

CHAPITRE 2

2.1. Introduction :

L'étude de la catégorisation automatique des textes remonte au début des années soixante ; puis, son utilisation principale était projetée pour l'indexation des documents scientifiques moyennant un vocabulaire contrôlé. C'était seulement dans les années 1990 que ce champ fut en pleine maturité avec la disponibilité d'un nombre croissant de documents textuels en format numérique et qui nécessite une organisation pour un usage plus facile.

Actuellement, la catégorisation automatique de textes est appliquée dans une variété de domaines, à savoir, le filtrage des spams, la catégorisation de pages Web, la génération automatique des métadonnées, la détection du genre des textes, etc.

2.2. Définition de la classification :

La classification est un processus qui permet d'organiser des données en classes homogènes dont le but est la simplification de la représentation de ces derniers.

La classification donc crée une hiérarchie qui améliore la recherche de documents ; elle génère une vue d'ensemble qui favorise la connaissance de l'environnement ciblé.

La classification d'un document est réalisée à partir d'une évaluation statistique de ses données. La figure 2.1 illustre ce mécanisme.

[Domaines d'information/ unités d'information]

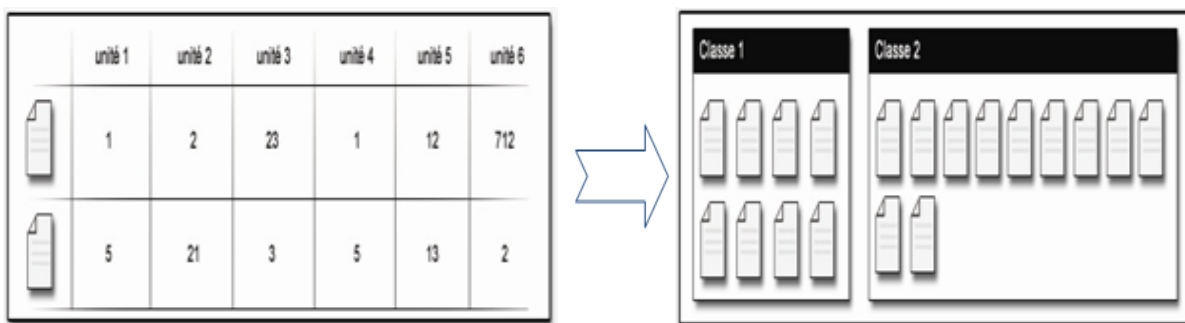


Figure 2.1 : Processus de classification illustré.

Prend en entrée des documents sous format vectoriel \longrightarrow retourne en sortie des documents classifiés.

Explicitement le processus de classification consiste à:

- Déterminer une unité d'information à considérer.
- Prendre en entrée des documents ou des portions de document.
- Comptabiliser les unités d'informations.
- Créer un vecteur de cooccurrence pour chacun des documents.
- Ordonner les vecteurs de manière à retourner en sortie des classes de similarités.

Les documents considérés constituent le domaine d'information tandis que l'union des différentes unités répertoriées forme l'ensemble des unités d'information [11].

2.3. Types de classification :

On distingue dans le domaine de la classification automatique deux types d'approches : la classification supervisée et la classification non supervisée.

Dans le cas de la classification non supervisée, les groupes de documents (classes) sont calculés automatiquement par la machine [12], tandis qu'ils sont, dans l'approche supervisée [13], définis par un expert.

2.3.1. La classification non supervisé : appelée aussi segmentation au clustering, les classes doivent ressortir comme résultat lors l'application du processus de classification.

2.3.2. La classification supervisée : dans la classification supervisée, appelée communément catégorisation les objets sont classés selon des catégories bien définies au préalable.

Notre mémoire concerne la classification supervisé (catégorisation) de document.

2.4. Définition de catégorisation automatique des textes :

Le but de la catégorisation automatique de textes est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu. Habituellement, les catégories font référence aux sujets des textes, mais pour des applications particulières, elles peuvent prendre d'autres formes. En effet, on peut résoudre, par des techniques de catégorisation, des problèmes tels que l'identification de la langue d'un document, le filtrage du courrier électronique pertinent ou indésirable, ou encore la désambiguïsation de termes. Un autre aspect du problème qui varie selon les applications est la présence ou non d'une contrainte concernant le nombre de catégories assignables à un document donné. Il se peut qu'on désire qu'un même texte ne soit associé qu'à une seule catégorie ou bien on peut permettre que plusieurs catégories accueillent un même document.

Aussi, une précision supplémentaire est à faire : dans le cadre de la catégorisation de textes, l'ensemble de catégories possibles est déterminé à l'avance. Il est à noter que le problème consiste à regrouper des documents selon leur similarité [14].

2.5. La catégorisation de documents et l'apprentissage :

Il y a deux approches principales pour la catégorisation des textes.

- La première, est l'approche d'ingénierie cognitive dans laquelle la connaissance d'experts au sujet des catégories est directement encodée dans le système, soit d'une façon déclarative, soit sous forme de règles de classification procédurales.
- L'autre est l'apprentissage automatique (machine learning) dans laquelle un processus inductif général établit un classifieur par apprentissage sur un ensemble d'exemples pré-classifiés.

La majeure partie des travaux récents sur la catégorisation est concentrée sur l'approche d'apprentissage automatique, qui exige seulement un ensemble d'exemples manuellement classifiés qui sont beaucoup moins coûteux pour les produire.

2.6. Catégorisation d'un texte monolingue :

La catégorisation des textes monolingue se rapporte à l'attribution des documents basés sur leurs contenus, à une ou plusieurs catégories prédéfinies.

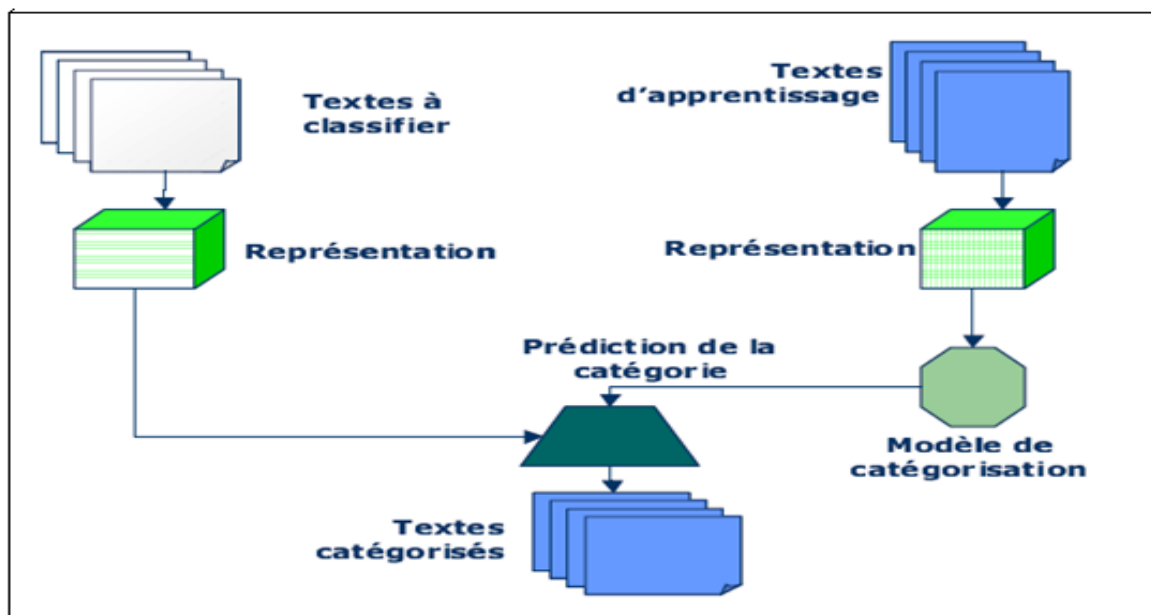


Figure. 2.2 : Schéma générale pour la catégorisation de textes « monolingue » [18].

2.6.1. Représentation du texte sous forme de vecteur:

L'exploitation et le traitement automatique des documents, en particulier pour les tâches de classification, nécessitent une première étape consistant à les représenter.

2.6.1.1. Prétraitement sur le texte :

Cette phase impose les étapes suivantes :

- Extraire le texte du document.
- Formater le texte : (virer la ponctuation, les chiffres et les caractères spéciaux).
- Nettoyage du texte : Filtrer le texte en le passant par les mots vides (stopList) ou les mots inutiles et qui n'ont aucune influence sur le sens (un, une, des ...).
- Normalisation du texte :
 - ✓ Stemmatisation : n'utiliser que les racines des mots.
 - ✓ lemmatisation : pour les verbes on opte pour sa forme infinitive, et pour les noms on prend le nom masculin singulier.

Exemple : Prenons comme exemple le texte suivant:

This award will support the planning activities for the development of the Global Ocean Ecosystems Dynamics Initiative of the NSF Global Geoscience Program and the U.S. Global Change Program for a period of two years. This initiative will address the role of physical ocean processes in the variability of populations of animals in the world's oceans and thereby assess the potential influences of changing global on animal abundance, distribution and production. Support will be used for workshops on technology conceptualization, theory and model international development and coordination of GLOBEC research; and for publications of the initiative.

Segmentation du texte (Tokenization) : Le premier travail est de diviser le texte en token. Ce qui constitue une liste de mots sans ponctuation :

This award will support the planning activities for the development of the Global Ocean Ecosystems Dynamics Initiative of the NSF Global Geoscience Program and the U S Global Change Program for a period of two years This initiative will address the role of physical ocean processes in the variability of populations of animals in the world s oceans and thereby assess the potential influences of changing global on animal abundance distribution and production Support will be used for workshops on technology conceptualization theory and model international development and coordination of GLOBEC research and for publications of the initiative

Elimination des mots outils (Stopwords Removal) : Le second travail consiste à supprimer les token « stopwords » qui sont des termes sans sens sémantique, et donc sans grand intérêt pour la génération de notre matrice :

award support planning activities development Global Ocean Ecosystems Dynamics Initiative NSF Global Geoscience Program U S Global Change Program period years initiative address role physical ocean processes variability populations animals world s oceans thereby assess potential influences changing global climate animal abundance distribution production Support workshops technology conceptualization theory model international development coordination GLOBEC research publications initiative

Stemmer : Afin de diminuer le nombre de token différents, il est intéressant de les groupés par racine lexicale. Les items au pluriel, au féminin, etc. sont ramenés à un état plus simple, permettant

de diminuer la variété et d'augmenter la fréquence. Le *stemmer* utilisé à également normaliser les lettres en minuscules.

award support plan activ develop global ocean ecosystem dynam initi nsf global geoscienc program u s global chang program period year initi
address role physic ocean process variabl popul anim world s ocean therebi assess potenti influenc chang global climat anim abund distribut
product support workshop technolog conceptu theori model intern develop coordin globec research public initi

2.6.1.2. Choix de termes :

2.6.1.2.1. Représentation en sac de mots :

C'est la plus simple représentation des textes introduite dans le cadre du modèle vectoriel, consiste à utiliser les mots sous leur forme originale comme descripteur pour ensuite représenter les textes sous forme vectorielle. Un prétraitement important peut consister à filtrer certains mots. Par exemple, des mots outils ou "stop words" (mots fonctionnels tels que les prépositions, articles, etc.) peuvent être supprimés. Le but est alors de seulement conserver les mots ayant une signification représentative du texte.

2.6.1.2.2. Représentation par les N-grammes de mots (phrases) :

Il existe beaucoup de mots ayant la même forme, mais des sens différents. Par exemple, "prix" n'a pas le même sens dans "prix Goncourt", "grand prix" ou "prix marchandise". Ces mots augmentent l'ambiguïté du sens des textes (classification erronée). En utilisant les N-grammes de mots (N mots consécutif), un sens parmi d'autres est favorisé. Par exemple, "actes biologie" (issu des données d'ITESOFT) est un bigramme de mots particulièrement pertinent.

Nous donnons ci-dessous des exemples (issus des données d'ITESOFT) de N-grammes de mots :

- N=1 (unigramme) : "biologie", "médicale", "frais", "accessoires" et "maladie".
- N=2 (bigrammes) : "biologie médicale", "frais maladies" et "malade no".
- N=3 (trigrammes) : "malade no facture", "prescription actes renseignements"[17].

En 2005 Paradis et Nie appliquent une telle représentation afin d'effectuer une classification de documents. La méthode consiste à classer les documents bruités en se fondant sur le filtrage du contenu avec les N-grammes de mots et les entités nommées sur des documents de type "appels d'offres" [15]. Les travaux de Tan et al en 2002 utilisent des bigrammes ou des unigrammes de mots comme descripteurs pour la représentation des données pour une tâche de classification [16]. Les résultats révèlent que l'utilisation des bigrammes sur ces données améliore les résultats de manière significative.

Notons que l'exemple donné précédemment montre clairement que l'utilisation de N-grammes de mots favorise un sens parmi d'autres. Cependant, en se fondant sur le groupement de mots, nous

pourrions dans certains cas dégrader la classification en introduisant une quantité supplémentaire de bruit. Par exemple dans le cas d'utilisation d'un trigramme de mots, les trois mots le composant vont être éloignés du sens de chacun des mots [17].

2.6.1.2.3. Représentation avec les N-grammes de caractères:

Dans cette méthode la représentation du document se fait par le N-gramme de caractères qui est une séquence de N caractères issus d'une chaîne de caractères.

Cette technique ayant plusieurs avantages comme [17]:

- Les N-grammes permettent de capturer automatiquement la racine des mots les plus fréquents. Il n'est pas nécessaire d'appliquer une étape de recherche de racine et/ou de lemmatisation.
- Ces descripteurs sont indépendants de la langue employée dans le corpus. Il n'est pas nécessaire d'utiliser des dictionnaires, ni de segmenter les documents en mots.
- Les N-grammes sont tolérants aux fautes d'orthographe et aux déformations causées lors de la reconnaissance de documents (système OCR (Reconnaissance Optique de Caractères)). 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.

2.6.1.2.4. Représentation avec les lemmes et stemms :

La lemmatisation consiste à représenter chaque mot par sa forme canonique. Ainsi, les verbes par leur forme infinitive et les noms par leur forme au singulier sont pris en compte. L'intérêt principal de la lemmatisation est la substitution des mots par leur racine ou leur lemme.

Le processus permet une réduction du nombre de descripteurs. Par exemple, le remplacement des mots : conductrice, conducteur par l'unique racine conducteur semble être avantageux tout comme le remplacement des formes conjuguées franchit et franchi par le lemme franchir.

La stemmatisation (radicalisation) consiste à supprimer tous les affixes d'un mot. Par affixes on entend : suffixe (défin-ition), préfixe (sur-consommation). Ces techniques sont principalement utilisées pour réduire l'espace de représentation des documents et faire ressortir les traits similaires entre les mots [17].

De nombreux travaux expérimentent ces techniques afin d'améliorer les performances de classification. En effet, avec la lemmatisation nous risquons de perdre des informations cruciales car les contextes des mots en forme singulière et plurielle (par exemple, les mots "président" et "présidents") peuvent se révéler différents suggérant des concepts distincts [17].

2.6.1.2.5. Représentation conceptuelle :

Cette technique consiste à représenter le document sous forme d'un ensemble de concepts, ces concepts peuvent être capturés en utilisant les réseaux sémantiques ou les sous arbres (un sous arbre représente une hiérarchie de concepts).

Cette méthode a comme avantage selon REHEL [14] de réduire l'espace de travail car les mots qui sont synonymes partagent au moins un concept. Cependant, l'inconvénient majeur de cette représentation est qu'il n'existe pas des bases lexicales pour toutes les langues.

2.6.1.3. Traitement numérique :

Nous présentons dans cette section divers traitements possibles afin d'obtenir une représentation quantifiable d'un terme [17].

2.6.1.3.1. Représentation fréquentielle :

Un traitement numérique simple et intuitif ("document frequency") consiste à calculer la fréquence des descripteurs dans chaque document. Une formule qui calcule le poids du mot t dans le document D est donnée ci-dessous :

$$\mathbf{TF(t, D) = \text{fréquence du descripteur } t \text{ dans le document } D.}$$

- TF : la fréquence d'apparition du mot dans le texte (fréquence locale) calculé selon la formule suivante :

$$TF(t, d) = f(t, d) / \text{Max}[f(t, d)] \quad (1)$$

Avec : $f(t, d)$: fréquence du terme t dans le document d (nombre d'apparition de t dans d).

Les travaux décrits ci-dessous proposent une mesure différente du TF afin d'attribuer un poids aux descripteurs.

2.6.1.3.2. Représentation avec TF.IDF :

Une autre mesure de poids connue sous le nom TF.IDF (Term Frequency Inverse Document Frequency). Elle permet de mesurer l'importance d'un mot en fonction de sa fréquence dans le document (TF = Term Frequency) pondérée par la fréquence d'apparition du terme dans tout le corpus (IDF = Inverse Document Frequency). Cette mesure permet de donner un poids plus important aux mots discriminants d'un document. Ainsi, un terme apparaissant dans tous les documents du corpus aura un poids faible.

Le poids d'un descripteur t_k dans un document D_j est calculé ainsi :

$$TF.IDF(t_k, D_j) = freq(t_k, D_j) * \log(N/docfreq(t_k)) \quad (2)$$

Avec : N est le nombre de documents dans le corpus (taille du corpus)

2.6.2. Choix de l'algorithme d'apprentissage :

Plusieurs algorithmes d'apprentissage peuvent être utilisés dans le domaine de la catégorisation des documents, notamment :

- Algorithme des k plus proches voisins kppv (en anglais K-NN)
- Algorithme de Naïve de Bayes
- Les SVM (Support Vector machines)
- Arbres de décision.
- Réseaux de neurones.

Dans le prochain chapitre, en va expliquer le fonctionnement de ces algorithmes.

2.6.3. Application de l'algorithme choisi :

2.6.3.1. Phase d'apprentissage :

- Prendre un échantillon de documents (corpus d'apprentissage).
- Déterminer l'ensemble de catégories.
- Affecter chaque document du corpus d'apprentissage à l'une des catégories prédéfinies.

2.6.3.2. Phase de classification :

- Indexation des documents (le corpus d'apprentissage) par des mots clés en les représentant par des vecteurs ou le poids représente la fréquence de chaque mot clé dans le document.
- Classer le nouveau texte en calculant la distance entre sa représentation vectorielle et celle de la base d'apprentissage.
- Appliquer l'algorithme d'apprentissage choisi.

2.7. Applications de catégorisation des textes :

La catégorisation de textes est utilisée dans de nombreuses applications. Parmi ces domaines figurent : l'identification de la langue, la reconnaissance d'écrivains, la catégorisation de documents multimédia, et bien d'autres.

La catégorisation de textes peut être un support pour différentes applications parmi lesquelles le filtrage, qui consiste à déterminer si un document est pertinent ou non (décision binaire), par exemple la détection de spam (les courriers électroniques indésirables) pour ensuite les supprimer, le routage qui permet d'affecter un document à une ou plusieurs catégories parmi les, par exemple la diffusion sélective d'information. Lors de la réception d'un document l'outil choisit à quelles personnes le faire parvenir en fonction de leurs centres d'intérêt. Ces centres d'intérêt correspondent à des profils individuels [18].

2.8. Catégorisation des textes multilingues :

2.8.1. Définition :

La catégorisation de textes multilingues consiste à catégoriser un texte rédigé dans une langue donnée, à partir d'un modèle de prédiction construit sur une base d'apprentissage dans une ou plusieurs langue cible [03].

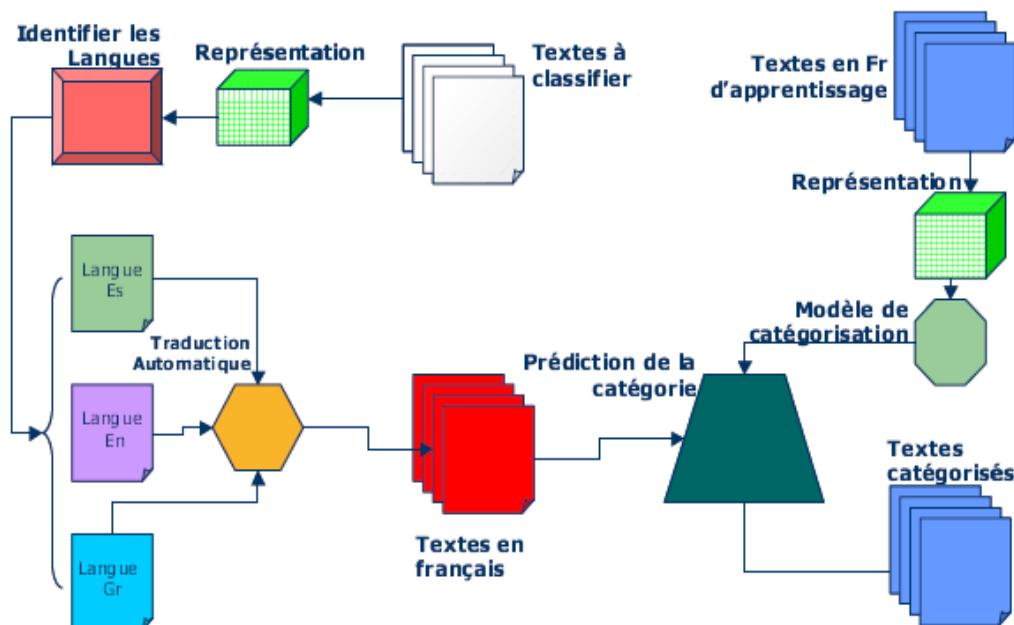


Figure 2.3 : Catégorisation de textes multilingue [18].

2.8.2. Les types de catégorisation des textes multilingues :

La catégorisation des textes multilingue peut être traitée selon différents types.

2.8.2.1. Catégorisation des textes par croisement de langues :

Dans La catégorisation des textes par croisement de langues, dite en anglais Cross-Language Text Categorization (CLTC), un ensemble de documents étiquetés est disponible dans une seule langue. Cet ensemble est utilisé pour catégoriser des documents non étiquetés exprimés dans une autre langue. Pour cela, deux manières différentes de traduction peuvent être employées [03].

2.8.2.1.1. Traduction des documents étiquetés : les documents étiquetés sont traduits dans la langue des documents non étiquetés afin d'être utilisé pour catégoriser ces derniers [03].

2.8.2.1.2. Traduction des documents à classer : Dans ce cas, c'est les documents non étiquetés qui sont traduit vers la langue des documents étiquetés. Le classificateur est donc construit en utilisant des documents non traduit [03].

2.8.2.2. Catégorisation des textes par multiples langues :

Dans ce cas, le classificateur est construit en utilisant un ensemble de documents étiquetés dans plusieurs langues afin de catégoriser des documents de différentes langues. Ce scénario exclu l'utilisation des stratégies de traduction donc, aucune perte l'information n'est faite [03].

2.8.2.3. Catégorisation des textes avec la langue universelle :

Ce scénario utilise une langue de référence universelle à laquelle tous les documents sont traduits. Cette langue devrait contenir toutes les propriétés des langues et doit être organisée d'une façon sémantique : les mots indiquant les mêmes concepts dans les langues devraient être traduits aux mêmes termes dans la langue universelle [19].

2.8.3. Identification de la langue :

Il est important de détecter avec précision la langue dans laquelle le texte à classer est rédigé, car une erreur à ce niveau voue à l'échec les étapes suivantes. Cette identification consiste à attribuer une unité textuelle, supposée monolingue, à une langue.

Il existe deux familles d'approches dans l'identification de la langue : linguistique ou statistique.

2.8.3.1. Approche linguistique : nécessite des connaissances linguistiques préalables, qui seront intégrées dans le programme informatique, par exemple la présence de certaines chaînes de caractères spécifiques et de certains mots [03].

2.8.3.2. Approche statistique : utilise des ressources construites automatiquement à partir d'un corpus textuel représentatif de la langue qui à pour objectif de capturer au moyen de modèles statistiques ou probabilistes par exemples les mots les plus fréquents, et les séquences de n-grammes les plus fréquentes [03].

2.8.4. Traduction automatique :

L'objectif de la traduction automatique (TA) du texte à classifier dans la langue du corpus d'apprentissage est de fournir un texte assurant une qualité de classement suffisante. Il est évident que le résultat obtenu dépendra du traducteur utilisé.

La traduction automatique propose des aspects très intéressants, en particulier l'espoir que l'ambiguïté sera moins prononcée dans les textes relativement longs. En effet, elle pourrait bénéficier de l'information contextuelle [03].

Un autre avantage est que les utilisateurs peuvent immédiatement recevoir les documents en leur langue préférée, ce qui leurs permet de les consulter directement. [20]

La TA consiste à saisir un texte puis le soumettre au traitement automatique et enfin récupérer en sortie une traduction brute sans intervention humaine.

Il existe dans [21] trois types de TA qui sont :

2.8.4.1. Traduction mot à mot : c'est une traduction directe qui se fait par des analyses linguistiques superficielles et sans compréhension, qui a été utilisée par les premiers traducteurs automatiques, c'est une méthode simple qui est réussie dans des domaines limitées, mais elle est utilisée seulement pour un couple de langue et pas d'analyses profondes.

2.8.4.2. Traduction par transfert : cette méthode commence par une représentation de la langue source qui se fait par une analyse grammaticale et dictionnaire de la langue source et qui se termine par la représentation de la langue cible qui se fait à leur tour par une synthèse grammaticale et dictionnaire grâce à des règles de transfert à partir d'un dictionnaire bilingue. C'est une méthode compliquée comporte une analyse linguistique importante qui affecte plusieurs niveaux linguistiques mais qui est une traduction unidirectionnelle.

2.8.4.3. Traduction par pivot : c'est une traduction multi-langue et par contre bidirectionnelle, qui utilise un langage pivot qui sert à la représentation sémantique qui fait l'abstraction des sens, qui peut lier des informations contextuelles et extralinguistiques et que cette méthode est utilisé dans plusieurs types d'applications.

2.8.5. Les difficultés particulières de la catégorisation des textes multilingue :

Les deux catégorisations des textes monolingue et multilingue sont confrontées aux mêmes problèmes qui sont les difficultés du langage naturel (polysémie, la redondance, l'ambiguïté et l'implicite). Ajoute à ces difficultés, la reconnaissance de la langue si celle-ci n'est pas connue. Et si au contraire la langue d'un texte est reconnue, il faut identifier les mots spécifiques utilisés. Dans beaucoup de langues, l'opération est facile parce que les mots sont séparés par des espaces, mais dans d'autres langues, les mots sont concaténés pour former de nouveaux mots. Dans les cas les plus difficiles, il n'y a pas d'espace entre les mots par exemple la langue japonaise ou chinoise. Or, la composante multilingue ajoute une complexité supplémentaire au processus de catégorisation qui est la traduction automatique. [03]

2.9. Conclusion :

Dans ce chapitre nous avons présenté le processus de la catégorisation des textes avec ses différentes phases, les notions importantes et quelques applications de la catégorisation des textes.

L'application des algorithmes d'apprentissage aux données textuelles introduit des difficultés supplémentaires. Nous avons cité : la redondance, l'ambiguïté, l'implicite et le sur-apprentissage. Dans le chapitre suivant nous présentons le processus de catégorisation des textes multilingues et ses types avec les deux étapes supplémentaires pour l'apprentissage et/ou le classement des textes par rapport à la catégorisation des textes monolingues.