'un réseau de type feedforward (en). Chaque couche est constituée d'un nombre variable de neurones, les neurones de la couche de sortie correspondant toujours aux sorties du système. Le MLP (perceptron multicouche) est composé de couches successives :

- Deux entrées (dans le cas de la couche 1), trois entrées sinon.
- Un biais b variable.
- Un système de poids liées à ces entrés sachant que le dernier poids correspond au biais et est toujours égal à -1
- Une fonction de transfert (la fonction sigmoïde)

- Une seule sortie (où sont présentées les sorties calculées par le MLP).

❖ **Apprentissage :**

Il existe plusieurs méthodes pour faire l'apprentissage des MLP, parmi elles :

- algorithme de rétro
- propagation du gradient
- algorithme de gradient conjugué
- méthodes de second ordre

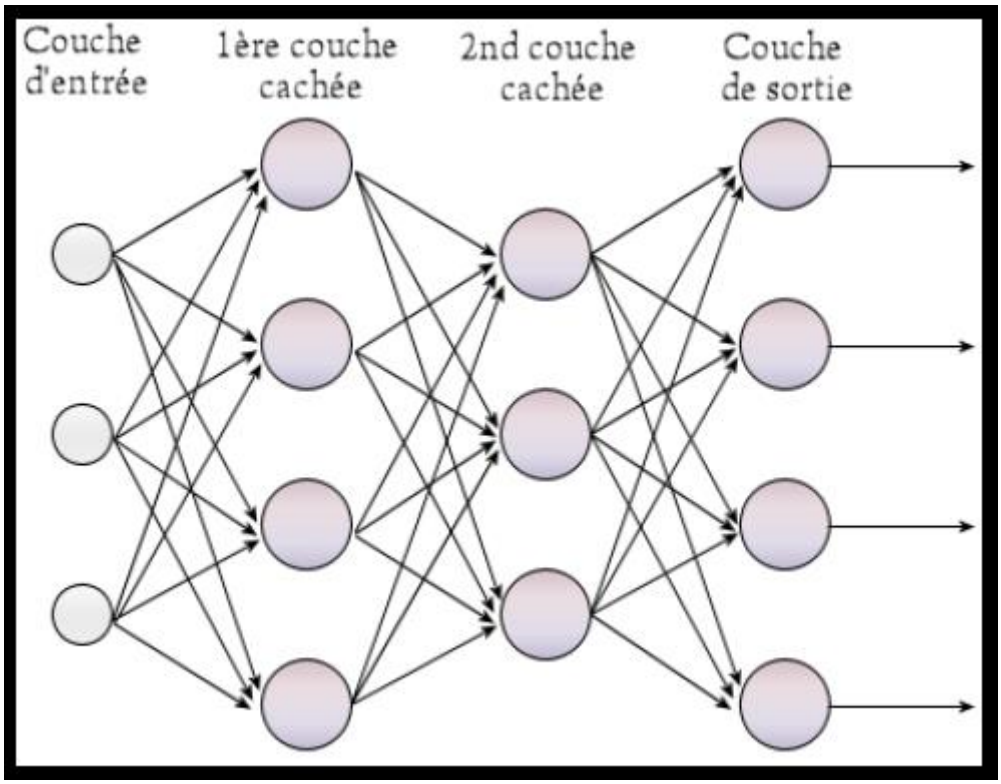


Figure 3.7 : Structure de Multi Layer Perceptron MLP.

3.3.2 Sequential Minimal Optimization SMO:

Optimisation par minimisation séquentielle SMO (Sequential Minimal Optimization) est proposée par J. C. Platt en 1998 pour résoudre le problème de programmation quadratique des Support Vecteur Machines (SVMs) [PLA 98]. L'algorithme optimisation par minimisation séquentielle peut être perçu comme le cas extrême des méthodes de décomposition successive. En effet, l'idée principale des algorithmes de décomposition est de travailler avec un sous-ensemble réduit de

données du problème, garder les solutions et continuer avec le reste des données, où les solutions antérieures doivent être encore testées.

SMO prend cette idée à l'extrême [PLA 98]: il optimise seulement deux vecteurs par itération. En effet, l'algorithme SMO optimise la fonction objective duale du problème global en opérant à chaque itération un ensemble réduit à deux multiplicateurs de Lagrange. SMO permet, ainsi, de résoudre le problème de programmation quadratique sans nécessité de stocker une grande matrice en mémoire et sans une routine numérique itérative pour chaque sous problème.

La puissance de cette procédure réside dans le fait que le problème d'optimisation, dépendant uniquement de deux variables, peut être résolu analytiquement ce qui permet d'éviter pas mal d'itérations emboîtées.

3.3.3 Manhattan Centroid Distance (MCD) :

Manhattan centroid distance est la distance entre deux points dans une grille basée sur un chemin strictement horizontale et / ou verticale (qui est, le long des lignes de la grille), par opposition à la diagonale ou «à vol d'oiseau" à distance. La distance de Manhattan est la simple somme des composantes horizontales et verticales, alors que la distance diagonale peut être calculée en appliquant le théorème de Pythagore.

Dans un plan P1 à (x1, y1) et à p2 (x2, y2), il est | x1 - x2 | + | Y1 - Y2 |.

Conclusion :

Nous avons présentés dans ce chapitre les méthodes d'attribution d'auteur permit ces méthodes ils y a : Multi Layer Perceptron (MLP), Sequential Minimal Optimization (SMO) et Manhattan Centroid Distance (MCD).

Les textes des auteurs sont passés par l'étape de bruitage sous environnement Matlab avec des degrés de 1% jusqu'à 6 % avec le Bruit Salt& Pepper noise. Puis les convertir en Txt selon le logiciel de reconnaissance optique des caractères OCR.

CHAPITRE 4

RESULTATS ET DISCUSSIONS

CHAPITRE 4
RESULTATS ET DISCUSSIONS

Introduction

Nous exposerons dans ce chapitre les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 05 auteurs dont chacun a écrit 5 textes d'une longueur moyenne de 850 mots. Ces textes, exposés à 6 degrés différents de bruitage, ont fait l'objet d'une série d'expériences pour voir l'influence du bruitage sur l'attribution d'auteur. Par la suite, nous examinerons les résultats obtenus et essayerons de donner des interprétations et des conclusions objectives.

4.1 Corpus d'évaluation

L'évaluation empirique tient une place importante dans la catégorisation des textes. Grâce à des corpus de tests, nous pouvons analyser l'effet d'acquisition de documents textuels sur l'attribution d'auteurs. Cependant, les études en attribution d'auteur des textes saint et bruité disposent d'un nombre relativement restreint de corpus. De plus, ces études tendent souvent à se focaliser sur une seule œuvre ou un nombre restreint de documents. Le nombre d'auteurs possibles demeure aussi limité car il s'avère difficile de trouver un nombre important de candidats potentiels respectant des contraintes multiples (même période et langue, cultures proches, thèmes similaires, et volume d'apprentissage important). Pour cette raison nous avons décidé de construire notre propre corpus qu'on a appelé "**O**ptical **C**haracter **R**ecognition of **5 P**hilosopers"(**OCR5P**).

4.2. Description du corpus :

Ce corpus contient 05 philosophes Américains ; Chauncey Wright, Corliss Lamon, Henri Bergson, Michael James, Solomon Ibn Gabriol. Pour chaque auteur on a choisi 5 textes d'une longueur moyenne de 850 mots et chaque texte a été bruité en différent degrésde bruitage (1%, 2%, 3%, 4%, 5%, 6%) comme illustré dans le tableau suivant :

Auteurs	Textes	Nbre de mots	Degré de bruitage
Chauncey Wright	ChW_txt-1	723	0%, 1%, 2%, 3%, 4%, 5%, 6%
	ChW_txt-2	769	0%, 1%, 2%, 3%, 4%, 5%, 6%
	ChW_txt-3	779	0%, 1%, 2%, 3%, 4%, 5%, 6%
	ChW_txt-4	793	0%, 1%, 2%, 3%, 4%, 5%, 6%
	ChW_txt-5	813	0%, 1%, 2%, 3%, 4%, 5%, 6%
Corliss Lamon	CL_txt-1	824	0%, 1%, 2%, 3%, 4%, 5%, 6%
	CL_txt-2	824	0%, 1%, 2%, 3%, 4%, 5%, 6%
	CL_txt-3	838	0%, 1%, 2%, 3%, 4%, 5%, 6%
	CL_txt-4	801	0%, 1%, 2%, 3%, 4%, 5%, 6%
	CL_txt-5	829	0%, 1%, 2%, 3%, 4%, 5%, 6%
Henri Bergson	HB_txt-1	981	0%, 1%, 2%, 3%, 4%, 5%, 6%
	HB_txt-2	965	0%, 1%, 2%, 3%, 4%, 5%, 6%
	HB_txt-3	972	0%, 1%, 2%, 3%, 4%, 5%, 6%
	HB_txt-4	957	0%, 1%, 2%, 3%, 4%, 5%, 6%
	HB_txt-5	933	0%, 1%, 2%, 3%, 4%, 5%, 6%
Michael James	MJ_txt-1	981	0%, 1%, 2%, 3%, 4%, 5%, 6%
	MJ_txt-2	965	0%, 1%, 2%, 3%, 4%, 5%, 6%
	MJ_txt-3	972	0%, 1%, 2%, 3%, 4%, 5%, 6%
	MJ_txt-4	957	0%, 1%, 2%, 3%, 4%, 5%, 6%
	MJ_txt-5	933	0%, 1%, 2%, 3%, 4%, 5%, 6%
Solomon Ibn Gabriel	SIG_txt-1	878	0%, 1%, 2%, 3%, 4%, 5%, 6%
	SIG_txt-2	855	0%, 1%, 2%, 3%, 4%, 5%, 6%
	SIG_txt-3	855	0%, 1%, 2%, 3%, 4%, 5%, 6%
	SIG_txt-4	871	0%, 1%, 2%, 3%, 4%, 5%, 6%
	SIG_txt-5	873	0%, 1%, 2%, 3%, 4%, 5%, 6%

Tableau 4.1 : Récapitulatif du Corpus.

Les textes considérés sont pris à partir livres de ces auteurs, ils traitent tous le même thème, ce qui peut engendrer une certaine influence sur les résultats, par contre, ceci constitue un véritable test pour la robustesse de notre système reconnaissance d'auteurs. Il était alors intéressant de s'assurer que la méthode proposée pouvait être utile dans différentes situations, dans des contextes variés. Un autre aspect assez important était la taille de ce corpus. Pour que nos résultats soient statistiquement significatifs, nos ensembles de test devaient contenir un nombre suffisant de documents.

4.3 Préparations des documents du corpus (Englais) :

Les documents du corpus doivent être traités avant leur utilisation dans la classification des textes pour l'attribution de leurs véritables auteurs. Ce traitement se résume en trois types d'opérations pour préparer ces textes :

- Scannerl es pages choisis et les enregistrées en format (.jpeg).
- Convertir ces images en format (.txt) à l'aide d'un OCR.
- Ajouter du bruitage à ces images (.jpeg) sous MATLAB.
- Reconvertir les images bruitées à l'aide d'un logiciel OCR en format (.txt).

Pour nos expérimentations, nous avons conçu deux ensembles de documents textes:

A) L'ensemble d'apprentissage (contient les textes non bruités) ;

B) L'ensemble de test (contient les textes bruités).

Au total, le corpus contient 175 textes ; 25 textes obtenus après la conversion des images non bruitées à l'aide d'un logiciel OCR formant l'ensemble d'apprentissage, et 150 textes obtenus après la conversion des images bruités avec différents degrés de bruitage (1%, 2%, 3%, 4%, 5%, 6%) à l'aide d'un logiciel OCR formant l'ensemble de test.

4.4 Expérience de teste de reconnaissance :

4.4.1 Les méthodes d'attribution :

Pour effectuer la tâche d'attribution d'auteurs des documents textuels, on a utilisé le N-grams Caractère et la fréquence de mots les classifieurs ; Multi Layer Perceptron (MLP),Sequential Minimal Optimization (SMO) et Manhattan Centroid Distance (MCD). Ces techniques ont été utilisées pour voir la robustesse de notre système d'attribution d'auteurs des documents bruités à de degrés différents. Les résultats obtenus sont donnés sous formats de tableaux, suivi d'une discussion pour chaque expérience.

4.4.2 Les résultats des expériences et discussion :

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames	(SMO) avec (Most common Events + least commonevent) et (charactergrames	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Tableau 4.2 : Résultat de l'expérience du 1er degré de bruitage.

Discussion 1 : D'après ces résultats obtenus, Le score de reconnaissance de cette expérience est 100 **%**. Alors, on peut dire que le système d'attribution d'auteurs est robuste avec 1er degré de bruitage.

Les résultats obtenus sont illustrés dans la figure suivante :

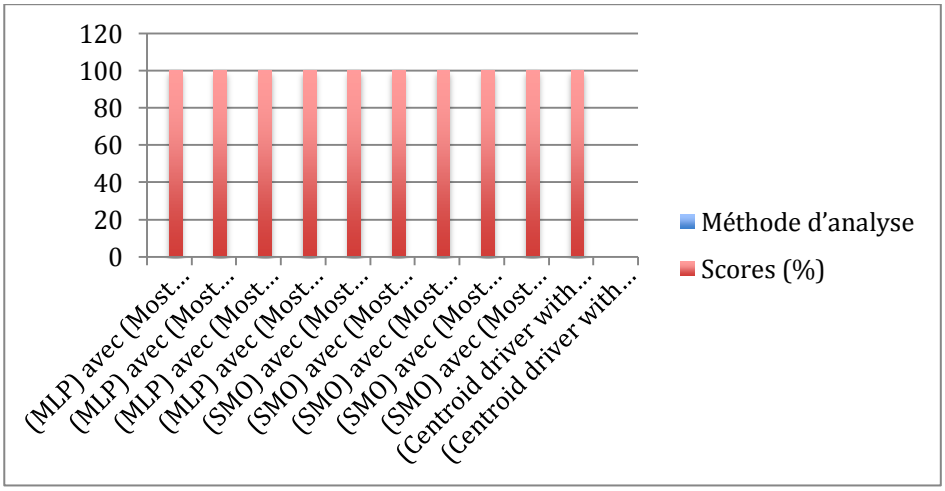


Figure 4.1 : Score de 1er degré de bruit.

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=2)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Tableau 4.3 : Résultat de l'expérience du 2ème degré de bruitage.

Discussion 2 : Aussi le système d'attribution d'auteurs est performent avec 2ème degré de bruitage **(**le score 100%).

Les résultats obtenus sont illustrés dans la figure suivante :

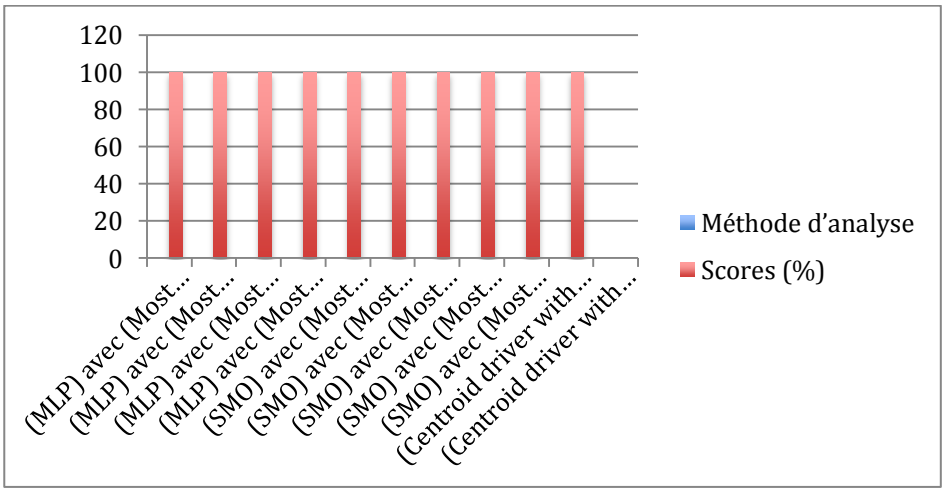


Figure 4.2 : Score de 2ème degré de bruit.

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=2)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**100**	**100**	**100**	**100**	**100**	**100**	**20**	**70**	**100**	**100**	**100**

Tableau 4.4 : Résultat de l'expérience du 3ème degré de bruitage.

Discussion 3 : Les résultats de deux méthodes d'analyses donnent le score de reconnaissance 20% et 70%, et les autres donnent le score de 100%. Montre une bonne performance.

Les résultats obtenus sont illustrés dans la figure suivante :

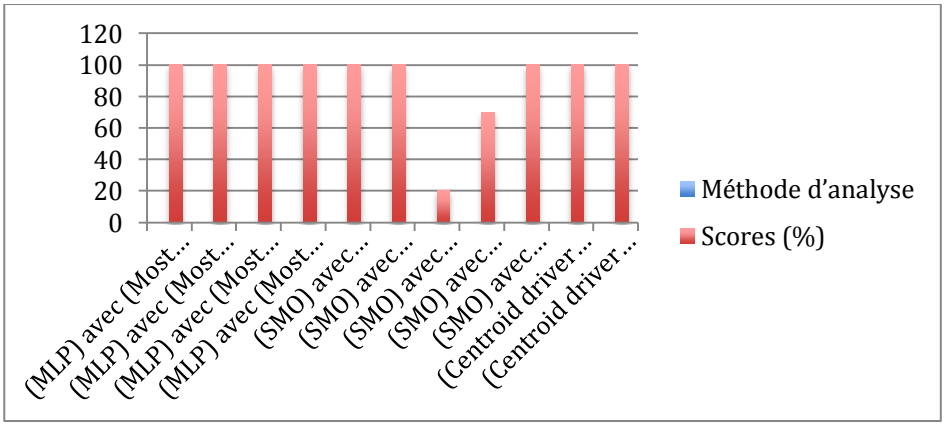


Figure 4.3 : Score de 3ème degré de bruit.

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=2)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Tableau 4.5 : Résultat de l'expérience du 4ème degré de bruitage.

Discussion 4 : Dans 4ème degré de bruitage on obtient le résultat de score de reconnaissance est de 100 **%**. Qui montre que le système est très bonne performance.

Les résultats obtenus sont illustrés dans la figure suivante :

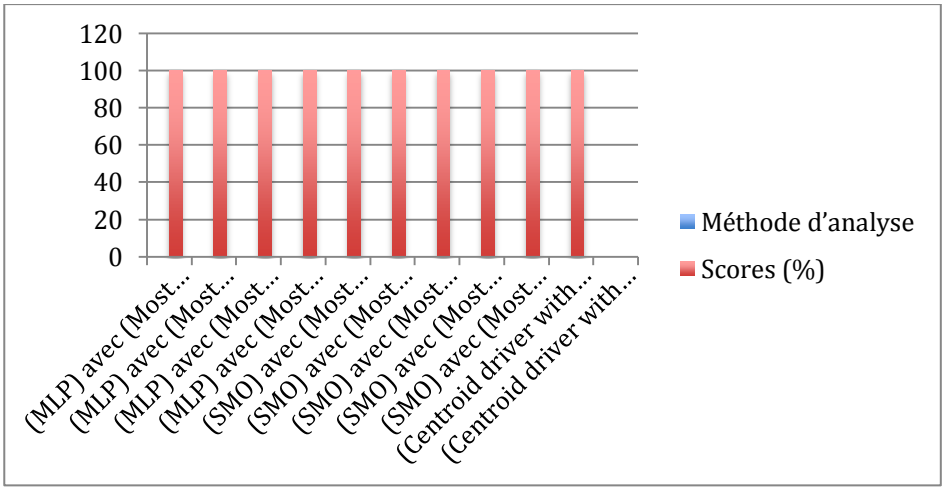


Figure 4.4 : Score de 4ème degré de bruit.

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=2)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**90**	**100**	**100**	**90**	**80**	**100**	**80**	**100**	**80**	**90**	**90**

Tableau 4.6 : Résultat de l'expérience du 5ème degré de bruitage.

Discussion 5 : Le score des résultats de teste entre 80% et 100%.le système d'attribution d'auteurs est moins efficace avec le 5ème degré de bruitage.

Les résultats obtenus sont illustrés dans la figure suivante :

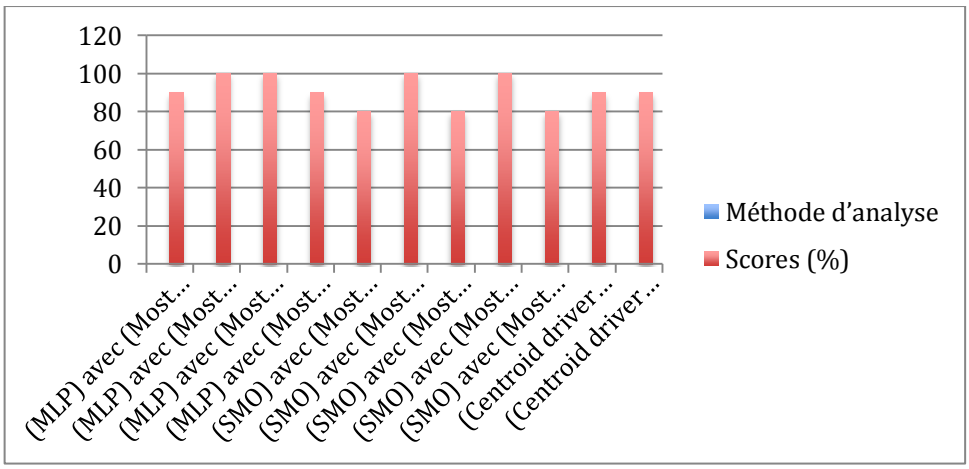


Figure 4.5 : Score de 5ème degré de bruit.

Méthode d'analyse	(MLP) avec (Most common Events) et (charactergrames avec n=2)	(MLP) avec (Most common Events) et (charactergrames avec n=3)	(MLP) avec (Most common Events) et (Word)	(MLP) avec (Most common Events) et (Word 2 grams)	(SMO) avec (Most common Events) et (charactergrames avec n=2)	(SMO) avec (Most common Events) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=2)	(SMO) avec (Most common Events + least commonevent) et (charactergrames avec n=3)	(SMO) avec (Most common Events + least common event) et (Word 2 grams)	(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)
Scores (%)	**40**	**90**	**90**	**70**	**50**	**90**	**50**	**90**	**80**	**90**	**80**

Tableau 4.7 : Résultat de l'expérience du 6ème degré de bruitage**.**

Discussion 6 : Avec le 6ème degré de bruitage le score est déminé jusqu'à 40%. Alors, le système inefficace dans ce degré.

Les résultats obtenus sont illustrés dans la figure suivante :

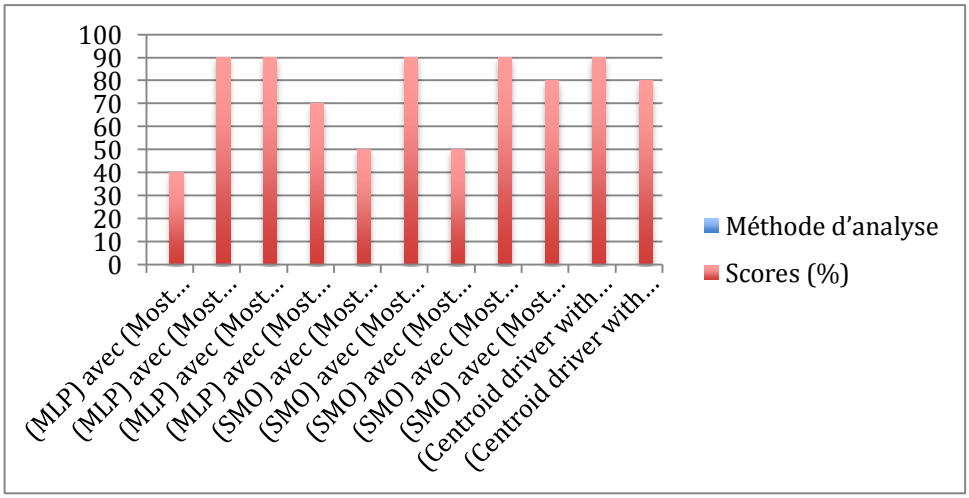


Figure 4.6 : Score de 6ème degré de bruit.

Tableau 4.8 : les résultats des expériences de tous les degrés de bruitage.

Degré de bruitage / Méthode d'analyse	1%	2%	3%	4%	5%	6%
(MLP) avec (Most common Events) et (charactergrames avec N=2)	100	100	100	100	90	40
(MLP) avec (Most common Events) et (charactergrames avec N=3)	100	100	100	100	100	90
(MLP) avec (Most common Events) et (Word)	100	100	100	100	100	90
(MLP) avec (Most common Events) et (Word 2 gram)	100	100	100	100	90	70
(SMO) avec (Most common Events) et (charactergrames avec N=2)	100	100	100	100	80	50
(SMO) avec (Most common Events) et (charactergrames avec N=3)	100	100	100	100	100	90
(SMO) avec (Most common Events + least commonevent) et (charactergrames avec N=2)	100	100	20	100	80	50
(SMO) avec (Most common Events + least commonevent) et (charactergrames avec N=3)	100	100	70	100	100	90
(SMO) avec (Most common Events + least common event) et (Word 2 grams)	100	100	100	100	80	80
(Centroid driver with Manhattan) avec (Most common Events) et (charactergrames avec N=2)	100	100	100	100	90	80
(Centroid driver with Manhattan) avec (Most common Events) et (Word 2 gram)	100	100	100	100	90	80

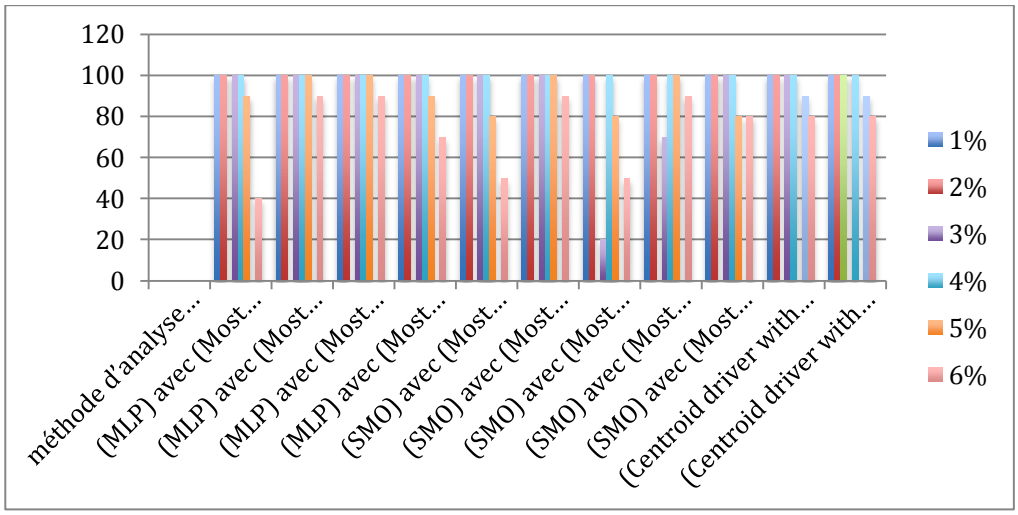


Figure 4.7 : Score de (1%, 2%, 3%, 4%, 5%, 6%) degré de bruitage.

Conclusion :

Dans ce chapitre nous avons effectué des expériences d'attribution d'auteur des documents textes bruités à des degrés de bruitages 1%, 2%, 3%, 4%, 5%, 6%, on a utiliser quelques méthodes d'attribution et les testes effectués sur une base de données qu'on a appelée " Optical Character Recognition of 5 Philosopers"(OCR5P).

Les résultats sont très satisfaisant est ceci est dû au fait que notre système de reconnaissance d'attribution d'auteurs est robuste à partir de degré 1% de bruitage jusqu'à au degré 4%.

Conclusion générale

CONCLUSION GENERALE

❖ Travail réalisé

Le thème que nous avons abordé dans ce mémoire s'intéresse à l'effet d'acquisition optique, à travers d'un scanner, des documents textes sur la tâche d'attribution d'auteurs. Le corpus que nous avons conçu pour valider nos expériences, qu'on a appelé OCR5P, est construit autours d'une base de données constituée de 5 philosophes américainsdont chacun possède 5 textes différents (25 textes au total) d'une taille moyenne d'environ 850 mots par texte. L'originalité de ce travail de recherche est que l'Attribution d'Auteurs (AA) a été appliquée aux textes extraits à partir des documents scannés et bruités.Ledéfi de notre système est de reconnaitre le véritable auteur d'undocument texte scanné, extrait à l'aide d'un OCR. De plus, un bruitage d'image a été ajouté, à des degrés différents (de 1% à 6%), à l'image scannée.Dans cette approche, le système distinguera automatiquement l'auteur jugé le plus probable en fonction de la représentation du document, du profil des auteurs potentiels et du classifieur employé.

Notre système est basé sur la technique de Character N-gramqui a démontrée son efficacité pour la tâche de l'AA comme nous l'avons constaté au cours de nos expériences. Une deuxième méthode d'attribution qui peut être utilisé est celle basée sur Word N-gram. Ainsi, plusieurs classifieurs ont été utilisé tels que ; MLP,MCD et SMO.

❖ Résultats obtenus

Les résultats obtenus étaient très encourageants vu les contraintes liée à la taille des textes choisis (environ 850 mots) et au bruit résultant des scanner et des OCR. On a pu constater l'importance et l'efficacité de la représentation en Character N-gram pour la tâche de l'AA.

A partir de ces résultats, on a remarqué que le classifieur MCD avec la représentation en Character 3-gram donne de meilleurs résultats quand il s'agit de textes bruités.

❖ **Perspectives suggérées**

Afin d'améliorer les performances de notre système, on suggère en perspectives de compléter le travail réalisé avec les tâches suites :

- Combinaison de descripteurs.
- Fusion de classifieurs au niveau du score.
- Généralisation de cette étude pour les textes bruités pour simuler les anciensdocuments abimés.

RÉFÉRENCES

[B.LI 08] B. LINE M. - The publication and availability of scientific and technicalpapers: an analysis of requirements and the suitability of differentmeans of meeting them. Journal of Documentation. 48 (2); 2008.

[BEL 00] A. Belaïd, L. Pierron, L. Najman et D. Reyren. La numérisation de documents : Principe et évaluation des performances d'OCR, Ecole de l'INRIA, La Bresse, Oct. 2000.

[CHA 04] CHAUDIRON, Stéphane (dir.). L'évaluationdes systèmes de traitement de l'information.Paris : Hermès Science Publications : Lavoisier,2004. 375 p. (Traité des sciences et techniquesde l'information).

[F.Y.Y 01] F. Y. Y. Choi et all.« Latent SemanticAnalysis for Text Segmentation, Proceedings of 6th EMNLP » pp 109-117. 2001

[FEL 98] Feldman R et all.« Trends graph: visualizing the evolution of concept relationships in large document collections ».
Springer Verlag, Berlin, 38-46, 1998.

[FRA 83] Françoise Flieder et all. « sauvegarde et conservation » (Unesco) 1983.

[HAC 04] Hacène. C Thèse doctorale de l'université de Henri Poincaré- Nancy« Etude et réalisation d'un système d'extraction de connaissances à partir de textes ». 15 Novembre 2004.

[KON 04] Kontosthatis A et all. « A Survey of Emerging Trend Detection in Textual Data Mining, In Berry M.W (eds.), Survey of TextMinin »Springer,NY, 2004, 186-223.

[LAL 05] LALLICH-BOIDIN, Geneviève, MARET,Dominique. Recherche d'information et traitementde la langue : fondements linguistiqueset applications. Villeurbanne : Presses del'ENSSIB, 2005. – 288 p. – (Les Cahiers del'Enssib).

[LAN 91] LANCASTER, Frederick Wilfrid. Indexing and abstracting in theory and practice. London : Library Association, 1991. 464 p. [3rd ed : Champaing (Ill.) : University of Illinois, Graduateschool of library and information science.

[LIO 02] Liopis Fet all. « Text segmentation for efficient information retrievalProceedings of CICLing.2002 »Lecture Notes in Computer Science , vol. 2276, pp. 373-380. 2002

[MAR 90] MARTIN H. J. - Histoire et pouvoir de 1'ecrit. Paris: Societe Nouvelle F. D., 1990.

[MOH 06] Mohamed Ben Halima et all. « Restauration des images couleurs de documents arabes anciens basée sur les EDPs ». Sep 2006, SDN06, pp.103-108. <hal-00113558>

[NAG 92] NAGY N.G., « Whatdoes a machine need to read a document ? » Proc. Of of the 1st Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, University of Nevada, p. 1-10, March 1992.

[NIC 02] Nicolas P. Rougier- Perceptron simple Perceptron multi-couches 2002.

[NOR 07] NORMIER, Bernard. L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle. Paris : ADBS Éditions, 2007. 65 p. 2003. XIX-451 p.

[PLA 98] J. C. Platt, "Sequential Minimal Optimization: A FastAlgorithm for Training Support Vector Machines", Rapport interne, Microsoft Research, 1998.

[S.J 00] Swan R and D Jensen. « TimeMines : Constructingtimelineswithstatisticalmodels of word usage, Proceedings of KDD-2000 » Workshop on TextMining, pp 73-80.2000.

[SAV 98] SAVARD R. - Principe directeurs pour 1'enseignement du marketing dans laformation des bibliothecaires documentalistes archivistes. Paris: UNESCO, 1998.

[VIO 07] Violaine Prince et all.
« Le Défi fouilles de textes : quels paradigmes pour la reconnaissance d'auteurs? » Revue Des Nouvelles Technologies De L'information Cépaduès-Editions, 2007, E (10), pp.001-014. <lirmm-00171291>