REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE
N°: …16.. /INST/ 2021

DOMAINE: SCIENCES ET TECHNOLOGIE
FILIERE: ÉLECTRONIQUE
OPTION: INSTUMENTATION

Mémoire présenté pour l'obtention
du diplôme de Master Académique

Par : BACHA Bochra et HADLI Imane

Intitulé

Attribution d'auteurs des textes arabes traduits en plusieurs langues en utilisant les traducteurs automatiques

Soutenu publiquement le : 20 /06/ 2021 devant le jury composé de:

Dr. LADJAL Mohamed	Université M'sila	Président
Dr. KHENNOUF Salah	Université M'sila	Encadreur
Dr. LAIB Abderrazak	Université M'sila	Co-Encadreur
Dr. DJERIOUI Mohamed	Université M'sila	Examinateur

Année universitaire : 2020 /2021

REMERCIEMENTS

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

Un grand merci à nos encadreurs Dr. KHENNOUF Salah et Dr. LAIB Abderrazak à qui nous devons beaucoup, pour leurs attentions, leurs disponibilités, leurs conseils et leurs sympathies.

Nos remerciements vont également aux membres du jury, chacun par son nom, pour avoir accepté de faire partie du jury d'évaluation de ce modeste travail.

Nous remercions 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.

B. BACHA & I. HADLI

DEDICACE

Je dédie ce modeste travail à :

Mes chers parents qui m'ont aidé et m'ont encouragé pendant toute ma vie d'étude et d'être ma source de bonheur et de réussite.

Mes chers frères et sœurs.

Mon marie

Mes chers amis(es).

Ma partenaire et ma chère amie avec qui je partage ce modeste travail :

Imane

Et à tous les collègues de ma promotion.

Bouchra

DEDICACE

Je voudrais dédie ce modeste travail

A ma mère pour ses sacrifices en témoignage de tout mon affection au long de mes études.

A mon marie Turki Houssam.

A ma sœur Khaoula que j'adore.

A mes frères: Houssam, Chokri, Alla Eddine, Charaf Eddine.

Mes chers amis(es).

Ma partenaire et ma chère amie avec qui je partage ce modeste travail :

Bouchra

Et à tous mes collègues de ma promotion.

Imane

Liste des abréviations

AA : L'Attribution d'Auteur

IA : Intelligence Artificielle

LDA : Linear Discriminant Analysis

LS : Langue Source

MLP : Multi Layer Perceptron

OCR : Optical Recognition Character

PAHO : Pan American Health Organization

RNA : Réseaux de Neurones Artificiels

SVM : Support Vector Machine

TA : Traducteur Automatique

TAO : Traduction Assistée par Ordinateur

THAM : Traduction Humaine Assistée par la Machine

TREC : Text **RE**trieval **C**onference

Liste des tableaux

Liste des figures

Table de la matière

Introduction générale

Introduction générale

L'Attribution d'Auteur (AA) consiste à prédire l'auteur d'un texte à partir d'un ensemble de candidats. La difficulté augmente quand les objets d'étude proviennent du Web où se côtoient différents genres textuels, styles et langues. Dès lors, les recherches en AA peuvent se concentrer sur certains de ces problèmes : le passage à l'échelle quand un grand nombre d'auteurs candidats est considéré ou l'indépendance vis-à-vis de la langue lorsque les ressources linguistiques sont rares ou manquantes.

Le développement technologique, y compris l'informatique et l'intelligence artificielle, a dû résoudre de nombreux problèmes différents de l'attribuer à l'auteur. La mention de l'auteur n'y est pas alors systématiquement présente. La fouille de données textuelles permet de classer les auteurs par catégorie (par genre, âge ou par opinion politique) ou en tant qu'individu. Ce dernier cas de figure est appelé le problème d'Attribution d'Auteur (AA). Cela consiste à deviner l'auteur de textes à partir d'un ensemble de candidats. Ainsi, cette tâche peut être vue comme un sous-domaine de l'apprentissage automatique supervisé. Techniquement cela consiste à définir une nouvelle paire reliant un texte à un auteur. Ces méthodes peuvent aussi être utilisées pour savoir si un auteur est facilement détectable via ses productions dans un flux de textes. Ce domaine est aussi connu sous le nom de writeprint, en référence aux termes anglais « écriture » (write) et « empreinte digitale » (fingerprint).

Le danger qui menace l'auteur à notre époque est l'attribution de son idée aux mémoires et écrits d'autrui, ou le prétendu vol scientifique, qui vise à voler les citations et les écrits de l'écrivain et à les lui attribuer.

La tâche d'AA est le plus souvent abordée sous l'angle de la stylométrie (ou étude du style). L'hypothèse est qu'un auteur laisse involontairement dans son message textuel des indices qui peuvent mener à son identification. Elle définisse un ensemble de traits (numériques) qui demeurent relativement constants pour un auteur donné et qui distinguent suffisamment son style d'écriture des autres auteurs.

Nos objectifs

Dans ce travail de recherche, on vise à faire une étude et analyse sur les performances des techniques d'identification d'auteur à partir de documents écrits. Et les textes corrigés et traduits en plusieurs langues pour voir la possibilité du programme d'identifier les textes volés Pour cela, plusieurs descripteurs seront utilisés pour modéliser le style de chaque auteur, et un classifieur SVM, MLP et LDA et connaitre la meilleure technique.

Pour les textes étudier on utilise l'OCR l'acronyme de Optical Character Recognition. Sa technologie permet de reconnaître automatiquement les caractères grâce à un mécanisme optique. Dans le cas des êtres humains, nos yeux sont un mécanisme optique. L'image vue par les yeux est entrée pour le cerveau. La capacité de comprendre ces entrées varie chez chaque personne en fonction de nombreux facteurs. L'OCR est une technologie qui fonctionne comme la capacité humaine de lecture. Bien que l'OCR ne soit pas en mesure de rivaliser avec les capacités de lecture humaines, l'OCR peut reconnaître à la fois le texte manuscrit et imprimé. Mais les performances de l'OCR dépendent directement de la qualité des documents d'entrée. L'OCR est conçu pour traiter des images composées presque entièrement de texte, avec très peu d'encombrement non textuel obtenir à partir d'une image capturée par une caméra mobile. Cette application est destinée au système d'exploitation mobile Android qui combine le moteur OCR open source de Google, Tesseract, texte moteur de reconnaissance OCR.

Dans ce travail, nous discuterons de l'effet de la TA sur l'attribution d'auteur, ces textes ont été obtenus après avoir reconnu les caractères (OCR) appliqués aux textes scannés :

- Une base de données texte est conçue pour valider les techniques que nous avons proposées.
- Développer une base de données conçue pour valider les techniques proposées.

Structure de la thèse

Ce mémoire est structuré en trois chapitres, comme suit :

Le premier chapitre la définition de l'attribution d'auteurs ainsi que les différents types de systèmes, les notions fondamentales de l'exploration de texte, la stylométrie et l'attribution de textes, la langue arabe.

Le deuxième chapitre aborde la méthodologie de recherche qui a été adopté dans ce mémoire ainsi que les approches et techniques proposées pour l'attribution d'auteurs des documents textuelles.

Dans le troisième chapitre on exposera les séries d'expériences d'attribution d'auteurs effectuées sur la base de données textuelle (ou Corpus) que nous avons conçu pour cette fin et qui contient deux catégories de textes.

CHAPITRE-1

GENERALITES SUR L'ATTRIBUTION D'AUTEUR

CHAPITRE-1
GENERALITES SUR L'ATTRIBUTION D'AUTEUR

Introduction

L'Attribution d'Auteur d'un texte inconnu ou douteux est l'un des plus anciens problèmes de la statistique appliquée à la littérature. Dans ce chapitre, nous présentons des généralités sur l'attribution d'auteur, ensuite la stylométrie, puis catégorisation automatique de textes. Enfin, on a présente des définitions et quelques types du plagiat.

1.1 Attribution d'Auteur

L'attribution d'auteur (AA) est le processus visant à identifier la paternité probable d'un document donné, compte tenu d'une collection de documents dont l'auteur est connu. L'attribution de la paternité devient un problème important car la gamme d'informations anonymes augmente avec une croissance rapide de l'utilisation d'Internet dans le monde entier. Application de la paternité de l'attribution comprend la détection du plagiat, déduire l'auteur de communications inappropriées qui envoyé de manière anonyme ou sous un pseudonyme, ainsi que la résolution de questions historiques paternité peu claire ou contestée [1].

L'attribution de la paternité est le moyen de déterminer l'auteur d'un texte lorsqu'il n'est pas clair qui l'a écrit. Il est utile lorsque deux personnes ou plus prétendent avoir écrit quelque chose ou quand personne ne veut (ou ne peut) dire qu'elle ou il a écrit la pièce.

1.2 Historique d'Attribution d'Auteur

Bien qu'elle n'ait pris une réelle importance en histoire de l'art qu'à partir du xix^{e} s., l'attribution avait déjà existé de temps en temps, au cours des siècles antérieurs, spécialement dans le milieu italien. Il suffira à cet égard de citer le " livre " de Vasari, ce recueil de dessins pour l'encadrement desquels il dessinait des éléments décoratifs, considérés comme caractéristiques du style de l'artiste auquel il attribuait le dessin. À les comparer entre elles, les attributions faites par des historiens comme Vasari, Baldinucci, Lanzi ou d'Agincourt permettent de connaitre l'idée que l'on se faisait jadis de maîtres comme Cimabue, Giotto ou Masaccio : à ce titre, elles nous intéressent surtout du point de vue de l'histoire du gout [2].

Parallèlement, la création de grands musées nationaux (londres, berlin) stimule la pratique de l'attribution. Pour permettre la rédaction des catalogues et une politique intelligente des acquisitions. Des recherches étroitement liées à celles des attributions se développent dans la 2ème moitié du siècle : on tente de définir la personnalité des artistes grâce aux œuvres qu'on leur attribue en fonction de leur style. C'est à la même époque que se généralise l'usage des "noms de commodité " (de l'allemand Notnamen), appellation qui fait bien ressortir la distinction nécessaire entre la personnalité civile et la personnalité esthétique de tel maître connu seulement par ses œuvres [2].

1.3 Etat de l'art

Plusieurs recherches ont été menées à l'AA au cours de ces dernières années. Avec la quantité croissante de documents sur Internet, et comme la plupart des écrits sont anonyme, l'attribution de la paternité devient importante. Les recherches portent sur différentes propriétés des textes. On distingue deux propriétés différentes des textes qui sont utilisés dans classification ; le contenu du texte et le style de l'auteur. L'analyse statistique du style littéraire complète la bourse littéraire traditionnelle car elle offre un moyen de saisir le caractère souvent insaisissable de l'auteur style en quantifiant certaines de ses caractéristiques. La majorité des études stylométriques utilisent des éléments de langage et la plupart de ces éléments sont à base lexicale [3].

1.4 Définition des traits d'un auteur

Les traits utilisés en AA peuvent être séparés en différents groupes :

- Valeurs numériques associées à des mots (nombre de mots dans les textes, nombre de caractères par mot, nombre de bi-grammes/tri-grammes de caractères au sein de ces mots) appelés aussi des traits lexicaux ;
- Valeurs associées à la syntaxe des phrases (effectifs des mots outils, des monogrammes/bi-grammes/tri-grammes de ces mots outils ou des séquences de parties du discours) ;
- Valeurs numériques associées à des unités plus grandes (nombre de paragraphes ou encore longueur moyenne des paragraphes), autrement dit des traits structurels ;
- Valeurs associées avec le contenu thématique (des sacs de mots, des n-grammes de mots clefs) ;
- Particularités en rapport avec les pratiques individuelles (telles que les fautes d'orthographe ou de frappe).

Parmi ces traits, certains sont spécifiques à des types de langue et de graphie. Si découper un texte en mots est aisé dans certains cas (en définissant un mot comme une chaîne de caractères entourée d'espaces), ce n'est pas une tâche triviale en chinois ou en japonais. Les approches exploitant les n-grammes de caractères apparaissent comme étant les plus simples pour traiter n'importe quelle langue, ainsi que les plus performantes [4].

1.5 Représentation des textes et des auteurs fondus sur les traits

Un même trait peut être attribuée à plusieurs pairs (texte, auteur) mais chaque texte et auteur ne partagent pas pour autant un grand ensemble de traits. Différents ensembles de traits peuvent être définis pour représenter des textes (et par extension, pour représenter des auteurs). Considérant les méthodes d'AA existantes, deux catégories principales de traits Peuvent être définies :

- Traits hors-ligne : traits a priori considérés pertinents pour cette tâche avec une connaissance préalable, comme deux largement décrit par Chaski (2001). Ils peuvent être définis quand le corpus à traiter n'est pas encore collecté.
- Traits en-ligne : traits définis pendant le traitement (dans le cas de méthodes supervisées, en fonction des corpus d'entraînement et de test, comme le modèle de langue de caractères décrit par Peng et al. (2003)). Ils ne peuvent être définis que lorsque le corpus à traiter est complet.

Les traits en-ligne renvoient naturellement à la notion d'indépendance vis-à-vis des langues, aucun a priori n'est émis avant le traitement du corpus et aucune ressource linguistique extérieure n'est exploitée. La méthode décrite dans cet article suit ce principe [4].

1.6 Etapes d'Attribution d'Auteur

- Un processus complet d'attribution de l'auteur consiste en rassemblement des textes qui sont les observations à classer.
- Une méthode d'extraction de caractéristiques qui calcule les informations numériques ou symboliques issues de ces observations.
- Un système de classification ou de catégorisation qui fait le classement à partir de ces observations [3].

1.7 Stylométrie

La technologie moderne et plus particulièrement l'informatique permet de nos jours d'analyser la trame stylistique d'un texte lorsque l'identité d'un auteur est contestée. Par

l'analyse stylométrie, il est maintenant possible de déterminer avec un haut degré de certitude si telle ou telle personne est l'auteur ou non de l'Ouvrage concerné [3].

1.7.1 Petit historique de la stylométrie

Les premières mentions de la stylométrie pour identifier des auteurs sont apparues en 1851. Mais, compte tenu de la difficulté des mesures à effectuer, les premières études crédibles ont dû attendre l'arrivée des ordinateurs modernes, pour leur précision de comptage et leur traitement à grande vitesse des données. Au début des années 1980, une équipe de chercheurs a travaillé pour affiner et rendre plus performantes les techniques de la stylométrie. Les travaux montrent que les méthodes de comptage et de comparaison entre différents textes ont été grandement améliorées.

La stylométrie continue à évoluer vers une fiabilité et une sensibilité toujours plus grandes, elle a atteint un niveau qui permet la mise en œuvre d'une technique de mesure rigoureuse qui donne des réponses fiables dans l'analyse des textes de plusieurs milliers de mots d'un même auteur, en flux libre [5].

1.7.2 Définition de la stylométrie

La stylométrie est l'étude quantitative du style littéraire a des méthodes informatiques de lecture distante. Elle se base sur l'observation fait que chaque auteur a tendance à écrire de façon relativement constante, reconnaissable et unique. C'est un ensemble de techniques à l'intersection de la linguistique et de la statistique, dont le but est d'identifier le style de documents textuels. Le style d'un texte est une caractéristique de son contexte d'écriture au sens large : son auteur, son époque, son « genre », etc…

La stylométrie tente de montrer qu'un texte est écrit dans un style différent d'une collection d'autres textes. Cette différenciation permet donc, dans une certaine mesure, de déterminer si un « anonyme » a été écrit par un auteur précis, et surtout de déterminer si un texte n'a pas été écrit par un auteur précis [5].

1.7.3 Caractéristiques utilisées dans la stylométrie

Chaque individu possède son propre vocabulaire, parfois riche, parfois limité. Bien qu'un vocabulaire étendu soit généralement associé à une littérature de qualité, ce n'est pas toujours le cas. Certaines personnes écrivent en phrases courtes, tandis que d'autres préfèrent les phrases complexes comportant plusieurs propositions. Il n'y a pas deux auteurs qui utilisent les points-virgules, les tirets et autres signes de ponctuation exactement pareil.

L'identification de l'auteur d'un texte anonyme constitue cependant l'une des applications les plus courantes de la stylométrie. Il est parfois possible de découvrir l'identité de l'auteur d'un texte en mesurant certaines caractéristiques de ce texte, comme la longueur moyenne des phrases ou le rapport entre le nombre d'articles définis et indéfinis [5].

Ces mesures sont ensuite comparées avec celles observées dans textes dont les auteurs sont connus Lorsque l'on parle de trame non –contextuelle, il s'agit de mots qui sont souvent interchangeables ou qui peuvent même être omis perte de la signification générale du texte. Ces mots contribuent peu à l'information contextuelle et sont souvent ignorés consciemment, aussi bien par le lecteur que par l'auteur.

Ces mots constituent typiquement 20a 45 % du texte total, Ce qui permet d'avoir un nombre important de choix statistique, et plus les mesures statistiques sont nombreuses, plus leurs résultats sont fiables [3].

1.8 Catégorisation automatique de textes

1.8.1 Définition de la catégorisation de textes

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également modèle de prédiction, est estimée par un apprentissage automatique (traduction de machine Learning method). Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit ensemble d'apprentissage, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'erreur en prédiction **[6]**.

1.8.2 Historique de la catégorisation de textes

C'est une discipline assez ancienne, en 1627, Gabriel Naudé propose un classement selon cinq grands thèmes : théologie, jurisprudence, histoire, sciences et arts, belles lettres. Le désir de maîtriser l'Univers se fait sentir dans la multiplication des encyclopédies. L'encyclopédie de Diderot (parue entre 1751 et 1772) est organisée selon l'ordre alphabétique avec des renvois associatifs alors que celle de Panckoucke (parue de 1776 à 1780) suit une organisation méthodique selon un ordre arborescent (Fayet & Scribe, 1997).

Le système de classification par thème, apparu dès les débuts de l'écriture et institutionnalisé à Alexandrie conduisit à la création par Dewey, en 1876, d'un système de classification « universel ». Il s'agit d'une classification documentaire de type

encyclopédique. Toutefois l'idée d'effectuer la classification de textes par des machines remonte au début des années 60 et qui a connu des progrès considérables à partir des années 90 avec l'apparition d'algorithmes beaucoup plus performants qu'auparavant.

Jusqu'au début des années 80, pour construire un classifier, il fallait consacrer d'importantes ressources humaines à cette tâche. Plusieurs experts éditaient des règles manuellement puis les affinaient au fur et à mesure des tests. L'avènement des de l'AA s'est donc traduit par un gain de temps conséquent. Il n'est plus nécessaire par exemple de reconfigurer tout le système en cas de changement d'arborescence. Ces évolutions technologiques et algorithmes avancées font aujourd'hui de la catégorisation un outil fiable.

Au début des années 90, les travaux proviennent essentiellement de la communauté de Recherche d'Information (RI). En effet, les méthodes de numérisation, les algorithmes de classification et les méthodologies de test ont été adaptés à la CT en particulier au cours des conférences TREC (**T**ext **RE**trieval **C**onference)

La communauté d'Apprentissage Automatique (AA) s'est intéressée elle aussi à ce problème il y a une dizaine d'années en le considérant comme domaine d'application à ces algorithmes de reconnaissance des formes. Actuellement, les méthodes de numérisation de texte restent largement inspirées de la RI alors que les classifier les plus performants sont issus de l'AA. Une autre communauté composée essentiellement de statisticiens et de linguistes, traite également le problème de la CT en s'appuyant sur les méthodes d'analyse de données. Le but ici n'est pas de créer un système qui classe automatiquement des documents sans intervention humaine mais d'extraire des informations synthétiques du corpus. Les problématiques traitées ici sont par exemple l'étude des genres littéraires ou la détermination de l'auteur d'un texte **[7]**.

1.8.3 Problématique de la catégorisation des textes

La problématique de la catégorisation peut se récapituler à trouver un prototype ou une fonction mathématique capable d'assigner automatiquement un document a une catégorie avec le plus grand taux de réussite possible, cette fonction se traduit par

Ω: D x C → {Vrais ; Faux}

Où : d représente l'ensemble des documents et C représente l'ensemble des catégories. Pour chaque couple (di, ci) appartenant à D×C, la fonction de catégorisation (Ω) renvoi vrai si le document appartient à la catégorie et Faux si non. Dans les systèmes de catégorisation

basés sur des méthodes d'apprentissage, la fonction de décision sera évaluée à l'aide d'un corpus d'entraînement. Cette fonction peut faire intervenir un grand nombre de valeurs numériques qu'un humain ne peut pas saisir. La détermination de cette fonction est appelée phase d'apprentissage, tandis que l'utilisation de cette fonction pour attribuer une catégorie à un document se fera pendant la phase de test [8].

1.8.4 Systèmes de Catégorisation

L'objectif de la Catégorisation Automatique des Textes (CAT) est de classer de façon automatique les documents dans des catégories qui ont été définies soit préalablement par un expert (catégorisation supervisée) ou classification, soit de façon automatique (catégorisation non supervisée) ou clustering (partitionnement de données) [8].

1.8.4.1 Catégorisation supervisée (Classification)

Ainsi, la classification de textes correspond à procédure d'affectation d'un ou de plusieurs catégories ou classes prédéfinies à un texte. Elle correspond à la catégorisation supervisée pour l'apprentissage automatique et à la discrimination en statistiques alors que la recherche d'informations utilise des termes plus proches de l'application concernée : filtrage ou routage.

Cette problématique a par ailleurs dernièrement trouvé de nouvelles applications dans les domaines du traitement du langage tels que: l'affectation de sujets en recherche d'information, l'aide de l'utilisateur pour l'indexation de documents (Hayes & Weinstein, 1990), la veille technologique, le filtrage personnalisé des documents intéressant (Lang, 1995) et l'amélioration de la recherche sur le web (Armstrong & all, 1995). Aujourd'hui, cette problématique utilise largement des méthodes issues de l'apprentissage automatique et beaucoup d'algorithmes d'apprentissage supervisé lui ont été appliqués (Naïve bayes, K-plus proches voisins, arbres de décision, machines à vecteurs support, réseaux de neurones, etc.…) [8]

1.8.4.2 Catégorisation non Supervisée (Clustering)

Quand l'ensemble des catégories n'est pas donné au départ, et qu'il s'agit de le créer en regroupant les textes en classes qui possèdent un certain degré de cohérence interne, on est dans un contexte de catégorisation non supervisée pour l'apprentissage automatique. Ce type de catégorisation non supervisée consiste à trouver de maniéré automatique une organisation cohérente a un groupe de documents homogènes pour construire des regroupements cohérents

(des classes ou clusters), elles correspondent aux statistiques et au clustering, qui est également le terme utilisé en recherche d'informations. Le clustering consiste donc, à diviser les objets (dans notre cas des textes) en groupes sans connaître à priori leurs classes d'appartenance. Les techniques pour réaliser de tels regroupements constituent un domaine d'étude très riche, qui a donné lieu à de multiples propositions dont le recensement n'est pas l'objet de cette étude [8].

1.8.4.3 Classification supervisée vs classification non supervisée

La classification supervisée consiste à identifier la classe d'appartenance d'un objet à partir de certains traits descriptifs. Cette approche permet l'affectation automatique de documents dans des classes préexistantes. L'objectif est de trouver une liaison fonctionnelle, que l'on appelle également modèle de prédiction, entre les textes à classer et l'ensemble des catégories. Pour estimer le modèle de prédiction, il faut disposer d'un ensemble de textes préalablement étiquetés, dit ensemble d'apprentissage, à partir duquel on estime les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire qui produit le moins d'erreurs en prédiction.

A la différence de la classification non supervisée où l'ordinateur doit découvrir lui-même des groupes de documents, la classification supervisée suppose qu'il existe déjà une classification de documents. C'est le cas par exemple d'une bibliothèque ou d'un moteur de recherche. Le but est alors de classer automatiquement un nouveau document. Il s'agit donc d'apprendre d'abord un modèle, ou classifier, à partir d'un ensemble d'entraînement composé de couples (objet, classe).

Contrairement à la classification non supervisée, la classification supervisée peut mesurer l'importance de chaque mot pour classer de nouveaux documents. Par exemple, une mesure (gain d'information) calcule la typicité d'un terme. Plus un mot est lié à une catégorie et pas aux autres, et plus il est important : si un nouveau document le contient, ce mot sera très discriminant. De nombreuses mesures semblables ont été mises au point.

Enfin, à l'inverse de la classification non supervisée, il est ici simple d'évaluer les résultats d'une classification. Parmi les N exemples de documents classés, on utilise une partie des documents pour l'entraînement, et le reste pour le test. Pendant la phase de test, on soumet chaque document à l'algorithme de classification et on regarde simplement si la machine trouve la bonne classe. Bien sûr, le résultat de ce test n'est en rien garanti lorsque la

machine aura à classer de nouveaux documents ! (Réussir le test est nécessaire, sans être suffisant) [9].

1.9 Les applications de la catégorisation des textes

La catégorisation de textes peut être un support pour différentes applications parmi lesquelles :

- L'identification de la langue.
- La reconnaissance d'écrivains et la catégorisation de documents multimédia.
- L'étiquetage de documents,
- Le filtrage (consistant à déterminer si un document est pertinent ou non (décision binaire)).
- Le routage (consistant à affecter un document à une ou plusieurs catégories parmi n [10] .

1.10 Démarche à suivre pour la catégorisation de textes

Pour réaliser l'opération de catégorisation automatique de textes comme nous l'avons défini, la démarche commune est la suivante : la première phase consiste donc 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).

Pour améliorer la performance des modèles, une évaluation de la qualité des classifieurs et la comparaison des résultats fournis par les différents modèles est effectuée en fin de cycle. La démarche d'une approche standard de classification 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 classification à partir des couples (Document, Classe).
- Évaluer les résultats du classifieur

La figure 1.2 illustre la démarche de catégorisation de textes avec ses trois étapes qui peuvent être schématisées comme suit [7] :

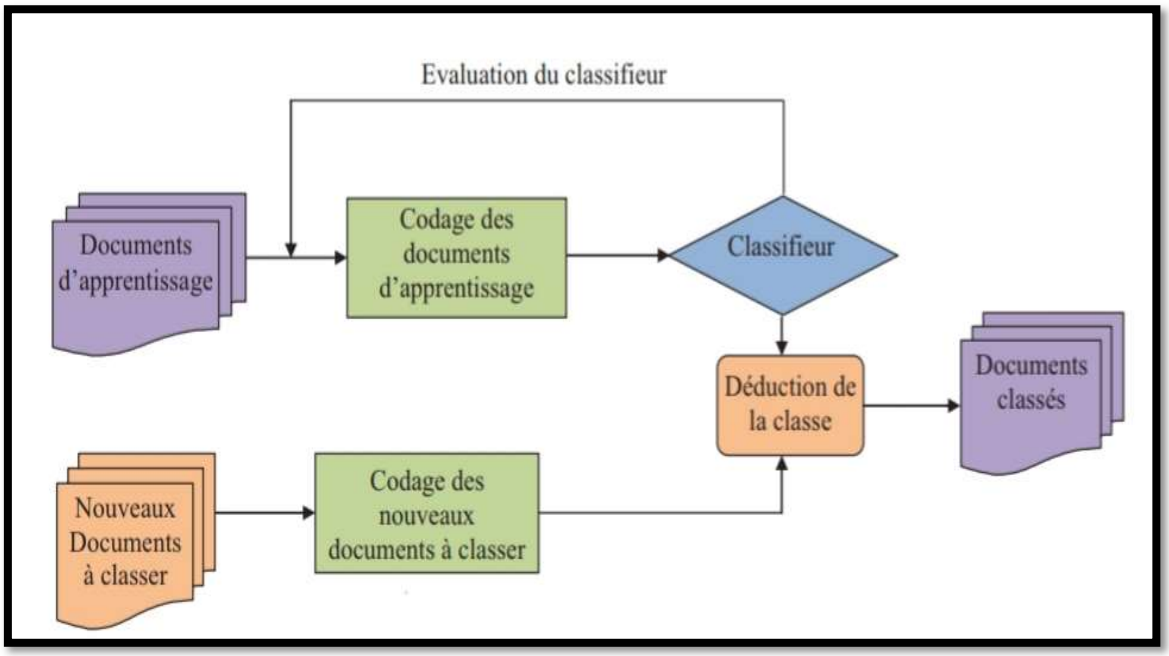


Figure-1.1: démarche de la catégorisation de textes [7]

1.11 Plagiat

1.11.1 Définition de plagiat

Le plagiat est une pratique que l'on rencontre dans tous les champs de l'activité humaine où s'exerce la création : la litté- rature, la peinture, la **musique**, la mode, etc. Il s'apparente à la contrefaçon et se caractérise par la reproduction d'éléments essentiels et caractéristiques d'une création, avec souvent une intention de tromperie dans un but intéressant. L'objectif est soit de faire croire que la contrefaçon est produite par un créateur connu dont le faussaire usurpe l'identité (signer Vermeer un tableau qui est une imitation), soit à l'inverse de s'attribuer une œuvre créée par un autre sans mentionner la source (signer de son nom une musique copiée d'Éric Satie). La science est une autre activité de création pour laquelle le plagiat s'apparente à ce dernier cas [11].

1.11.2 **Types** de plagiat

Nous formulerons une définition générale du plagiat que nous compléterons par la présentation de neuf types de plagiat qui s'échelonnent sur un continuum de gravité, allant du plagiat conscient au plagiat inconscient. Lorsque cela sera jugé nécessaire, des distinctions supplémentaires seront apportées pour préciser des conduites apparentées au plagiat sans en être réellement [11] .

1.11.2.1 Plagiat classique

Le plagiat est une question de faits: pour être protégeable, l'« œuvre ne peut pas être une simple idée. Il faut qu'elle soit une forme tangible d'expression ». De plus, l'œuvre plagiée doit être « originale» au sens où elle tire son origine du fruit du travail de l'auteur plagié [11].

1.11.2.2 Paraphrase abusive

La distinction entre une paraphrase acceptable et une paraphrase abusive n'est pas facile. Dubois suggère de définir la paraphrase acceptable comme « l'énoncé d'une idée dans des mots différents mais de même longueur». Dans la mesure où il subsiste un écart acceptable entre les deux textes, l'auteur est alors dispensé des guillemets, mais non de la référence. La notion d'écart est une affaire de jugement [11].

1.11.2.3 Vol de paternité

Le vol de paternité regroupe un ensemble de conduites voisines du plagiat classique, mais plus difficilement identifiable. Il se produit soit entre collègues, soit entre maîtres et élèves [11].

1.11.2.4 Plagiat par omission des références secondaires

La notion de référence secondaire se rapporte à deux situations différentes. La première est acceptée et prévue par la communauté scientifique. Elle consiste à citer un auteur en se basant uniquement sur ce qu'un autre auteur en dit. Seul le deuxième auteur apparaît alors dans la liste des références puisque c'est le seul qui a été consulté explicitement [11].

1.11.2.5 Plagiat par omission de citation

Le plagiat par omission de citation constitue en quelque sorte le pendant du plagiat par omission des références secondaires. Ici, l'auteur passe sous silence plus ou moins délibérément la provenance réelle des idées des autres qui meublent son texte. La faute ne vient pas de l'utilisation de ces idées mais de la non-reconnaissance de leur provenance qui laisse alors l'impression qu'il en est bien l'unique auteur **[11]**.

1.11.2.6 Auto plagiat

Le terme auto plagiat recouvre un ensemble de pratique, les unes largement répandues et acceptées, voire souhaitables les autres réprouvées et à bannir. En fait, on peut établir trois

catégories s'échelonnant sur un continuum de gravité. A l'une des extrémités, on retrouve l'autoplagiat intentionnel inacceptable, au centre une large zone grise plus ou moins bien définie et à l'autre extrême, l'autoplagiat correct [11].

1.11.2.7 Plagiat oral

Le plagiat ne s'applique pas seulement aux écrits. Les principes de la reconnaissance de la paternité des écrits et des idées concernent aussi les exposés oraux incluant discours, conférences et enseignements [11].

1.11.2.8 Plagiat inconscient

Le mode de fonctionnement de la science implique que les chercheurs sont nécessairement en contact avec plusieurs idées plus ou moins reliées à leur champ d'intérêt. Une idée peut dès lors jaillir sans que le chercheur soit conscient qu'il l'a déjà lue dans une revue scientifique ou entendue lors d'un congrès ou d'une conversation informelle [11].

1.11.2.9 Plagiat traduction

La traduction d'un texte constitue une situation idéale pour le paraphrase et une tentation supplémentaire pour un éventuel plagiaire. L'auteur tente alors de faire passer le texte traduit pour un texte de son cru. La détection de ce type de plagiat est particulièrement difficile surtout lorsque le plagiat se content de traduire de courts extraits difficilement repérables dans un texte. Par contre, celui qui tenterait de s'attribuer la paternité de la traduction d'une œuvre entière serait plus facilement repérable. À cet égard, il y a quelques cas célèbres. Par exemple, au XIXe siècle (1862), Bertholow a publié dans le Journal of Médicine un article qui fut reconnu comme une traduction littérale d'un essai du français Topinard [11].

1.12 Conclusion

Dans ce chapitre, nous avons exposé quelques généralités sur l'attribution d'auteur(AA) bref historique on a montré traits d'un auteur et les étapes de l'attribution d'auteur. Par la suite, des définitions de la stylométrie ainsi qu'un bref historique de cette dernière sont présentées. D'autre part, nous avons cité la catégorisation automatique des textes finalement nous avons discuté la définition et quelques types du plagiat.

Dans le prochain chapitre, nous allons présenter TA et la méthodologie de recherche ainsi que les techniques utilisées dans ce travail de recherche.

CHAPITRE-2

TRADUCTION AUTOMATIQUE ET METHODES PROPOSEES

CHAPITRE-2
TRADUCTION AUTOMATIQUE ET METHODES PROPOSEES

2.1 Introduction

Dans ce chapitre on va découvrir comment fonctionne ce que l'on appelle la Traduction Automatique (TA). Nous verrons dans ce qui suit quels sont les enjeux et défis de la TA en commençant par son historique, puis les différents systèmes de la TA et enfin les approches (méthodes) proposées pour l'AA.

2.2 Historique

Dans les années cinquante, la compétition entre les États-Unis et l'Union Soviétique est à son comble. Les autorités américaines sont inquiètes des progrès techniques de l'autre grande puissance. Les Américains pensent alors que si les milliers de pages de documents publiés par les Russes pouvaient être traduits en anglais, alors ils pourraient prévoir leurs avancées techniques. Les domaines des technologies de l'espace, de la physique atomique, et d'une manière générale de la défense les intéressent au premier chef. Mais traduire beaucoup et vite dans des domaines extrêmement spécialisés ne pourrait guère se faire de façon traditionnelle. Ainsi est née l'idée de la « machine à traduire »

Dans la même période se développent des « machines à calculer » qui n'ont plus rien à voir avec les calculatrices de Pascal. L'électronique entre dans les transmissions : la télévision, d'expérimentale devient quasi quotidienne dans la décennie. Elle est aussi utilisée pour les grands calculs dans des centres spécialisés. L'idée d'associer électronique (plus tard informatique) et traitement du langage (en particulier traduction) est née dès cette époque.

Dans un contexte de course à l'espace, la traduction rapide des documents soviétiques devient un enjeu non seulement scientifique mais politique. Le premier Spoutnik est mis en orbite le 4 octobre 1957. Le 12 avril 1961 Gagarine fait le tour de la terre à bord du vaisseau Vostok. Le 20 avril 1961 le président Kennedy (arrivé à la Maison Blanche en janvier 1961) demande au vice-président Johnson de faire le point sur l'astronautique Américaine. Le 30 mai 1966 la première sonde américaine est posée sur la Lune, le 21 décembre 1968 le premier vol habité américain fait le tour de la Lune (Apollo 8), et le 21 juillet 1969 le premier Américain laisse les traces de ses bottes sur le sol lunaire. Quel a été le rôle de la traduction dans ces événements ? On ne le sait. Mais ce qui

est évident, c'est que l'atmosphère de cette décennie est propice aux projets ambitieux, même s'ils paraissent irréalistes à certains.

Ainsi, aux États-Unis, le gouvernement et les industries sont prêts à investir dans la recherche en vue de la traduction automatique. L'université de Georgetown à Washington D.C. se lance la première dans cette entreprise. Et le travail est d'abord confié à des linguistes qui essayent de coder la connaissance linguistique. Les ingénieurs se joignent ensuite à eux. Ces efforts mènent aux systèmes Mark 1 et Mark 2 d'IBM. Ceux-ci ont été utilisés par l'US Air Force. Ils conduisent aussi au « système Georgetown » qui a été utilisé par la Commission de l'Énergie Atomique aux USA, et plus tard par le Centre de Recherche Atomique d'Ispra. Ces deux systèmes étaient, bien entendu, des systèmes de traduction russe-anglais. Ils ont fonctionné jusque vers la fin des années soixante.

Pendant cette même décennie, Peter Toma, ingénieur qui avait quitté la Hongrie après la seconde guerre mondiale et avait travaillé un temps comme officier de liaison entre la 3e armée en Bavière et la Croix Rouge hongroise, comprend rapidement à quel point il était important d'améliorer les communications entre Russes et Anglais. Il décide, comme premier pas, de maîtriser lui-même le russe. En 1956 il s'installe en Californie, et cherche à appliquer ses connaissances pratiques de langues à la technologie des ordinateurs. Son but est de produire un système de traduction automatique pragmatique. Contrairement aux autres scientifiques de l'époque, Toma ne croit pas que la linguistique puisse fournir une solution adaptée au traitement du langage par l'ordinateur. Il est convaincu que le traitement du langage doit être adapté aux possibilités de l'ordinateur plutôt que l'inverse.

C'est donc au début des années soixante qu'il commence à appliquer ses hypothèses aux ordinateurs au Californian Institute of Technology, à Pasadena. Il y a d'abord produit *AUTOTRAN,* ensuite *TECHNOTRAN.,* et finalement, *SYSTRAN,* sur l'IBM 360, en 1963-64.

Cependant, en 1967 le rapport ALPAC (Automatic Language Processing Advisory Committee*)* recommande une nette réduction des travaux de recherche en vue de la traduction automatique aux États-Unis. Peter Toma estime que l'aspect pragmatique des choses n'était pas du goût des théoriciens savants qui ont eu à expertiser les progrès de

cette recherche. Bien que ce rapport ait signifié une coupure totale des budgets américains dans ce domaine, Toma réussit à convaincre la Deutsche Forschungs gemeinschaft à Bonn et la même année, il a un prototype de système de traduction automatique fonctionnant sur IBM 360-30. La dernière main y a été mise à l'Université de Bonn sur un IBM 3460-50 entre septembre et novembre 1968.

En 1974, Toma applique les résultats du programme d'analyse de l'anglais à un prototype anglais – français. Celui-ci est testé par l'Office des traductions fédéral canadien, et aussi par les services centraux canadiens de Ford et General Motors. General Motors continue à donner son appui au développement du système [12].

2.3 Les différents systèmes de TA

Le travail de nombreuses équipes aboutit à une diversification des méthodes. Si la traduction directe – d'abord conçue comme un mot-à- mot à peine amélioré, puis intégrant une part variable d'analyse syntaxique- domine largement, s'ajoutent deux autres approches. D'où la typologie suivante, qui perdurera

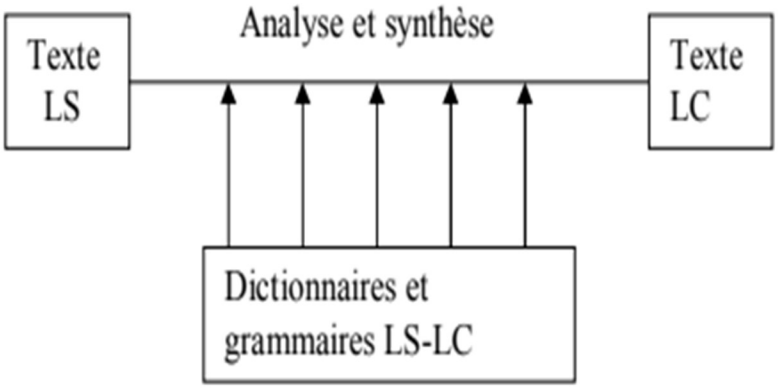


Figure-2.1 : Système de traduction direct

Un système direct n'opère que les analyses indispensables pour la traduction d'une langue dans une autre : l'analyse de la syntaxe et du vocabulaire de la langue source n'est pas poussée au-delà de ce qui est indispensable pour choisir les expressions appropriées et l'ordre des mots de la langue (figure2-2).

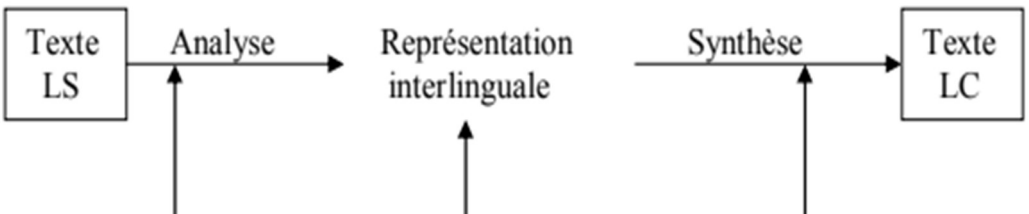


Figure2 -2: Système interlingual

La base conceptuelle de ce travail réside dans la conviction que la traduction interlinguale (textuelle), intralinguale (métatextuelle), intersémiotique (extratextuelle), intertextuelle et intratextuelle peut être décrite sur la base d'un seul modèle de processus de traduction, et qu'en définitive, il faille des présupposés pour décrire la culture comme processus global de traduction totale.

Littérature en traduction. Connotations culturelles dans les traductions interlinguales (polonaise, française et anglaise) des « Chroniques de l'oiseau à ressort » interprétation-accourt, construction, explication, interprétation - adaptation - adaptation, version - adaptation, interprétation, transposition, version - crib, interlingual reddition, rendering, translation, version - traduction - translating, translation - thème - prose - sous-titrage - subtitles, subtitling - romanisation – transposition C'est le code qui permet la traduction, avec chaque composante qu'elle comporte : interlinguale, intralinguale et intersémiotique.

Nous ne devrions cependant pas nous hâter de considérer la traduction intralinguale comme le « fondement » des deux autres genres de traduction cités par Jakobson, à savoir la traduction interlinguale et la traduction intersémiotique (c'est-à-dire entre des systèmes

de signes différents, par exemple un système de signes verbaux et un système de signes iconiques), comme si toute traduction n'était que la simple actualisation, sinon de significations données d'avance, du moins d'un système de signes existant.

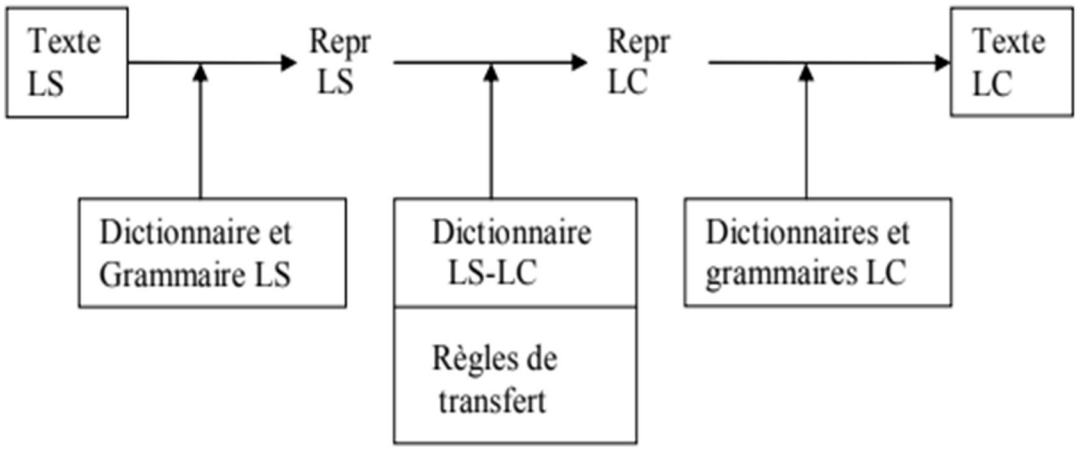


Figure-2.3 : Système par transfert

On appelle système par transfert un système qui comporte trois étapes, la première transforme le texte source en représentations de transfert de la langue source, la seconde transforme celles-ci en représentation de transfert de la langue cible, à partir desquelles est opérée la synthèse du texte en langue cible**.[13]**

2.4 Approches proposées pour l'attribution d'auteurs

On emploie le terme de « traduction » aussi bien pour la poésie que pour les romans, la publicité, les ouvrages scientifiques, les rapports et manuels techniques, les nomenclatures de pièces détachées, alors qu'il conviendrait, au moins, de distinguer : - la « recréation », par exemple la traduction d'Edgar Allan Poe par Baudelaire, qui vise avant tout à transmettre l'aspect subjectif, fût-ce au prix d'une légère transformation du contenu; - la « localisation », largement pratiquée pour les manuels de micro-ordinateurs, qui vise à adapter un contenu à un environnement culturel particulier; - la « traduction-diffusion » , en particulier la traduction de documentations techniques dont le contenu doit être strictement rendu, sans ajout ni omission, même si le style « sent la traduction »; et

la « traduction rapide » enfin, dans laquelle nous rangerons la « traduction dépistage » de textes écrits et l'interprétation simultanée. Quelle automatisation de la traduction peut-on souhaiter et réaliser sur des stations de travail individuelles dans les dix prochaines années? Après avoir examiné les types d'automatisation envisageables, et les situations traductionnelles, qui ne se limitent pas à l'image d'Épinal d'une « photocopieuse-

traductrice», nous conclurons qu'il vaut mieux en général ne pas parler de « stations de travail », mais plutôt d'environnements ou d'outils, qui ne pourront être spécifiques que dans un cas, celui de la traduction professionnelle en groupe, avec ou dans traduction automatique**.**

2.4.1 La TAO

actuelle II n'est absolument pas envisageable pour l'instant d'automatiser la traduction récréation ni la traduction-localisation plus que par la mise à disposition d'outils d'aide au traducteur humain. Par contre, la « fonction traduisante » est automatisable pour la traduction-diffusion et la traduction-dépistage. Le terme «TAO» (Traduction assistée par ordinateur) recouvre aujourd'hui l'ensemble des techniques d'automatisation de la traduction. L'ancien terme « TA » (Traduction automatique) est réservé aux techniques d'automatisation de la fonction traduisante, qu'il y ait ou non préédition, postédition ou interaction, tandis que celui de « THAM » (Traduction humaine assistée par la machine ») concerne les outils ou environnements d'aide au traducteur ou au réviseur. Avant de se demander quelle pourrait être la « station de travail du traducteur » en l'an 2001, il n'est pas inutile de voir où nous en sommes aujourd'hui.

- **TA pour le dépistage**

Vers 1949, les États-Unis, puis l'URSS, ont lancé des programmes de « TA » motivés par le besoin de renseignements. C'est ce que nous appellerons la TAO pour le veilleur. Il s'agit de traduction automatique, dont on attend des traductions grossières, produites rapidement, en grand volume et à bas coût. La qualité de ces traductions n'est pas essentielle. Elles servent en effet à filtrer des documents, dont les plus intéressants seront, si nécessaire, traduits ou communiqués à des spécialistes bilingues. Préédition et postédition doivent être absentes ou très limitées (ex: séparer les phrases, les formules, Les systèmes SYSTRAN sont essentiellement de ce type (par exemple, le système russe-

Anglais installé depuis 20 ans à la Wright-Patterson Air Force Base traduit, d'après nos informations, environ 18 millions de mots par an, avec une qualité tout à fait satisfaisante pour l'usage visé). Ce besoin est toujours actuel. Cependant, il s'agit maintenant plus de veille scientifique, technique, économique et financière que de renseignement militaire. À titre d'exemple, on peut citer l'accès en anglais à des bases de données japonaises depuis la Suède.

- **TA pour la diffusion**

Une quinzaine d'années plus tard, on a commencé à travailler sur la TAO pour le réviseur. Il s'agit de produire automatiquement des traductions brutes, destinées à être révisées. Dans cette optique, la machine doit remplacer le traducteur, qui est promu réviseur. Cela n'est possible que si l'on restreint convenablement le style et le domaine Christian Boitet des textes à traduire (approche par « sous-optimisation », pour reprendre le terme de L. Bourbeau). Les décideurs (politiques, scientifiques et industriels) comme le grand public n'envisagent souvent que cette possibilité, et ce sans doute à tort. En effet, il existe des systèmes qui peuvent répondre à des besoins de ce type, mais seulement dans des situations convenables. Sinon, l'échec est garanti. Voyons cela un peu plus en détail. Il existe actuellement près d'une quinzaine de systèmes. Il y a surtout des systèmes japonais (AS-Transac de Toshiba, ATLAS-II de Fujitsu, PIVOT de NEC, HICAT de Hitachi, SHALT d'IBM-Japon, PENSÉE d'OKI, MU de l'Université de Kyoto et du JICST...), qui traitent presque uniquement les couples japonais anglais. On peut encore citer des systèmes américains (LOGOS, METAL) ou français (Ariane/ aéro/F-E de SITE-B'VITAL, fondé sur les outils informatiques et les méthodes linguistiques du GETA) centrés sur l'anglais, l'allemand ou le français, bien que des maquettes ou des prototypes existent sur de nombreuses autres langues. Que peut-on espérer de tels systèmes? Essentiellement, de répondre à des besoins de plus en plus importants en traduction technique. Typiquement, en moyenne industrielle, une page de 250 mots est traduite en une heure et révisée en 20 minutes. Dans l'idéal, avec quatre personnes, on passerait donc de trois pages à l'heure à douze pages à l'heure, et on multiplierait donc la productivité par quatre. Il s'agit en fait d'une limite, le chiffre le plus raisonnable étant plutôt de huit pages à l'heure, en comptant une révision plus lourde, de 30 minutes par page, et ce avec des réviseurs formés. Quand les utiliser? Cela n'est actuellement envisageable que pour

de gros flux de textes homogènes et informatisés, comme des manuels d'utilisation ou de maintenance. Dans ces conditions, un système à 1 MF (400 KF de base et 600 KF de spécialisation au vocabulaire et au type de texte) doit pouvoir être amorti en deux ans, pour un flux de 10 000 pages par année (en comptant 10 %/an de maintenance, 60 F/page de coût machine, et 100 F/page de révision, contre 150 F/page de traduction et 70 F/page de révision pour la méthode manuelle classique, soit 60 F/page de gain pour amortir 1,2 MF). Comment les utiliser? Une condition essentielle de succès de ce type de TAO est de

constituer une équipe de développement et de maintenance des linguiciels (dictionnaires, grammaires) qui soit en liaison constante avec l'équipe de révision, et si possible avec les auteurs des documents à traduire. C'est ce qu'a su réussir la PAHO (Pan American Health Organization), avec ses systèmes ENGSPAN et SPANAM. Dans le « contre-rapport ALP AC » du JEIDA comme au MTS-II à Munich en août 1989 par exemple, Fujitsu a clairement reconnu avoir fait une erreur en distribuant largement ATLAS-II : seules sont en effet rentables les traductions effectuées chez Fujitsu, soit pour sa documentation, soit dans le cadre d'un contrat avec la CEE, ce dernier ne demandant qu'une révision minimale, car il s'agit en fait de veille technologique. Peut-être est-il approprié de faire ici un parallèle avec les systèmes experts, qui peuvent être développés par des tierces parties, mais qui doivent ensuite être totalement maîtrisés par leurs utilisateurs, seuls à même de les faire évoluer de façon adéquate**.[14]**

2.5 Traduction automatique (T.A) et intelligence artificielle (I.A)

La Traduction automatique (TA) est un logiciel capable de traduire des mots, phrases ou textes d'une langue source (LS) à une langue cible (LC) à l'aide de l'Intelligence Artificielle (IA). Aujourd'hui, on peut utiliser la TA sur des sites gratuits, du moment où on a accès à internet. On peut aussi se servir de logiciels de Traduction Assistée par Ordinateur (TAO), qui eux utilisent des systèmes de mémoire sans IA. Le succès de cet outil, encore à ses balbutiements, n'a pourtant pas été évident lorsqu'il a été créé par Georges Artsrouni.

Il est l'inventeur des trois premières machines de traduction créées entre 1932 et 1937. Ces appareils étaient mécaniques et fonctionnaient grâce à une « mémoire » entièrement dispensée par l'homme sur des bobines de carton sur lesquelles il notait les mots dans les

langues cibles choisies. Ces dispositifs sont le point de départ des recherches d'A.D. Booth qui veut les perfectionner en y ajoutant le principe de la calculatrice numérique. Son

initiative se concrétise en 1954 avec l'expérience Georgetown-IBM lorsque leur machine traduit plus de quarante phrases de l'anglais au russe. Cette réussite encourage l'engagement de grands investisseurs, comme le gouvernement américain. L'US Air Force

et la NASA s'allient avec la première entreprise spécialisée dans la traduction automatique, SYSTRAN, fondée en 1968. Leur objectif est de dépasser les frontières linguistiques dans le contexte de la Guerre en créant un logiciel de traduction du russe vers l'anglais. Un

intérêt d'abord politique bien qu'il s'élargisse à des domaines plus pratiques comme celui de la météo avec le projet canadien TAUM-Météo. D'autres sociétés se sont établies depuis et des pays s'engagent dans ces recherches, comme la France depuis 1959 avec l'aide du CNRS, dans l'espoir de diversifier les utilisations de cette technique**.[15]**

2.6 Méthodologie de recherche proposée

Notre méthodologie de recherche est basée sur quatre étapes. La première étape consiste à convertir les textes scannés en textes modifiables à l'aide d'un système OCR. La deuxième étape est consacrée aux corrections des textes obtenus afin de les préparés pour l'utilisation dans l'attribution d'auteurs.

Dans la troisième étape on fait l'extraction des caractéristiques pertinentes (dans notre cas les caractères n-grammes) et construction du modèle de chaque auteur. Enfin, la quatrième étape est dédiée aux méthodes de classification (MLP et SVM et LDA) pour réaliser le processus d'identification.

2.6.1 Conversion des textes scannés

Les textes scannés sont convertis en textes modifiables en utilisant un système de Reconnaissance Optique de Caractères (en anglais : Optical Recognition Character)

(OCR). 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**.[16]**

Figure-2.4 : Conversion des textes scannés en textes modifiables à l'aide d'un OCR.

La technologie d'OCR a été appliquée ces dernières années à travers tout le spectre d'industries en train de révolutionner le processus de gestion des documents. Les systèmes d'OCR ont permis à des documents numérisés de se transformer en documents entièrement consultables avec le contenu du texte qui est reconnu par les ordinateurs. Cependant, après plus de deux décennies de recherche sur la numérisation des documents, ces systèmes peuvent encore laisser quelques imperfections pour parvenir à une réédition du document ce qu'il peut être dû aux différents problèmes dont la qualité du document et de l'impression, la discrimination de la forme, le type d'acquisition, les variations des dimensions, le nombre de scripteurs, la taille du vocabulaire, etc.

2.6.2 Prétraitement des textes obtenus par la conversion OCR

Les textes convertis à l'aide d'un OCR comportent deux types d'erreurs ; caractères insignifiants ou (bruits) (caractères spéciaux, des chiffres, etc.) et caractères incorrects. Les caractères insignifiants sont des caractères qui n'ont pas un sens bien défini dans la langue arabe et qui apparaissent par erreur dans les textes convertis à l'aide d'un système OCR (i.e. ", %, &, £, *, #, $, 0, 1, …, 9, etc.). Or, les caractères incorrects sont des caractères qui sont mal convertis ou convertis par erreur en autres caractères que les vrais caractères En conséquence, le prétraitement appliqué .Dans cette phase consiste à :

- Supprimer les caractères insignifiants,
- Supprimer les diacritiques arabes,
- Supprimer les multiples espaces de mots.
- Supprimer les signes

Pour les caractères incorrects sont laissés afin de tester la robustesse de notre méthode proposée pour AA

N°	Caractères	Noms
1	a b c … y z	Lettre français minuscules
2	A B C … Y Z	Lettre français majuscules
3	0 1 2 3 4 5 6 7 8 9	Chiffres en Français
4	٠١٢٣٤٥٦٧٨٩	Chiffres en Arabe
5	‘ ‘ "" « » “ ”	Guillemets
6	* # @	Etoile, dièse, arobas

Tableau-2.1. Liste des caractères insignifiants pour la stylométrie

2.6.3 La traducteur automatique des textes scannés

Dans cette partie, les textes scannés par OCR on les traduits automatiquement en deux langues français et anglais

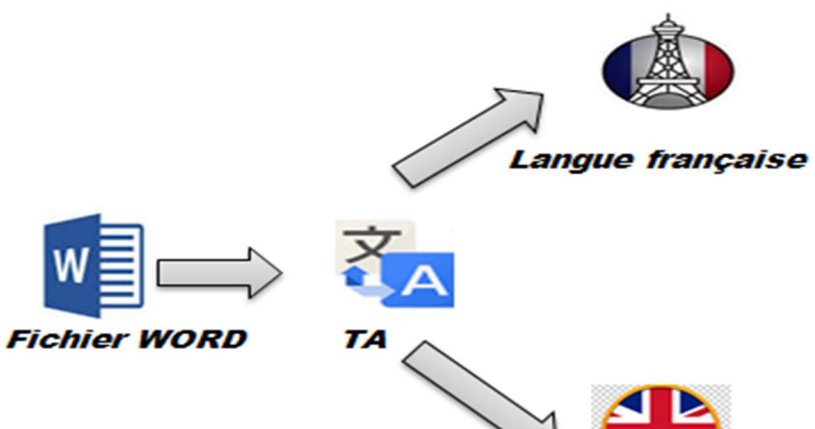


Figure-2.5 : Convertir des textes d'une langue à une autre à l'aide d'un TA

2.6.4 Extraction des caractéristiques

La notion de N-grammes de caractères a été utilisée de manière fréquente dans l'identification de la langue ou dans l'analyse de corpus oraux. L'utilisation de profils de fréquence N-gramme, qui est une tranche de N caractères d'une chaîne de caractères, est un moyen simple et fiable de classification des documents dans un large éventail de tâches de catégorisation.

Les N-grammes sont tolérants aux fautes d'orthographe et aux déformations causées lors de la reconnaissance de documents (système OCR). Lorsqu'un document est reconnu à l'aide du système OCR il y a souvent une part non négligeable de bruit. Par exemple, il est possible que le mot "feuille" soit lu "teuille". Un système fondé sur les N-grammes prendra en compte les autres N-grammes comme "eui", "uil", etc. Dans quelques travaux les N-grammes de caractères sont appliqués pour la classification de petits documents tels que les polluriels (SPAM), courriers électroniques, SMS. D'autres travaux

utilisent les N-grammes pour la classification de langues complexes. Les expérimentations fondées sur diverses valeurs de N (de 2 à 7-grammes),

Les types de caractéristiques qui ont été proposées et utilisées dans ce travail sont N grammes (avec N= 2, 3, 4,5 ,6,7) comme illustré ci-dessous:

- Caractères bi-grammes (n=2),

- Caractères tri-grammes (n=3),
- Caractères tétragrammes (n=4),
- Caractères pentagrammes (n=5).

Pour utiliser ces caractéristiques, une liste de tous les mots est extraite du texte, puis les caractères n-grammes de chaque mot sont pris, ainsi un profil de caractères n-grammes est créé (contenant les caractères n-grammes et leurs fréquences).

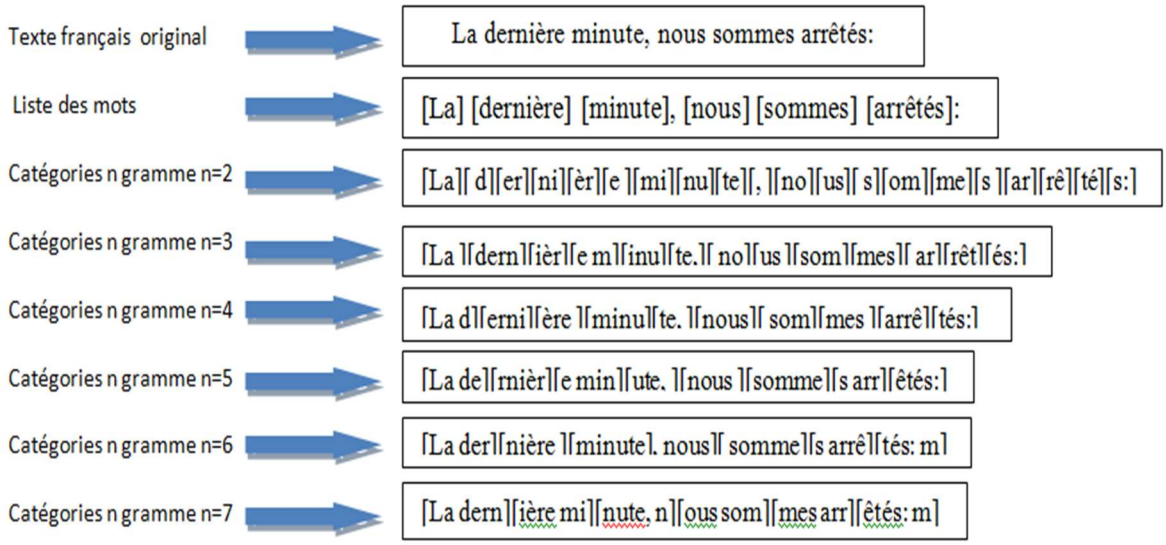


Figure-2.6 : Exemple d'extraction des caractères N-grammes d'un texte.

2.6.5 Approches proposées pour l'attribution d'auteurs

Les réseaux de neurones artificiels (RNA) constituent une nouvelle méthode d'approximation de systèmes complexes, particulièrement utile lorsque ces systèmes sont difficiles à modéliser à l'aide des méthodes statistiques classiques. Une classe de modèles

de RNA appelée perceptrons multicouches (PMC) a, ces dernières années, été privilégiée pour la prévision de phénomènes hydrologiques. La théorie et le langage connexionniste restent, malgré cette percée, encore peu connus de la communauté des hydrologues. **[18]**

2.6.5.1 Les réseaux de neurones artificiels

Les RNA sont des modèles mathématiques non linéaires, de type 'boîte noire', capables de déterminer des relations entre données par la présentation (l'analyse) répétée d'exemples (à savoir, des couples constitués par une information d'entrée et une valeur de sortie que l'on voudrait approcher par le modèle). La modélisation à l'aide de RNA (appelée 'phase d'apprentissage') suppose l'adaptation des paramètres du réseau, afin de mettre en évidence les relations qui portent sur les exemples présentés. Les RNA sont constitués d'un ensemble d'éléments de calcul (neurones artificiels), organisés dans une structure spécifique (par exemple, celle présentée sur la figure 2-5), les paramètres du réseau (les poids) étant représentés par les valeurs associées aux connections de ces éléments de calcul. Un élément de calcul du RNA comporte une ou plusieurs entrées et une sortie. La valeur de sortie est obtenue par l'application d'une relation mathématique (fonction d'activation) sur la somme pondérée d'entrées.

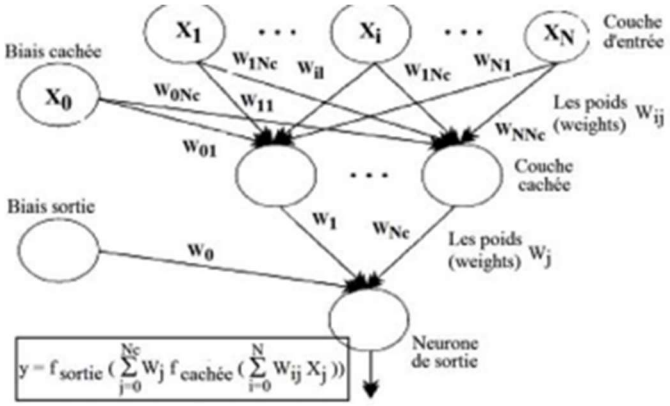


Figure-2.7: Structure du perceptron multicouche

Celui-ci comporte la couche d'entrée, une ou plusieurs couches intermédiaires (cachées) et la couche de sortie. Chaque couche contient des unités de calcul– neurones– connectés à

d'autres neurones par la voie des poids. Les flèches (connexions des éléments de calcul) indiquent le sens de propagation des données . Dans la modélisation à l'aide de réseaux de neurones artificiels, on peut choisir le type de fonctions d'activation, le nombre de neurones et l'arrangement de leurs connexions (à savoir, la structure du réseau). Généralement, on utilise des fonctions d'activation de type 'sigmoïde'**.[19]**

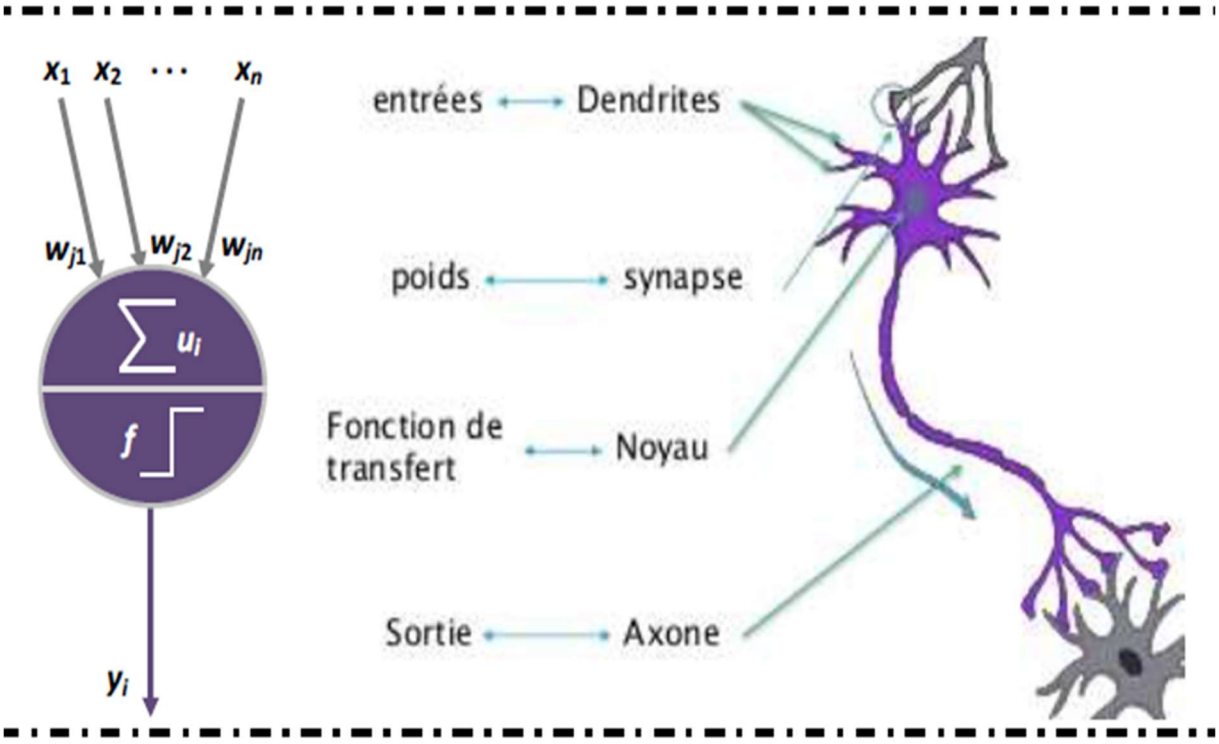


Figure-2.8 : Analogie entre neurone biologique et neurone artificiel

2.6.6 Phases d'apprentissage et de test du réseau

Le calage des paramètres du modèle (essentiellement le poids des liaisons entre les différents neurones) est réalisé d'après un algorithme de calcul qui utilise la présentation répétée d'un ensemble de plusieurs couples entrée – sortie connus (exemples qui constituent l'ensemble d'apprentissage). L'objectif de ce calcul est la minimisation d'une fonction d'erreur entre la réponse désirée et la réponse obtenue à la sortie du modèle.

Par exemple, l'algorithme de 'rétro propagation' estime le gradient de la fonction d'erreur par rapport aux poids du modèle et réalise l'adaptation de ces paramètres successivement de la couche de sortie vers la couche d'entrée. La validation du modèle se réalisera ensuite sur des exemples (ensemble de test) non utilisés dans le calcul des poids. La performance du réseau est déterminée en fonction du nombre de succès et d'échecs dans la discrimination. Les paramètres d'ajustement du réseau sont le nombre de neurones cachés et les fonctions d'activation. Ce travail d'apprentissage et de test est donc opéré sur un nombre important de configurations possibles, lesquelles sont classées en fonction de leur performance

2.6.7 Apprentissage des RNAs

2.6.7.1 Apprentissage supervisé (ou à partir d'exemples étiquetés) :

L'objectif de ce type d'apprentissage est de construire un modèle prédictif d'une grandeur numérique qui permet d'apprendre au mieux la fonction inconnue qui génère des données aléatoires, indépendantes et identiquement distribuées et dont nous ne disposons que de quelques exemples. Il s'agit, dans ce cas, d'un problème de classification. [20]

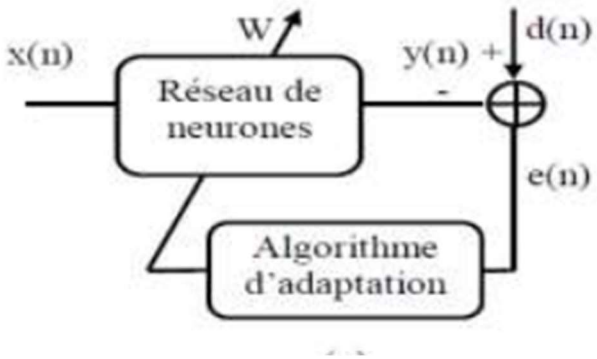


Figure-2.9: Apprentissage supervisé

2.6.7.2 Apprentissage non supervisé :

Il vise à apprendre certaines informations sur des données non étiquetées dans un but de les rassembler en des groupes homogènes. Dans ce cas, on présente une entrée au réseau et on le laisse évoluer librement jusqu'à ce qu'il se stabilise. La règle

d'apprentissage, n'étant pas fonction du comportement de la sortie du réseau, mais plutôt du comportement local des neurones. Il existe de nombreux cas où on ne possède aucune information sur les classes de l'ensemble d'apprentissage. Il s'agit, dans ce cas, d'un problème de clustering.

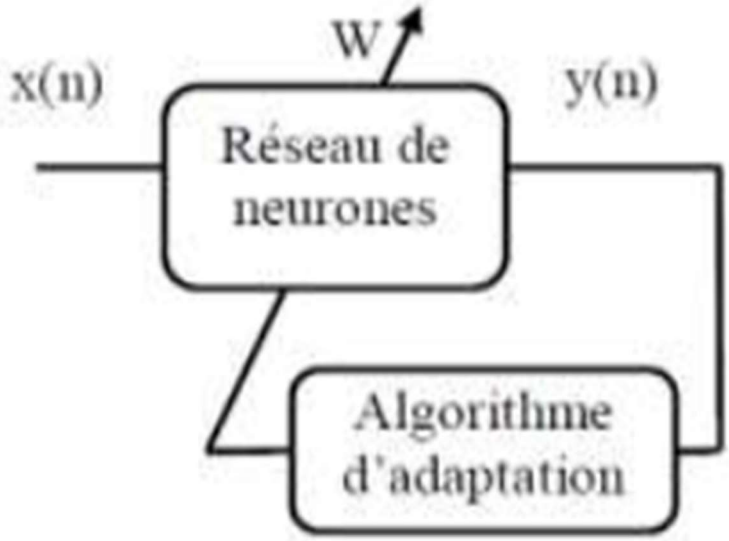


Figure -2.10 : Apprentissage non supervisé

2.6.7.3 Apprentissage par renforcement :

L'apprentissage par renforcement, c'est apprendre `a agir par essai et erreur. Dans ce paradigme, un agent peut percevoir sont états et effectuer des actions. Apres chaque action, une r´récompense numérique est donnée. Le but de l'agent est de maximiser la récompense totale qu'il reçoit au cours du temps. Une grande vari´et´e d'algorithmes ont proposes, qui sélectionnent les actions de façon `a explorer l'environnement et a graduellement construire une stratégie qui tend a obtenir une récompense cumulée maximale. Ces algorithmes ont appliques avec succès a des problèmes complexes, tels que

les jeux de plateau, l'ordonnancement de tâches , le contrôle descendeurs et, bien sur, des taches de contrôle moteur, simulées ou réelles. **[21]**.

2.7 Les méthodes Proposés

2.7.1 Multi Layer Perceptron MLP

Le Perceptron multicouche est un Classifier 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 feed forward .

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 ouches successives :

- Deux entrées (dans le cas de la couche 1), trois entrées sinon..
- Un système de poids liées à ces entrés sachant que le dernier poids)
- Une seule sortie (où sont présentées les sorties calculées par le MLP).

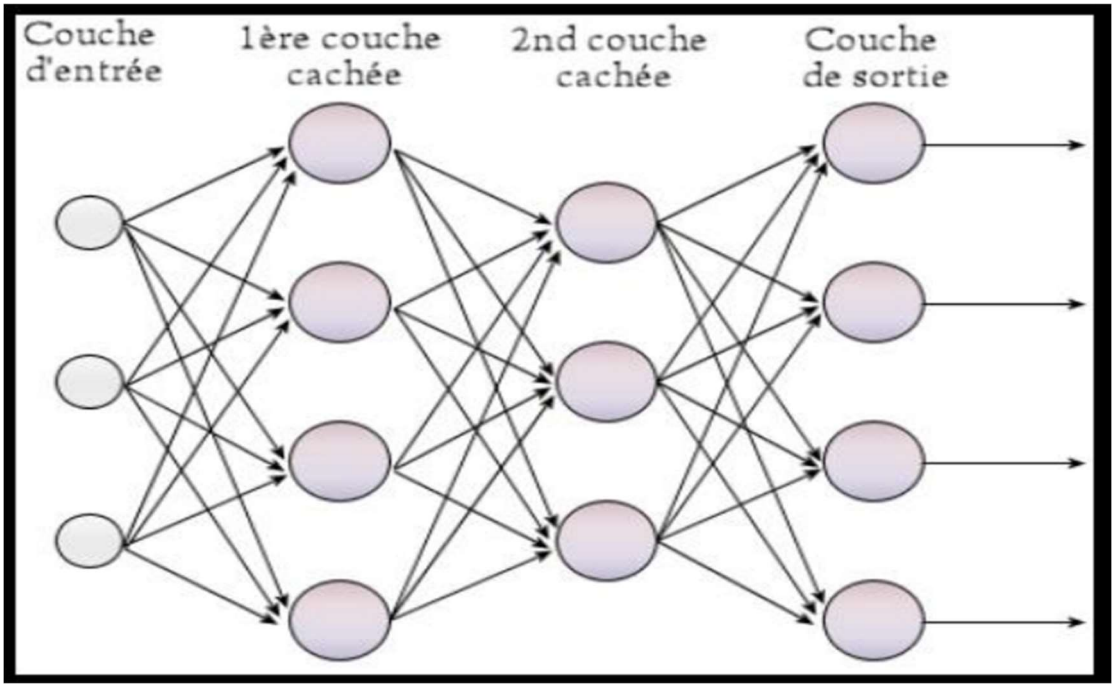


Figure -2.11 : Structure de Multi Layer Perceptron MLP.

ETAPE 1 : Initialisation au hasard ou aléatoire des poids des connexions et $\boldsymbol{vij}$et $\boldsymbol{wjk}$ •
ETAPE 2 : Propagation des entrées xi= ei

$Y_j = f(\sum_{i=1}^{m} x_i v_{ij} + x_0)$

Puis de la couche cachée vers la couche de sortie :

$KJ = f(\sum_{i=1}^{n}. Y_j W_{kj} + Y_0)$

Les valeurs 0 x et 0 y sont des biais, f est la fonction d'activation qu'on a choisie où on a défini notre réseau MLP.

• ETAPE 3 : Rétro-propagation de l'erreur pour chaque neurone de la couche de sortie, on calcule l'erreur, c'est-à-dire la différence entre la sortie désirée k s et la sortie réelle (obtenue) k z .

$E_k = Z_k(1 - Z_k)(S_K - Z_K)$

On propage cette erreur sur la couche cachée ; l'erreur de chaque neurone de la couche cachée est donnée par :

$F_j = YJ(1-YJ)\sum_{k=1}^{p} Wkj E_K$

ETAPE 4 : Correction des poids des connexions : Il reste maintenant la modification des poids des connexions et aussi les biais. **[22]**

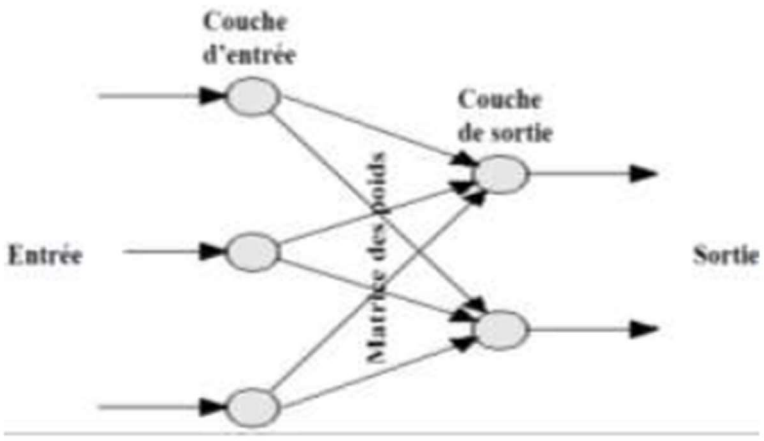


Figure- 2.12 : Architecture globale du perceptron (a) et du perceptron multicouche

Dans cette étude, nous allons utiliser un réseau de neurones de type perceptron multicouches "MLP" (Multi-layer Perceptron), car ce dernier est le plus utilisés en reconnaissance de formes. Le MLP est un réseau de neurones composé de plusieurs couches successives et chaque couche est connectée à la suivante. Il comporte en général une couche d'entrée, une couche de sortie et une ou plusieurs couches dites cachées.

La fonction d'activation des neurones de ce réseau est la fonction sigmoïde. Dans ce type de réseau, le superviseur fournit à l'entrée un ensemble de couples (entrée, sortie désirée) .

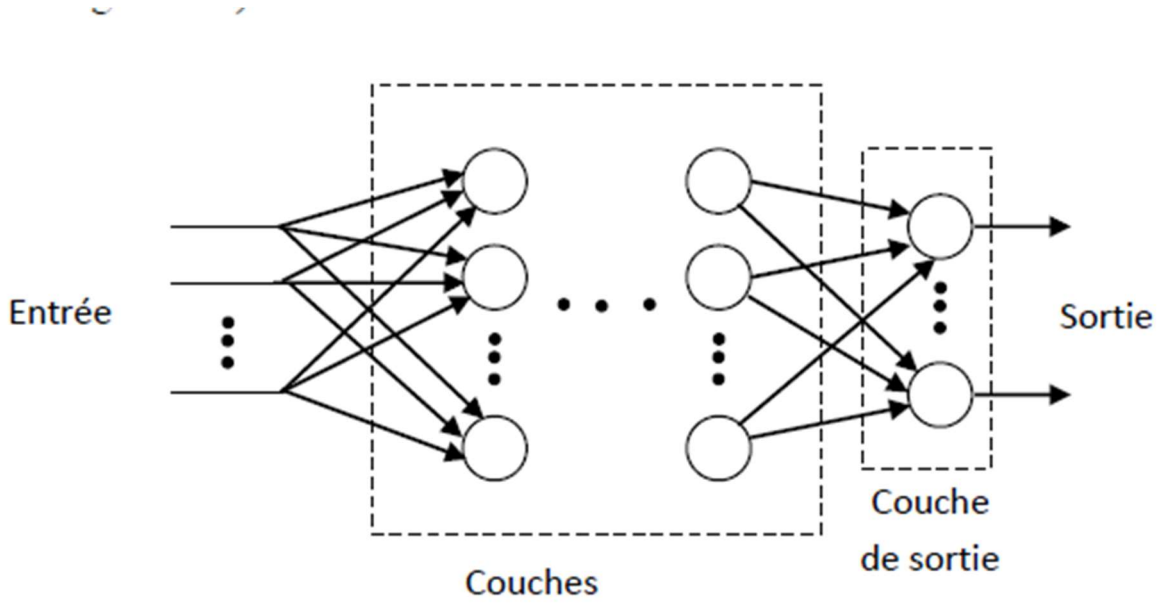


Figure-2.13 : Architecture d'un perceptron multicouche MLP.

Les neurones d'entrée sont organisées en une seule couche appelée couche d'entrée sans effectuer aucune opération sur ces signaux. Les neurones qui effectuent les calculs et les traitements intermédiaires sont les neurones des couches cachées. Le théorème des approximations universelles montre que la structure élémentaire à une seule couche cachée est bien suffisante pour prendre en compte les problèmes classiques de modélisation ou d'apprentissage statistique. Il existe plusieurs méthodes pour faire l'apprentissage des réseaux de neurones (MLP), parmi elles on peut citer :

- Algorithme de rétro-propagation du gradient.
- Algorithme de gradient conjugué.
- Méthodes de second ordre. **[23]**

2.7.2 Notions de base des SVMs

La reconnaissance de formes est un domaine fort intéressant de l'intelligence artificielle. Pour résoudre les problèmes de reconnaissance de formes, des classifies sont construits en utilisant des prototypes de données à reconnaître ainsi que leur classe d'appartenance. On parie d'apprentissage supervisé. Aujourd'hui, face aux importants volumes de données disponibles, le coût de l'étiquetage des données devient très exorbitant. Ainsi, il est impraticable, voir impossible d'étiqueter toutes les données disponibles. Mais puisque, nous savons que la performance d'un classifie est liée au nombre de données d'apprentissage, la principale question qui ressort est comment améliorer l'apprentissage d'un classifier en ajoutant des données non étiquetées à l'ensemble d'apprentissage. La technique d'apprentissage issue de la réponse à cette question est appelée apprentissage semi- supervisé.

La machine à vecteurs de support(SVM) et sa variante Least-Squares SVM (LS-SVM) sont des classifies particuliers basés sur le principe de la maximisation de la marge qui leur confère un fort pouvoir de généralisation. Au cours de nos travaux de recherche, nous avons considéré l'apprentissage semi-supervisé de ces machines. Dès lors, nous avons proposé diverses techniques d'apprentissage de ces machines pour accomplir cette tâche.

Dans un premier temps, nous avons utilisé l'inférence bayésienne pour estimer les paramètres du modèle et les étiquettes. Ainsi, nous avons élaboré des formulations bayésiennes à un et deux niveau(x) d'inférence, qui sont par la suite appliquées aux SVMs et aux LS-SVMs dans le contexte de l'apprentissage semi-supervisé

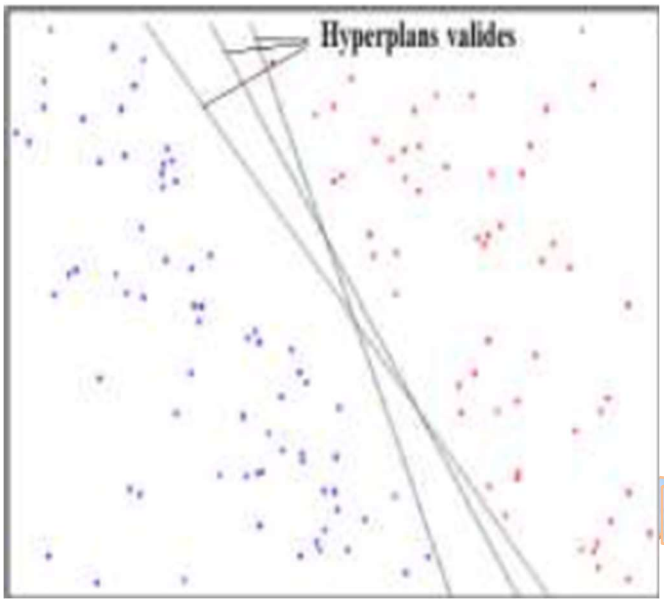
Dans un second temps, nous avons proposé d'améliorer la technique d'auto-apprentissage, en utilisant un classifier d'approche générative pour aider le principal classifier discriminant entraîné en semi-supervisé à étiqueter les données. Nous nommons cette

stratégie Apprentissage soutenu (Help-Training), et nous l'avons appliqué avec succès aux SVMs et à sa variante LS-SVM.Nos divers algorithmes d'apprentissage semi-supervisé ont été testés sur des données artificielles et réelles et ont donné des résultats encourageants. Cette validation a été appuyée par une analyse montrant les avantages et les limites de chacun des méthodes développées.

Le principe des SVMs consiste à projeter les données de l'espace d'entrée dans un espace de plus grande dimension appelé espace de caractéristiques, de façon à ce que les données deviennent linéairement séparables. Dans cet espace, on construit un hyperplan optimal séparant les classes tel que :

 Les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan.

 La plus petite distance entre les vecteurs et l'hyperplan (la marge), soit maximale**.[23]**



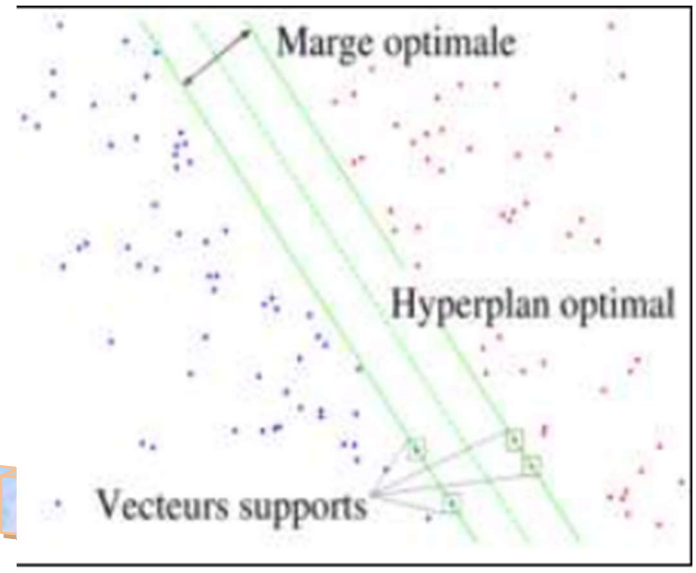


Figure- 2.14 : Hyperplan optimal, Marge optimale et vecteurs de support

2.7.2.1 La marge maximale

On se place désormais dans le cas où le problème est linéairement séparable. Même dans ce cas simple, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque_empirique est le même), mais dont les performances en généralisation peuvent être très différentes. Pour résoudre ce problème, il a été montré[11], qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur.

Il existe des raisons théoriques à ce choix. Vapnik a montré que la capacité des classes d'hyperplans séparateurs diminue lorsque leur marge augmente.

La marge est la distance entre l'hyperplan et les échantillons les plus proches. Ces derniers sont appelés vecteurs supports. [24]

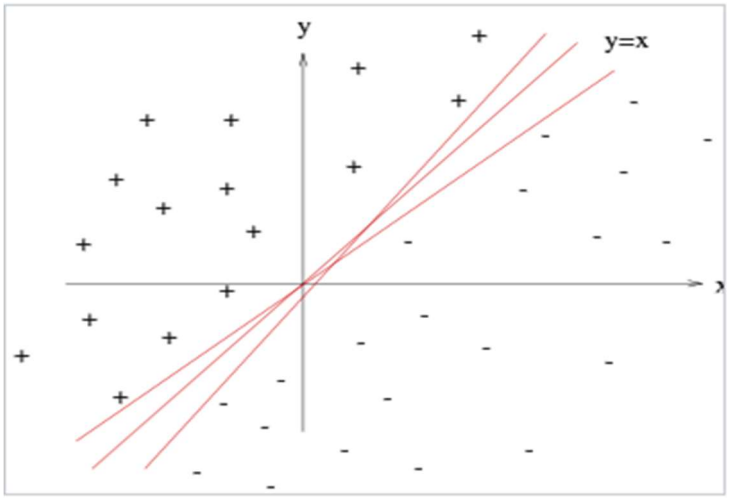


Figure-2.15 : Pour un ensemble de points linéairement séparables

, il existe une infinité d'hyperplans séparateurs

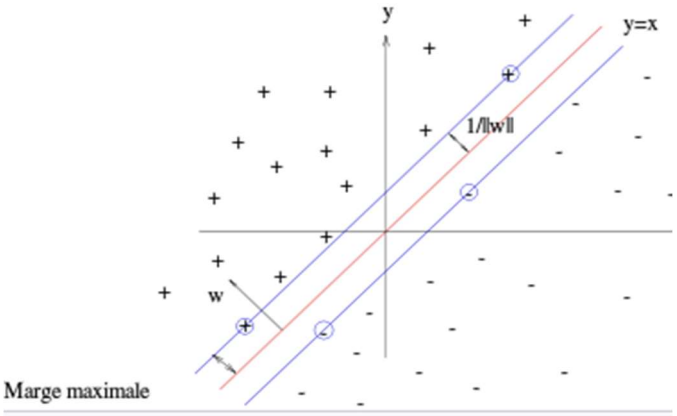


Figure-2.16 : L'hyperplan optimal (en rouge) avec la marge les échantillons entourés sont des *vecteurs supports*

A. Hard marger

Étant donné un ensemble de données d'apprentissage {(xi, yi)}n

i=1 qui est linéairement séparable, SVM à marge dure cherche le plan affine qui sépare

les deux classes avec la marge maximale [1]. Cela revient à résoudre le problème d'optimisation suivant :

$Ø^N$=min llwll$^2_2$

Soit $\hat{w}_H$ et $\hat{b}_H$ résoudre le problème ci-dessus, alors la marge dure classificateur appliqué à une observation invisible x est donné par LH (x) = signe($\hat{w}^T_H x + \hat{b}_H$).

B. Soft marger

Si les données ne sont pas linéairement séparables, les contraintes du Les SVM à marge dure ne peuvent pas tous être satisfaits ensemble. Par conséquent, le coût du problème d'optimisation des marges dures est infini, puisque le minimum sur un ensemble vide est par convention . Dans de tels paramètres, une alternative consiste à utiliser la marge

souple SVM qui par construction tolère que certaines données d'entraînement sont mal classés mais paie le coût de chaque observation mal classée en ajoutant une borne supérieure au nombre de observations d'entraînement mal classées. Plus formellement, le SVM soft margin équivaut à résoudre le problème d'optimisation.

2.7.3 Approximation de la densité locale (LDA)

L'approximation de la densité locale LDA (Local Density Approximation) est l'approximation sur laquelle repose pratiquement toutes les approches actuellement employées. Elle a été proposée pour la première fois par Kohn et Sham, mais la philosophie de cette approximation était déjà présente dans les travaux de Thomas et Fermi. Pour comprendre le concept de LDA rappelons d'abord comment l'énergie cinétique d'un système de particules indépendantes Ts[n] est traité dans l'approximation de Thomas et Fermi[16]. Dans un système homogène, il est bien connu que :

$$\prod_s^{hom}(n) = \frac{3h}{10m}(3\pi^2)^{0,66}\mathrm{n}^{1,66}$$

Avec l'approximation $\prod_n^{ho}$ [n]≈TS LDA [n], la valeur trouvée pour l'énergie cinétique était très inférieure à celle trouvée par traitement de Ts en termes d'orbitales donné par les équations de Kohn-Sham, mais à partir d'ici le concept de LDA s'est tourné vers une autre composante de l'énergie totale pour être très utile et efficace: c'est le terme d'échange qui va être maintenant traité par LDA. L'approximation LDA consiste alors à utiliser directement le résultat d'énergie exacte pour le terme d'échange par particule d'un gaz d'électrons homogène, pour la détermination de l'énergie d'échange d'un gaz d'électrons inhomogène en remplaçant la densité n= constante par n(r) dans l'expression de l'énergie d'échange du gaz d'électrons homogène. On considère le gaz d'électrons inhomogène comme localement homogène, ce qui revient à négliger les effets des variations de la densité. En d'autres termes, elle repose sur l'hypothèse que les termes d'échange ne dépendent que de la valeur locale de n(r). L'énergie d'échange s'exprime alors de la manière suivante :

$$E_{xc}^{LDA} = \int \varepsilon_{xc}[n(r)]n(r)dr$$

Où $[(r)]$ est l'énergie d'échange et de corrélation par particule d'un gaz d'électrons uniforme, qui a été paramétré pour différentes valeurs de la densité électronique.

On pourrait s'attendre à ce qu'une telle approximation, qui ne repose pas sur des critères physiques, ne donne des résultats corrects que dans des cas assez particuliers, où la densité varie peu. L'expérience a montré qu'au contraire, elle permet d'obtenir dans de très nombreux cas une précision équivalente, voire meilleure, que l'approximation de HartreeFock**. [26].**

2.8 Conclusion

Dans ce chapitre nous avons présenté les différentes méthodologies suivies afin de réaliser les objectifs visés dans ce travail et aussi, nous présentons l'évaluation empirique de toutes les approches et techniques proposées pour l'attribution d'auteurs en utilisant le corpus que nous avons conçu pour cette fin.

Dans prochain chapitre nous allons exposer les différentes expériences réalisées et illustrer les résultats obtenus ainsi que les discussions et conclusions aboutées.

CHAPITRE-3

EXPERIENCES ET RESULTATS

CHAPITRE-3
EXPERIENCES ET RESULTATS

3.1 Introduction

Dans ce chapitre nous allons exposer les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 23 auteurs (15 masculins et 8 féminins) dont chacun a écrit 6 textes d'une longueur moyenne de 2000 mots. Ces textes, numéroté de 1 à 6 et qui ont été obtenus après une opération de Reconnaissance Optique de Caractères (OCR), sont classés en deux classes ; textes corrigés et textes non-corrigés chaque classes traduit en deux langues français et anglais deus selon le type de prétraitement appliqué.

Ces textes ont fait l'objet d'une série d'expériences pour voir l'effet sur le Taux d'Attribution d'Auteurs (TAA). Par la suite, les résultats obtenus ont été examinés et discutés et des interprétations et des conclusions objectives ont été données.

3.2 Corpus d'évaluation

3.2.1 Description du Corpus

L'évaluation expérimentale occupe une place importante dans la classification des textes. A l'aide des corpus de tests, nous pouvons voir l'effet TA sur l'attribution d'auteurs. Cependant, les études en attribution d'auteur des textes obtenus après une opération OCR disposent d'un nombre relativement restreint de corpus, encore moins pour les textes corrigés, et non corrigés, traduit en deux langues français et anglais.

De plus, 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 **23** **C**ontemporary **A**rab **W**riters (**OCR23CAW**).

3.2.2 Constituants du Corpus

Le corpus que nous avons conçu contient 23 écrivains arabes contemporains (8 Féminins et 15 Masculins) qui sont : Ghada Saman, Houda Barakat, Kolite Sohil, Nazik Malaika, May Ziada, Nawel Saadawi, Assia Djebar, Latifa Zayat, Mahmoud Akad, Djebran Khalil , Mikhail Naima, Abdelkader Mazini, Sadak Rafie, Taha Hocin, Toufik

Hakim, Hana Mina, Haider Haider, Youcef Idriss, Ibrahim sonaa Allah, Najib Mahfoud, Lotfi Manfalouti, Nadjib Mahmoud et Kharat Edward.

On choisit un livre pour chaque auteur, puis on fait extraire aléatoirement un certain nombre de pages contenant le nombre de mots choisi. Pour chaque auteur on sélectionne 6 textes d'une longueur moyenne de 2000 et plus mots et on les classe en deux catégories ; textes corrigés, textes non-corrigés, et on les traduire en 2 langues (Anglais et Français).

Les textes utilisés pour l'opération d'apprentissage (qui sont les textes numéros ; 3, 4, 5 et 6 pour chaque auteur. On utilise les deux catégories traduites en deux langues. Cependant, chacun des textes utilisés aussi pour l'opération de test (qui sont les textes numéros1 et 2 pour chaque auteurs). Dans ce cas on trouve quatre types de textes (c'est-à-dire on a texte-An _Corrigé, texte-An_ non-Corrigé, texte-Fr _Corrigé, texte-Fr_ non-Corrigé).

Les textes considérés ont été pris à partir des romans de ces écrivains. Les détails des informations sur les écrivains et les textes de notre corpus sont donnés dans les tableaux suivants

Tableau-3.1 : Récapitulatif du Corpus (Ecrivains Féminins)

Ecrivains	Pays de Naissance	Période	Nbre de livres	Langue	Textes	Nbre de mots / texte	Utilisation
Assia Djebar	Algérie	1936-2015	26 -livres	AR/FR	Asia-1	2130	Test
					Asia-2	1956	Test
					Asia-3	2408	Apprentissage
					Asia-4	2361	Apprentissage
					Asia-5	2161	Apprentissage
					Asia-6	2510	Apprentissage
Nazik Malaika	Irak	1923-2007	25-livres	AR	Nazik-1	3211	Test
					Nazik-2	1682	Test
					Nazik-3	1035	Apprentissage
					Nazik-4	886	Apprentissage
					Nazik-5	1015	Apprentissage
					Nazik-6	840	Apprentissage
Ghada Saman	Syrie	1942 – à ce jour	46-livres	AR	Ghada-1	3088	Test
					Ghada-2	1825	Test
					Ghada-3	2960	Apprentissage
					Ghada-4	1697	Apprentissage
					Ghada-5	3273	Apprentissage
					Ghada-6	3151	Apprentissage
Houda Barakat	Liban	1952 – à ce jour	12-livres	AR	Houda-1	2098	Test
					Houda-2	1984	Test
					Houda-3	2073	Apprentissage
					Houda-4	2116	Apprentissage
					Houda-5	2069	Apprentissage
					Houda-6	1929	Apprentissage
Kolite Sohil	Syrie	1931 – à ce jour	29-livres	AR	Kolite-1	2329	Test
					Kolite-2	1608	Test
					Kolite-3	1995	Apprentissage
					Kolite-4	1465	Apprentissage
					Kolite-5	1664	Apprentissage
					Kolite-6	1969	Apprentissage
Latifa Zayat	Egypte	1923-1996	12-livres	AR	Latifa-1	2209	Test
					Latifa-2	1865	Test
					Latifa-3	2232	Apprentissage
					Latifa-4	1983	Apprentissage
					Latifa-5	1993	Apprentissage
					Latifa-6	1753	Apprentissage
May Ziada	Palestine	1886 – 1941	19-livres	FR-EN-ES-IT-	May-1	1536	Test
					May-2	1617	Test
					May-3	1580	Apprentissage
					May-4	1612	Apprentissage
					May-5	1638	Apprentissage
					May-6	1606	Apprentissage
Nawel Saadawi	Egypte	1931- 2021	34-livres	AR	Nawal-1	1878	Test
					Nawal-2	1698	Test
					Nawal-3	1673	Apprentissage
					Nawal-4	1736	Apprentissage
					Nawal-5	2811	Apprentissage
					Nawal-6	2238	Apprentissage

Tableau-3.2: Récapitulatif du Corpus (Ecrivains Masculins)

Ecrivains	Pays de Naissance	Période	Nbre de livres	Langue	Textes	Nbre de mots / texte	Utilisation
Najib Mahfoud	Egypte	1911-2006	49 livres	AR	Najib -1	2749	test
					Najib -2	3115	test
					Najib -3	2491	Apprentissage
					Najib -4	2426	Apprentissage
					Najib -5	2638	Apprentissage
					Najib -6	2042	Apprentissage
Kharat Edward	Egypte	1926-2015	30 l livres	AR	Kharat-1	3471	test
					Kharat-2	3049	test
					Kharat-3	2578	Apprentissage
					Kharat-4	3840	Apprentissage
					Kharat-5	3524	Apprentissage
					Kharat-6	3735	Apprentissage
Abdelkader Mazini	Egypte	1889 - 1949	19 livres	AR-EN	Mazini-1	2075	test
					Mazini-2	2024	test
					Mazini-3	2088	Apprentissage
					Mazini-4	1979	Apprentissage
					Mazini-5	2078	Apprentissage
					Mazini-6	1974	Apprentissage
Djebran Khalil	Liban	1883-1931	163 livres	AR-EN-FR	Djebran-1	2323	test
					Djebran-2	2525	test
					Djebran-3	2463	Apprentissage
					Djebran-4	2345	Apprentissage
					Djebran-5	2413	Apprentissage
					Djebran-6	717	Apprentissage
Haidar Haidar	Syrie	1936- à ce jour	40 livres	AR	Haider-1	1911	test
					Haider-2	2282	test
					Haider-3	2018	Apprentissage
					Haider-4	2196	Apprentissage
					Haider-5	2710	Apprentissage
					Haider-6	2105	Apprentissage
Hana Mina	Syrie	1924-2018	21 livres	AR	Hana-1	2026	test
					Hana-2	2315	test
					Hana-3	1984	Apprentissage
					Hana-4	1960	Apprentissage
					Hana-5	2363	Apprentissage
					Hana-6	2069	Apprentissage
Ibrahim Sonaa Allah	Egypte	1937- à ce jour	47 l livres	AR	Sonaa-1	3302	test
					Sonaa-2	3220	test
					Sonaa-3	3287	Apprentissage
					Sonaa-4	3652	Apprentissage
					Sonaa-5	2486	Apprentissage
					Sonaa-6	2042	Apprentissage

Tableau-3.2: Récapitulatif du Corpus (Ecrivains Masculins) **(Suite)**

Ecrivains	Pays de Naissance	Période	Nbre de livres	Langue	Textes	Nbre de mots / texte	Utilisation
Mahmoud Akad	Egypte	1889 - 1964	689 livres	AR	Akad-1	2033	test
					Akad-2	2011	test
					Akad-3	2023	Apprentissage
					Akad-4	2036	Apprentissage
					Akad-5	2093	Apprentissage
					Akad-6	2072	Apprentissage
Mikhail Naima	Liban	1889-1988	76 livres	AR	Mikhail-1	2488	test
					Mikhail-2	2460	test
					Mikhail-3	1858	Apprentissage
					Mikhail-4	2453	Apprentissage
					Mikhail-5	2403	Apprentissage
					Mikhail-6	2331	Apprentissage
Nadjib Mahmoud	Egypte	1905-1993	110 livres	AR	Nadjib-1	2213	test
					Nadjib-2	1992	test
					Nadjib-3	2174	Apprentissage
					Nadjib-4	2096	Apprentissage
					Nadjib-5	2086	Apprentissage
					Nadjib-6	2126	Apprentissage
Sadak Rafei	Egypte	1880-1973	167 livres	AR	Rafei-1	2088	test
					Rafei-2	1974	test
					Rafei-3	2024	Apprentissage
					Rafei-4	2042	Apprentissage
					Rafei-5	1952	Apprentissage
					Rafei-6	2187	Apprentissage
Taha Hocin	Egypte	1889-1973	381 livres	AR-FR-LAT	Taha-1	2028	test
					Taha-2	2034	test
					Taha-3	2101	Apprentissage
					Taha-4	2082	Apprentissage
					Taha-5	2069	Apprentissage
					Taha-6	2035	Apprentissage
Toufik Hakim	Egypte	1898-1987	233 livres	AR	Toufik-1	2010	test
					Toufik-2	1973	test
					Toufik-3	2012	Apprentissage
					Toufik-4	1999	Apprentissage
					Toufik-5	2024	Apprentissage
					Toufik-6	1982	Apprentissage
Lotfi Manfalouti	Egypte	1876-1924	107 livres	FR	Lotfi-1	2691	test
					Lotfi-2	2384	test
					Lotfi-3	2779	Apprentissage
					Lotfi-4	2809	Apprentissage
					Lotfi-5	1929	Apprentissage
					Lotfi-6	2672	Apprentissage
Youssef Idriss	Egypte	1927-1991	108 livres	AR	Idriss-1	3630	test
					Idriss-2	3724	test
					Idriss-3	3439	Apprentissage
					Idriss-4	2748	Apprentissage
					Idriss-5	2737	Apprentissage
					Idriss-6	3599	Apprentissage

3.2.3 Préparation des documents du corpus

Les documents du corpus doivent être préparés avant leur utilisation pour l'attribution de leurs véritables auteurs. La phase de préparation se résume en opérations pour préparer ce texte :

- Scanner les pages choisis et les enregistrées en format (.jpeg)
- Convertir ces images en fichier Word (txt) à l'aide d'un OCR.
- Faire les opérations de prétraitement mentionnées dans la section (2.6.2).
- Les documents textes obtenus corrigés et non corrigés sont traduits en deux langues Français et Anglais et les sauvegarder dans des fichiers Word
- Les documents textes obtenus sont enregistrés sous forme UTF-8 (Encodage basé sur l'Unicode qui peut être codé sur 4 octets).

Il est à noter, qu'on a utilisé l'encodage UTF-8 pour encoder tous les textes du corpus, car ce dernier couvre un très grand nombre de caractères, et qui est implicitement capable d'encoder la majorité des langues vu qu'il est encodé sur 4 octets. En revanche, l'utilisation de cet encodage est payée en termes de temps de calcul et en termes de mémoire.

Par la suite, le corpus est divisé en deux sous-ensembles (apprentissage et Test) selon la règle (2/3 pour l'apprentissage et 1/3 pour le test) appliquée dans les bases de données;

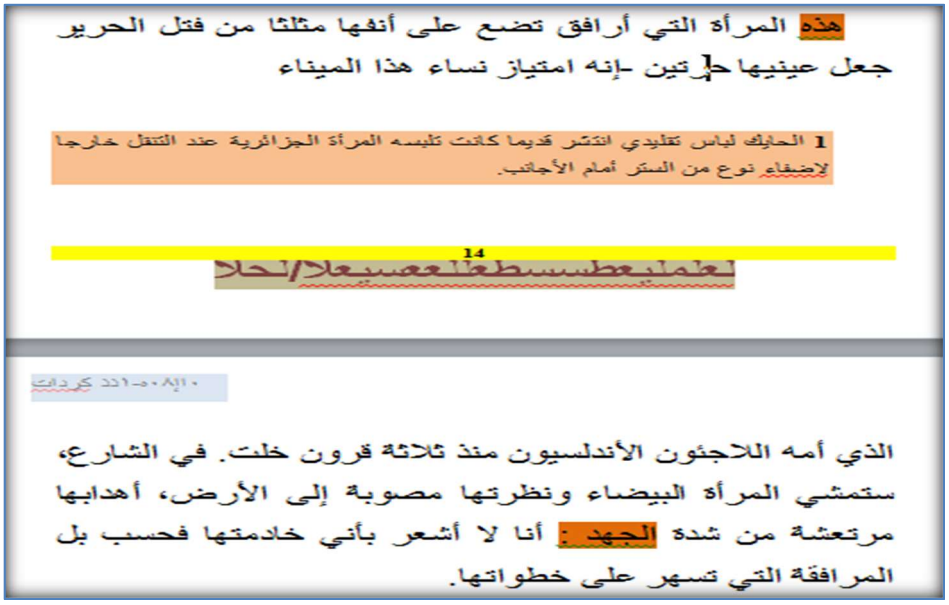
- L'ensemble d'apprentissage est constitué des textes (numéros 3, 4, 5 et 6 pour chaque auteur) de chaque catégorie Français et Anglais.
- L'ensemble de test est constitué des textes (numéros 1 et 2 pour chaque auteur) de chaque catégorie Français et Anglais.

Au totale, le corpus contient 552 textes (Anglais / Français) divisés comme suit ; 368 textes traduits corrigés et non corrigé (pour l'apprentissage) et 184 textes pour le test. La correction est faite manuellement, en corrigeant les erreurs apparus après l'étape d'OCR uniquement avant la phase de traduction automatique en utilisant (Google traduction). Les erreurs de la traduction automatique ne sont pas traitées.

3.2.4 Exemples de textes obtenus après une opération OCR

Après le processus de scandes documents en PDF, les résultats obtenus sont considérés comme des documents modifiables (format Word) pour corriger les erreurs, ajouter ou supprimer tout ce qui est supplémentaire. Ci-dessous nous passons en revue des

exemples de textes que nous avons obtenus après l'opération OCR afin de les utiliser dans les prochaines expériences (chaque couleur exprime un type d'erreurs).


هذه المرأة التي أرافق تضع على أنفها مثلثا من فتل الحرير
جعل عينيها حرتين -إنه امتياز نساء هذا الميناء

1 الحايك لباس تقليدي انتشر قديما كانت تلبسه المرأة الجزائرية عند التنقل خارجا لإضفاء نوع من الستر أمام الأجانب.

14

الذي أمه اللاجئون الأندلسيون منذ ثلاثة قرون خلت. في الشارع، ستمشي المرأة البيضاء ونظرتها مصوبة إلى الأرض، أهدابها مرتعشة من شدة الجهد : أنا لا أشعر بأني خادمتها فحسب بل المرافقة التي تسهر على خطواتها.

Figure-3.1: Exemple de texte Arabe non-corrigé.

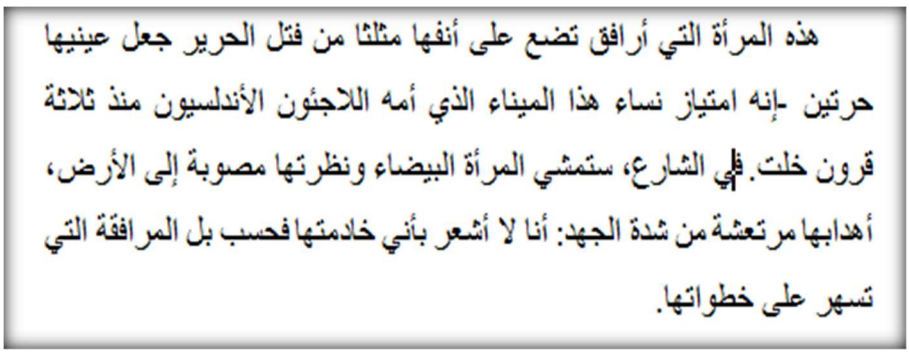

هذه المرأة التي أرافق تضع على أنفها مثلثا من فتل الحرير جعل عينيها حرتين -إنه امتياز نساء هذا الميناء الذي أمه اللاجئون الأندلسيون منذ ثلاثة قرون خلت. في الشارع، ستمشي المرأة البيضاء ونظرتها مصوبة إلى الأرض، أهدابها مرتعشة من شدة الجهد: أنا لا أشعر بأني خادمتها فحسب بل المرافقة التي تسهر على خطواتها.

Figure-3.2: Exemple de texte Arabe corrigé.

3.2.5 Exemples de textes obtenus après une opération Traduction Automatique

Finalement les textes Word obtenues après une opération OCR On les traduire correctement en français et anglais, Ci-dessous nous passons en revue des exemples de textes que nous avons obtenus après l'opération TA afin de les utiliser dans les prochaines expériences.

Cette femme que j'accompagne met sur son nez un triangle de mèches de soie, qui lui a libéré les yeux - c'est l'apanage des femmes de ce port.

1 Le haïk est un vêtement traditionnel qui s'est répandu dans le passé, et les femmes algériennes le portaient lors de leurs voyages à l'étranger afin de mettre une sorte de voile devant les étrangers.

14

Savoir, éternuer, chercher du miel / pour une solution

Qui était la mère des réfugiés andalous il y a trois siècles. Dans la rue, la femme blanche marchera les yeux pointés vers le sol, ses franges frémissantes avec l'intensité de l'effort: je sens non seulement que je suis sa servante, mais la compagne qui regarde ses pas.

Figure-3.3: Exemple de texte Français Non corrigé.

Cette femme que j'accompagnais portait sur le nez un triangle de mèches de soie qui lui rendait les yeux libres - c'est l'apanage des femmes de ce port qui fut la mère des réfugiés andalous il y a trois siècles. Dans la rue, la femme blanche marchera les yeux pointés vers le sol, ses franges frémissantes avec l'intensité de l'effort: je sens non seulement que je suis sa servante, mais la compagne qui regarde ses pas.

Figure-3.4: Exemple de texte Français corrigé.

This woman whom I accompany puts on her nose a triangle of silk wicks, which has made her eyes free - it is the prerogative of the women of this port.

1 The haik is a traditional dress that spread in the past, and Algerian women used to wear it when traveling abroad in order to put a kind of veil in front of foreigners.

14

To know, to sneeze, to get honey / for a solution

Who was the mother of Andalusian refugees three centuries ago. On the street, the white woman will walk with her eyes pointed at the ground, her fringes quivering from the intensity of the effort: I feel not only that I am her servant but the companion who watches her steps.

Figure-3.5: Exemple de texte Anglais Non corrigé.

This woman whom I accompanied wore a triangle of silk wicks on her nose that made her eyes free - it is the prerogative of the women of this port that was the mother of Andalusian refugees three centuries ago. On the street, the white woman will walk with her eyes pointed at the ground, her fringes quivering with the intensity of the effort: I feel not only that I am her servant, but the companion who watches her steps.

Figure-3.6: Exemple de texte Anglais corrigé

Ces textes sont obtenus à partir d'une opération OCR, puis ont été traduits en utilisant Google traduction (voir figure ci-dessus)

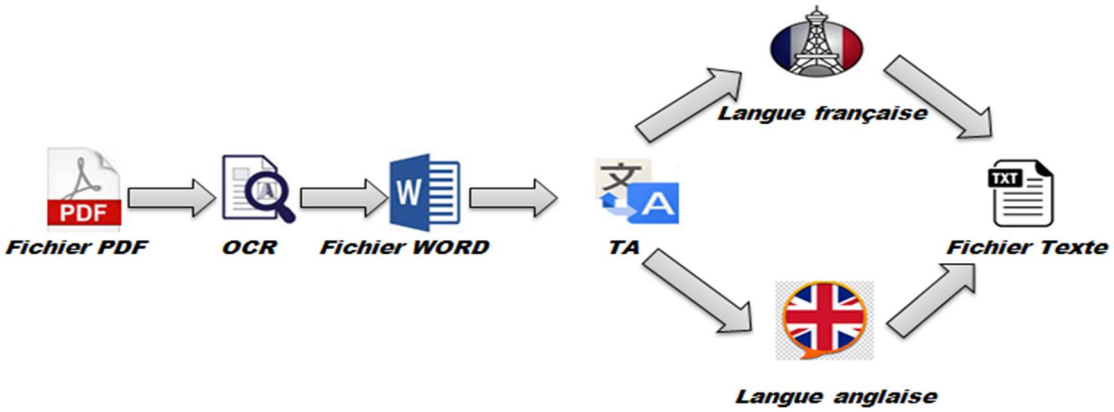


Figure-3.7: Processus de conversion des textes scannés en textes traduits.

3.3 Travaux d'expérimentation

3.3.1 Protocole expérimental

Dans ce mémoire, la tâche d'attribution d'auteurs est effectuée en utilisant le N-grams caractères comme caractéristique et trois classifieurs ; MLP, LDA et SVM. Ces techniques ont été utilisées pour voir si la tâche d'attribution d'auteurs tient toujours en utilisant les textes obtenus par une opération OCR et traduits en d'autres langues or que la langue utilisée par l'auteur original. Le Taux d'Attribution d'Auteurs (TAA) est défini par la relation suivante :

$$TAA = \frac{nombre\ documents\ correctement\ attribués}{nombre\ document\ testé} \times 100$$

Ce travail expérimental est organisé en quatre séries d'expériences et chaque série comporte plusieurs cas d'applications selon la valeur de N-grams.

- Dans la première série, les textes utilisés dans la phase d'apprentissage et les textes utilisés dans la phase de test sont tous les deux des textes corrigés.
- Dans la deuxième série, les textes utilisés dans la phase d'apprentissage sont des textes corrigés et les textes utilisés dans la phase de test sont des textes non corrigés.
- Dans la troisième série, les textes utilisés dans la phase d'apprentissage et les textes utilisés dans la phase de test sont tous les deux des textes non corrigés.

3.3.2 Séries d'expériences et résultats obtenus

3.3.2.1 1ère série : Utilisation des textes corrigés pour l'apprentissage et pour le test

Cette série d'expériences vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes traduits et corrigés en anglais/Français. On utilise 4 textes pour l'apprentissage et 2 textes pour le test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 300 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

A) Textes traduits en Anglais

A.1) Auteurs Masculins

Tableau-3.3: TAA pour les textes traduits en Anglais des auteurs Masculins

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	86	83	80	83	70	80
MLP	90	86	86	90	86	90
LDA	80	86	76	83	83	76

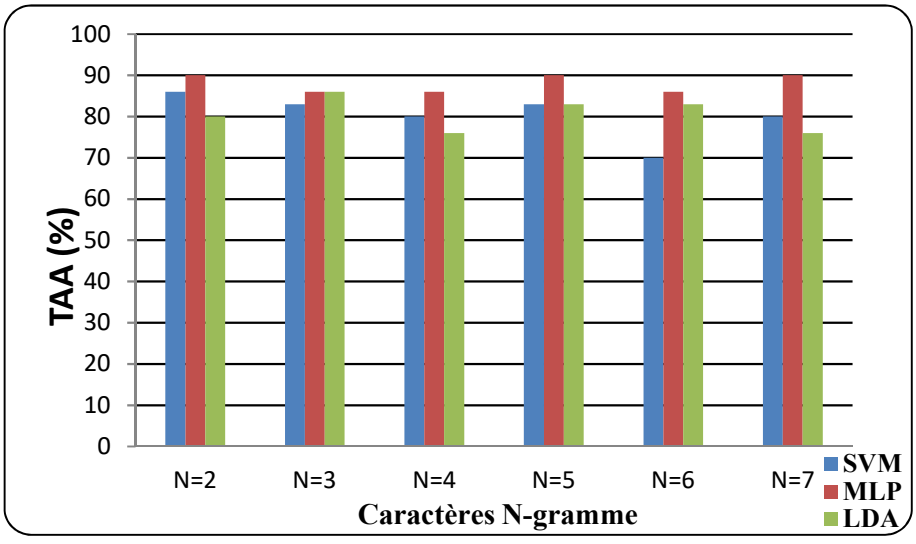


Figure -3.8: TAA pour les textes traduits en Anglais des Auteurs Masculins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 70-86% pour le SVM, 86-90 % pour MLP et 76-86% pour le LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant

d'une opération OCR qui a subi un prétraitement non complet, ainsi que l'opération TA en Anglais.

A.2) Auteurs Féminins

Tableau-3.4: TAA pour les textes traduits en Anglais des auteurs Féminins

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	81	81	81
MLP	81	87	75	81	81	81
LDA	87	87	87	81	87	81

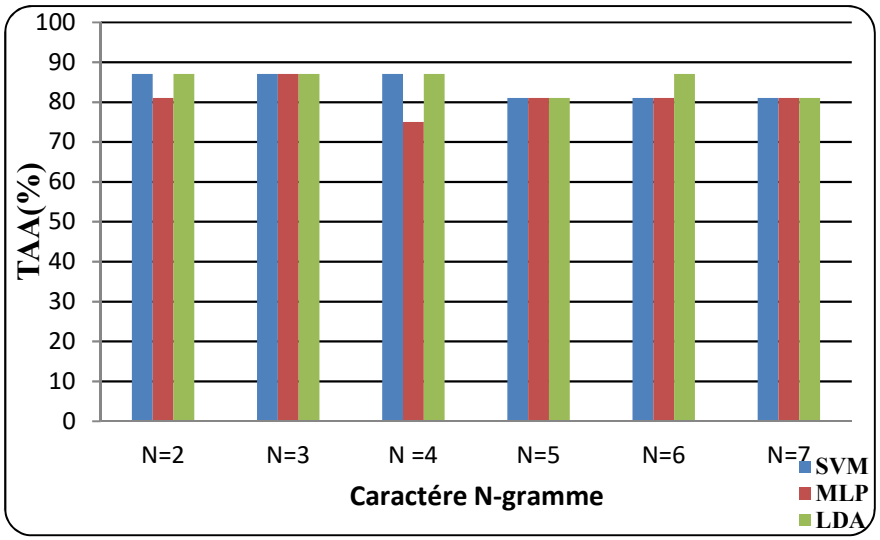


Figure- 3.9 : TAA pour les textes traduits en Anglais des Auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 81-87% pour le LDA et le SVM, et entre 75-87% pour le MLP. Ceci peut être expliqué, aussi, par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet, ainsi que l'opération TA en Anglais.

B) Textes traduits en Français

B.1) Auteurs Masculins

Tableau-3.5 : TAA pour les textes traduits en Français des Auteurs Masculins

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	90	86	86	-	-	73
MLP	86	90	93	90	80	80
LDA	93	90	80	76	83	76

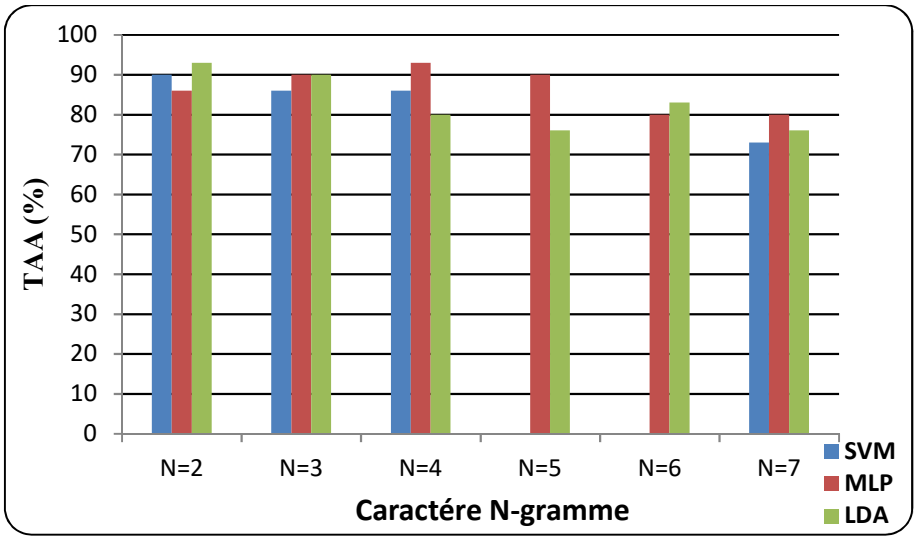


Figure-3.10 : TAA pour les textes traduits en Français des Auteurs Masculins.

D'après les résultats obtenus, on peut voir clairement que le TAA de cette expérience est entre 73-90% pour le SVM, 80-90% pour MLP et 76-93% pour LDA. Il est à noter que les TAA de (N=5 et N=6) sont non-identifiés.

B.2) Auteurs Féminins

Tableau-3.6 : TAA pour les textes traduits en Français des Auteurs Féminins.

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	87	87	87
MLP	87	87	87	87	81	81
LDA	87	87	87	87	87	81

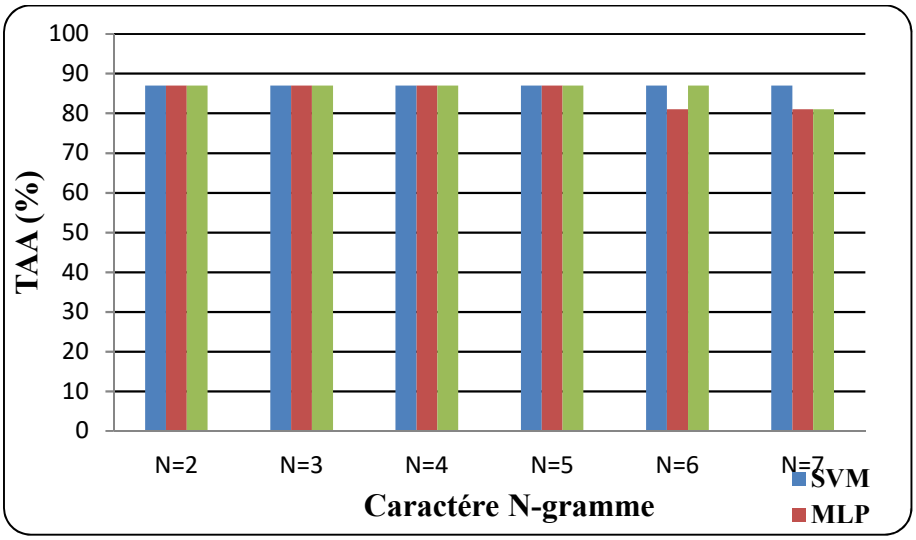


Figure-3.11 : TAA pour les textes traduits en Français des Auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 81-87% pour le SVM, MLP et LDA. . Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Français.

3.3.2.2 2ème série : Utilisation des textes corrigés pour l'apprentissage et des textes non corrigés pour le test

Cette série d'expérience vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes Anglais/Français. On utilise 4textes corrigés pour l'apprentissage et 2 textes non corrigés pour le test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 300 »).

A) Textes traduits en Anglais

A.1) Auteurs Masculins

Tableau-3.7 : TAA pour les textes traduits en Anglais des auteurs Masculins

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	86	83	76	76	76	76
MLP	93	80	90	90	87	97
LDA	83	80	83	83	76	76

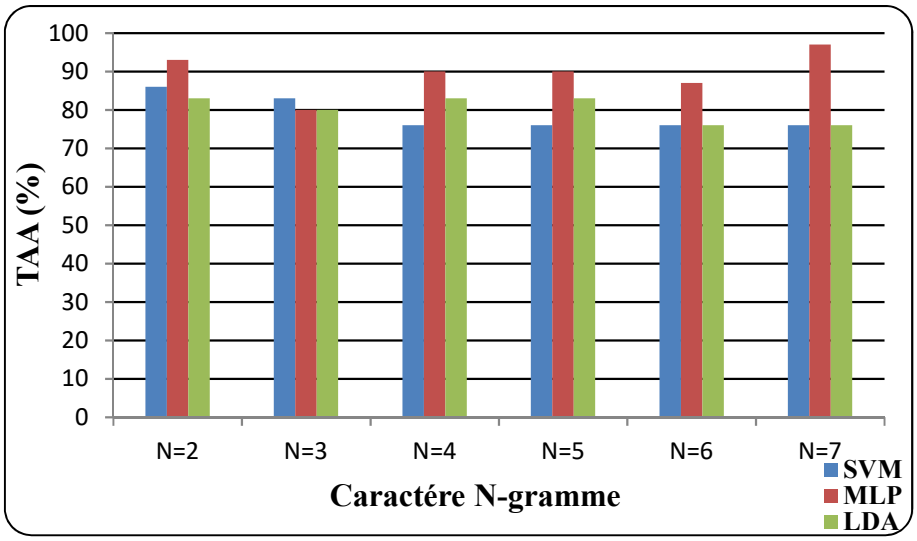


Figure-3.12 : TAA pour les textes traduits en Anglais des auteurs Masculins**.**

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 76-86% pour le SVM, 80-97% pour MLP et 76-83% pour LDA. . Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Anglais.

A.2) Auteurs Féminines

Tableau 3.8 : TAA pour les textes traduits en Anglais des auteurs Féminins.

classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	81	81	81
MLP	81	87	87	87	87	87
LDA	81	87	87	87	87	75

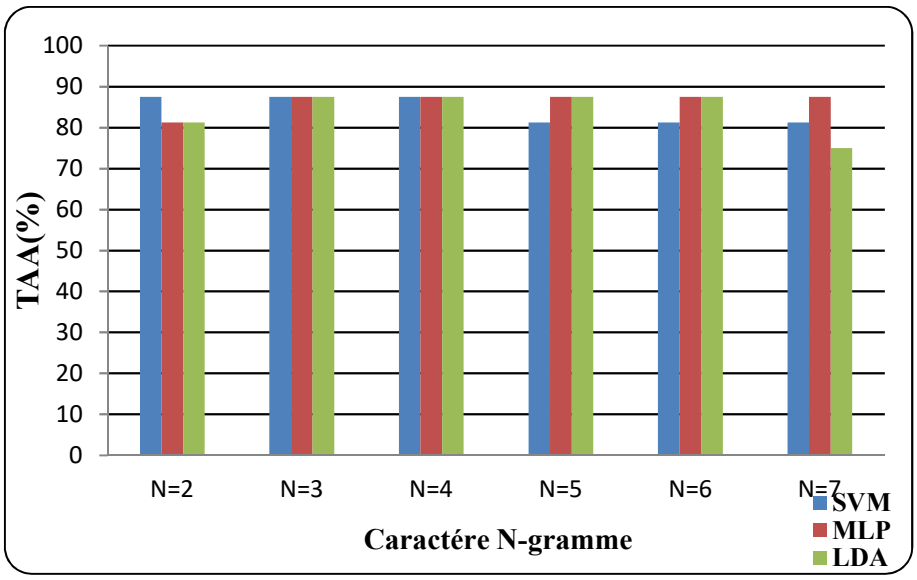


Figure-3.13 : TAA pour les textes traduits en Anglais des auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 81-87% pour le SVM et MLP et 75-87% pour LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Anglais.

B) Textes traduits en Français

B.1) Auteurs Masculine

Tableau-3.9 : TAA pour les textes traduits en Français des Auteurs Masculins**.**

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	pas identifié	80	93	pas identifié	pas identifié	63
MLP	83	90	93	86	76	80
LDA	90	90	93	80	86	76

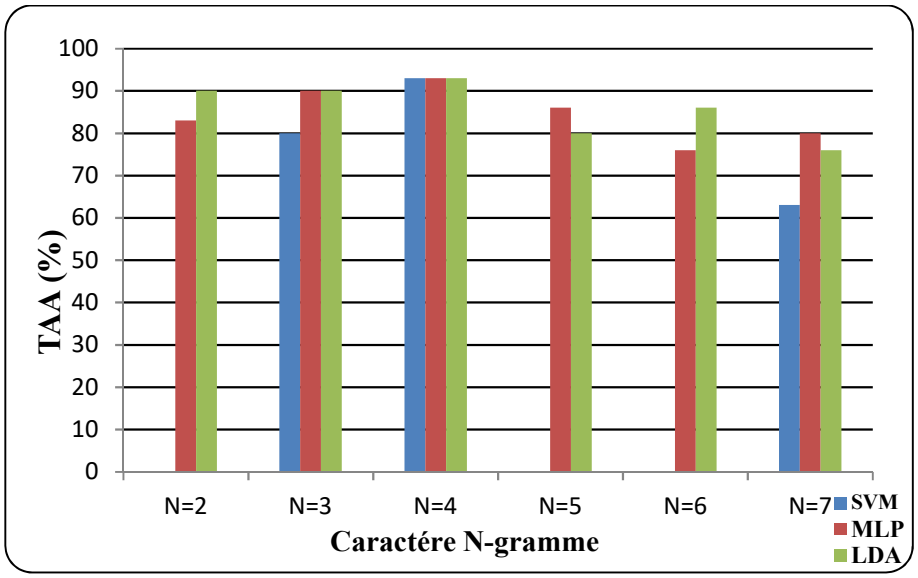


Figure-3.14 : TAA pour les textes traduits en Français des Auteurs Masculins**.**

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 63-93% pour le SVM, 76-93% pour le MLP et 76-93% pour le LDA. . Il est à noter que les TAA de (N=2, N=5 et N=6) sont non-identifiés.

B.2) Auteurs Féminines

Tableau-3.10 : TAA pour les textes traduits en Français des Auteurs Féminins.

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	87	87	81
MLP	87	87	87	81	81	81
LDA	76	87	87	81	87	87

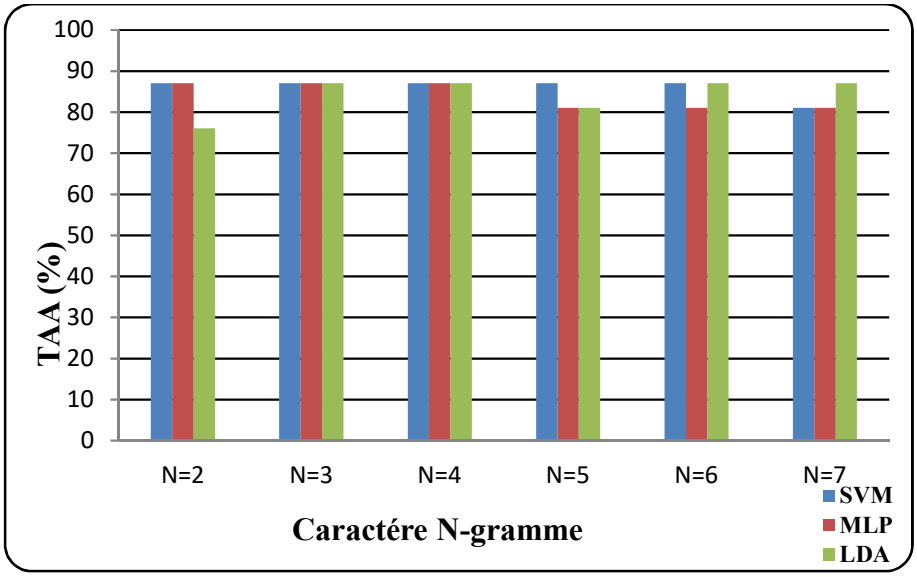


Figure-3.15 : TAA pour les textes traduits en Français des Auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 81-87% pour le SVM, 81-87% pour le MLP et 76-87 % pour le LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Français.

3.3.2.3 3ème série : Utilisation des textes non corrigés pour l'apprentissage et pour le test

Cette série d'expérience vise à déterminer le nombre approprié de caractères (N-gramme) pour avoir le meilleur taux TAA en utilisant les textes Anglais/Français. On utilise 4 textes Non corrigés pour l'apprentissage et 2 textes Non corrigés pour le test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 300 »).

A) Textes traduits en Anglais

A.1) Auteurs Masculine

Tableau-3.11 : TAA pour les textes traduits en Anglais des auteurs Masculins**.**

Classifies	N=2	N=3	N=4	N=5	N=6	N=7
SVM	90	76	76	73	80	76
MLP	86	80	93	86	76	86
LDA	76	86	83	80	73	76

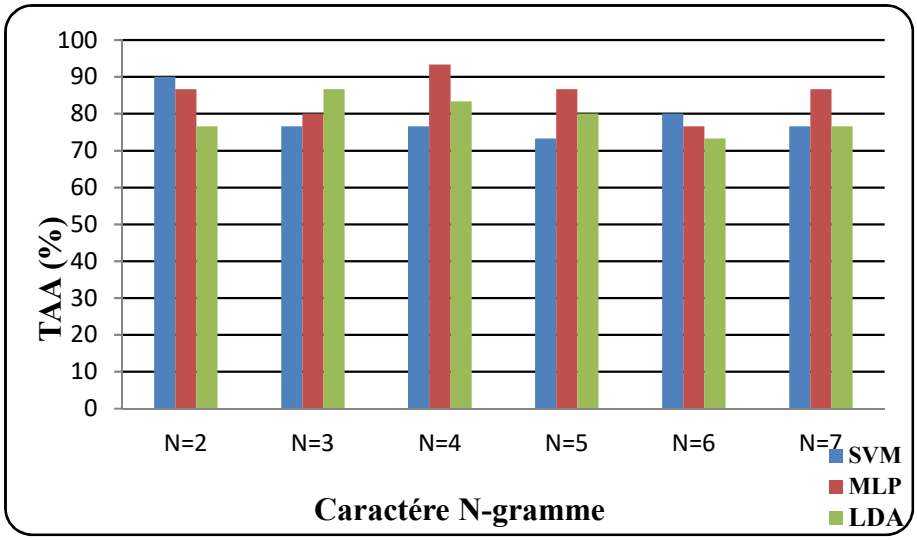


Figure-3.16 : TAA pour les textes traduits en Anglais des auteurs Masculins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 73-90% pour le SVM, 76-93% pour MLP et 73-86 % LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Anglais.

A.2) Auteurs Féminins

Tableau-3.12 : TAA pour les textes traduits en Anglais des auteurs Féminins.

Classifie	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	87	87	81
MLP	81	87	75	81	81	81
LDA	87	87	81	81	81	81

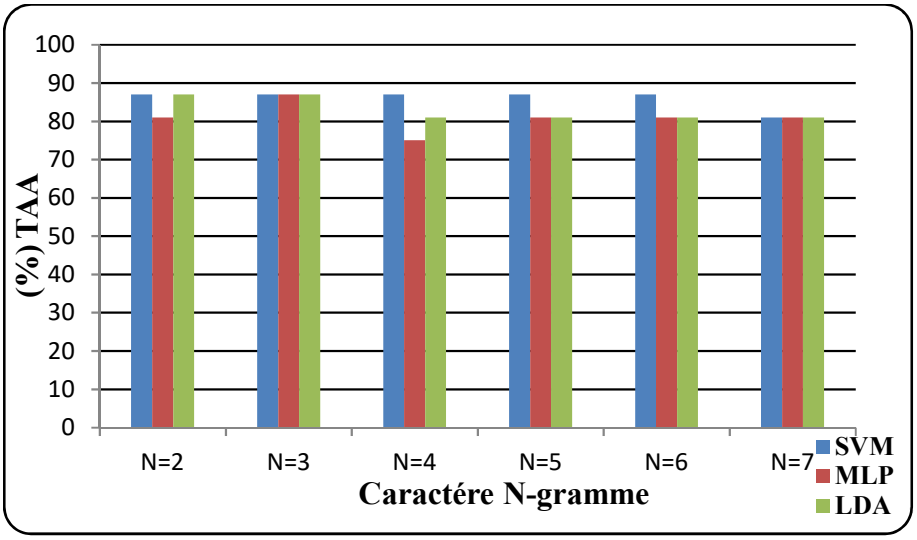


Figure-3.17 : TAA pour les textes traduits en Anglais des auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 81-87% pour le SVM, 75-87% pour MLP et 81-87 % LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Anglais.

B) Textes traduits en Français

B.1) Auteurs Masculine

Tableau-3.13 : TAA pour les textes traduits en Français des Auteurs Masculins.

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	86	83	83	80	73	63
MLP	90	83	86	83	83	86
LDA	80	86	86	80	83	73

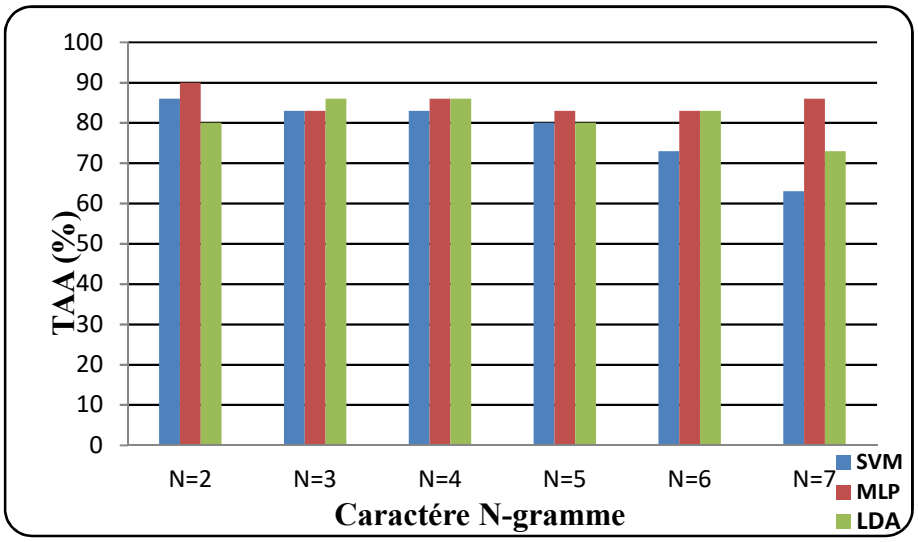


Figure-3.18 : TAA pour les textes traduits en Français des Auteurs Masculins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 63-86% pour le SVM, 83-90% pour MLP et 73-86 % LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Français.

B.2) Auteurs Féminins

Tableau-3.14 : TAA pour les textes traduits en Français des Auteurs Féminins.

Classifier	N=2	N=3	N=4	N=5	N=6	N=7
SVM	87	87	87	87	87	75
MLP	87	75	87	86	89	89
LDA	75	81	86	87	75	81

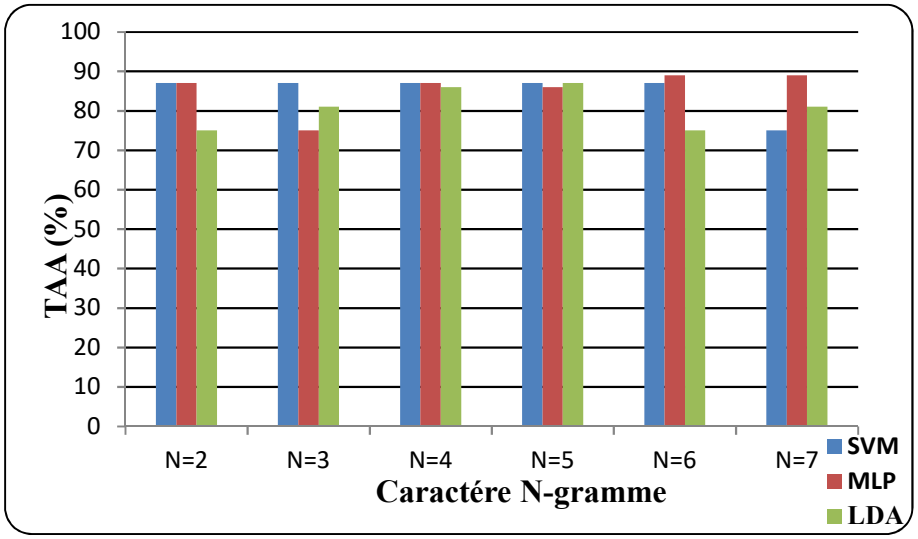


Figure-3.19 : TAA pour les textes traduits en Français des Auteurs Féminins.

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 75-87% pour le SVM, 75-89% pour MLP et 75-87 % LDA. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant d'une opération OCR qui a subi un prétraitement non complet. Puis, opération TA en Français.

3.4 Conclusion

Dans ce chapitre nous avons effectué des expériences d'attribution d'auteur des documents textes écrits et textes traduit automatique obtenus après une opération OCR. L'évaluation expérimentale a été réalisé en utilisant une base de données qu'on conçu pour cette fin et qu'on a appelé « Optical Character Recognition 23 Arabs Contemporary Writers » (OCR23ACW). Les résultats obtenus ont montré que les erreurs dans la conversion OCR et le traducteur automatique affectent de manière significative le taux d'attribution TAA.

La méthode utilisée pour l'attribution d'auteurs est basée sur l'utilisation des N-grammes et classifier (MLP SVM et LDA). Les expériences sont effectuées sur une base de données (Corpus). On peut conclure que la meilleure technique pour les textes traduits est MLP pour différents types de n-gramme.

Conclusion générale

Conclusion générale

Travail réalisé

Le thème que nous avons étudié dans ce mémoire s'intéresse à l'effet de traduction automatique, après acquisition à l'aide d'un OCR, des documents textes sur la tâche d'attribution d'auteurs. Dans ce travail, nous avons abordé l'attribution d'auteurs des textes anonymes, en particulier, nous nous sommes intéressés par les textes issus d'une opération de reconnaissance optique de caractères, ensuite subis une opération de traduction automatique à plusieurs langues. Pour ce type de texte, une application particulière d'identification des textes OCR a été effectuée.

Le corpus que nous avons conçu pour réaliser nos expériences, est construit autours d'une base de données constituée de 23 écrivains dont on a choisi un livre pour chaque auteur, puis on fait l'extraction aléatoire d'un certain nombre de pages contenant (~2000) mots. Pour chaque auteur on a sélectionné 6 textes d'une longueur moyenne de 2000 et on les a classés en deux catégories ; textes corrigés, textes non-corrigés, et on les a traduits automatiquement en deux langues ; Anglais et Français. Les textes utilisés pour l'opération d'apprentissage (qui sont les textes numérotés ; 3, 4, 5 et 6 pour chaque auteur. En utilisant les deux catégories traduites en deux langues. Cependant, chacun des textes utilisés aussi pour l'opération de test (qui sont les textes numérotés 1 et 2 pour chaque auteurs). Dans ce cas on trouve quatre types de textes (c-à-d on a texte-An _Corrigé, texte-An_ non-Corrigé, texte-Fr_Corrigé, texte-Fr_ non-Corrigé).

Ce mémoire avait pour ambition d'étudier le style des auteurs afin de trouver le véritable auteur, en appliquant des caractéristiques telles que caractère N-grammes et des classifieurs telles que le MLP et SVM et LDA. L'originalité de ce travail de recherche est que la tâche d'Attribution d'Auteurs (AA) a été appliquée aux textes traduits qui ont été reporté fidèlement en les comparants avec les fichiers et des textes (inconnues) et non pas été écrits directement par les auteurs.

- **Les résultats obtenus**

Les résultats obtenus étaient très encourageants vu la contrainte liée à la taille des textes choisis (2000 mots uniquement), d'où on a vu l'importance et l'efficacité de la représentation en n-grammes pour la tâche de l'AA.

A partir de ces résultats, on remarque un changement de la valeur TAA à partir de N=2 jusqu'à N=7 tel qu'on a constaté que le classifieur MLP a donné d'excellents résultats pour les textes masculins et la représentation en n-grammes sont les plus appropriés aux tâches d'AA quand il s'agit de textes de taille réduite, d'autre part le classifieur SVM avec caractère n-grammes adonné un bon résultat pour textes féminins. D'après cette étude on peut constater que la langue la plus proche de l'Arabe est l'Anglais.

Références bibliographiques

Références bibliographiques

[1] Bozkurt, I. N., Bağlıoğlu, Ö., & Uyar, E. (2007). Authorship attribution: performance of various features and classification methods. In 22nd International Symposium on Computer and Information Sciences, ISCIS 2007-Proceedings (pp. 158-162). IEEE

[2] https://www.google.com/url?q=https://www.larousse.fr/encyclopedie/peinture/attribution/

[3] MENASRI, R., & YAKOUBI, M. (2020). Etude et analyse des effets d'acquisition optique à l'aide d'un OCR des textes arabes sur l'attribution d'auteurs (Doctoral dissertation, Univ M'sila).

[4] : Brixtel, R., Lecluze, C., & Lejeune, G. (2015, June). Attribution d'Auteur: approche multilingue fondée sur les répétitions maximales. In Actes de la 22e conf\'erence sur le Traitement Automatique des Langues Naturelles. Articles longs (pp. 208-219).

[5] : La technique de la stylométrie appliquée au Livre de Mormon. Agnès Boltoukhine

[6] : Jalam, R. (2003). Apprentissage automatique et catégorisation de textes multilingues. PhD Tesis, Université Lumiere Lyon, 2.

[7] : MATALLAH, H. (2011). Classification Automatique de Textes Approche Orientée Agent (Doctoral dissertation).

[8] : BENYAHIA, A. (2019). Etude et analyse sur les performances des techniques d'identification d'auteurs à partir des documents écrits et des documents transcrits (Doctoral dissertation, UNIVERSITE MOHAMED BOUDIAF-M'SILA).

[9] : SAHRAOUI, S. (2013). Identification de la langue et catégorisation thématique de textes d'un corpus multilingue en utilisant les réseaux de neurones artificiels RNA (Doctoral dissertation, FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE-UNIVERSITE DE M'SILA).

[10] : Mataalah Hocine « classificat ion automatique de textes Orienté Agent » faculté des sciences –algerie2010-2011

[11] Larivée, S. (1995). La notion de plagiat scientifique. Les Cahiers de Propriété Intellectuelle, 8(1), 159-190.

[12] https://books.openedition.org/septentrion/74864?lang=f

[13] Fayçal, K. A. LA TRADUCTION AUTOMATIQUE: état de l'art et les problèmes inhérents (Arabe-Français-Anglais).

[14] Chuquet, H., & Paillard, M. (1987). Approche linguistique des problèmes de traduction anglais-français. Editions Ophrys.

[15] https://actualitte.com

[16] GAJOUI, K., & ALLAH, F. A. Vers un système de reconnaissance optique des caractères dans des documents multilingues: Français-Amazighe

[17] Laïdi, M., & Hanini, S. (2012). Approche neuronale pour l'estimation des transferts thermiques dans un fluide frigoporteur diphasique. Revue des Energies Renouvelables, 15(3), 513-520

[18] Laïdi, M., & Hanini, S. (2012). Approche neuronale pour l'estimation des transferts thermiques dans un fluide frigoporteur diphasique. Revue des Energies Renouvelables, 15(3), 513-520

[19] Touzet, C. (1992). les réseaux de neurones artificiels, introduction au connexionnisme. EC2.

[20] Oukacine, N. (2012). Utilisation des réseaux de neurones pour la reconstitution de défauts en évaluation non destructive (Doctoral dissertation, Université Mouloud Mammeri).

[21] Bouzy, B. (2005). Apprentissage par renforcement (3). Cours de d'apprentissage automatique

[22] Mohamed, M. E. Z. I. D. (2019). Approximation du modèle géométrique inverse d'un robot manipulateur par les réseaux de neurones artificiels (Doctoral dissertation, UNIVERSITE MOHAMED BOUDIAF-M'SILA).

[23] Adankon, M. M. (2009). Apprentissage semi-supervisé pour les SVMS et leurs variantes (Doctoral dissertation, École de technologie supérieure).

[25] (Kammoun, A., & AlouiniFellow, M. S. (2021). On the precise error analysis of support vector machines. IEEE Open Journal of Signal Processing, 2, 99-

[26 Laour, H., & Lakehal, S. (2019). Etude théorique de la Triapine

ملخص

إن إسناد نص مجهول لمؤلف معين هو واحد من أكثر المشاكل المطروحة منذ القدم، وقد حاولنا في هذا العمل البحثي القيام بإسناد نصوص أدبية مترجمة إلى عدة لغات لأصحابها الأصليين. هذه النصوص قد تم الحصول عليها باستعمال برنامج التعرف الضوئي على الحروف (OCR)، وترجمتها آليا إلى الفرنسية والإنجليزية. في هذه الدراسة قمنا بإنشاء قاعدة بيانية جديدة لهذا الغرض كما قمنا باقتراح خوارزميات إحصائية لحل مشكلة التصنيف الأوتوماتيكي للمؤلفين والتعرف على الكتاب الأصليين.

الكلمات المفتاحية : التعرف على الكاتب، الترجمة الآلية، التعرف الضوئي على الحروف.

Résumé

L'attribution d'un texte inconnu à un auteur donner est l'un des problèmes les plus anciens. Dans ce travail de recherche, nous avons essayé d'attribuer des textes littéraires traduis en plusieurs langues à leurs propriétaires respectifs. Ces textes ont été obtenus à l'aide d'un logiciel de reconnaissance optique de caractères (OCR), en suite sa traduction automatiquement en Français et en Anglais. Dans cette étude, nous avons créé une nouvelle base de données à cet effet, et nous avons proposé des algorithmes statistiques pour résoudre le problème de la classification automatique des auteurs et de l'identification des auteurs originaux.

Mots-clés : Attribution d'Auteur, Reconnaissance Optique des Caractères (OCR), traduction automatique.

Abstract

The attribution of an unknown text to a given author is one of the most ancient problems. In this research work, we tried to attribute the translations of unknown literary texts to their respective writers. These texts were obtained using Optical Character Recognition (OCR) software, and then translate them automatically into French and English. In this study, we created a new database for this purpose, and we proposed statistical algorithms to solve the problem of automatic classification of authors and attribution of original authors.

Keywords : Author attribution, French, English, Optical Character Recognition (OCR), automatic translation.