Table des matières

Liste des Figures

Liste des Tableaux

Introduction Générale

La communication est un processus nécessaire pour l'être humain. Cette communication peut être orale (parole) ou écrite. Actuellement, l'information peut avoir comme support trois média de base : le texte, le son et l'image. Notre travail se focalisera sur les textes, ainsi le terme « document » induit directement qu'il s'agit d'un document textuel.

Vu l'apparition d'internet et le nombre important de collections de documents multilingues, il est devenu indispensable au utilisateur du web de trouver les documents pertinents, quelles qu'en soient leurs langues. Ce qui a donné naissance à un nouveau domaine qui est le domaine de catégorisation des textes multilingue.

Ce mémoire traite la représentation conceptuelle pour la catégorisation des textes multilingue, qui consiste d'abord à représenter les documents des deux corpus, l'étiqueté rédigé dans la langue L1 et non l'étiqueté décrit en langue L2 avec une bonne méthode de représentation. L'objectif principal est de comparer la représentation « sac de mots » avec la représentation « conceptuelle » dans un processus de catégorisation des textes multilingues.

La structure de la suite de ce mémoire est comme suit : en premier lieu, le chapitre I vise à présenter le processus de la catégorisation des textes et les principales méthodes d'apprentissage ayant fait leurs preuves dans ce domaine, aussi, les difficultés liées à la catégorisation des textes. Puis, le chapitre II expose les types de la catégorisation des textes multilingue, et un état de l'art dans ce domaine. Le chapitre III est dédié à la description des approches implémentées ainsi que les résultats obtenus.

Chapitre I : Catégorisation des Textes

I. Introduction

Comme déjà cité dans l'introduction générale, internet a fournit à ses utilisateurs une base gigantesque de documents textuels. Afin de trouver les documents pertinents dans un temps raisonnable, il est nécessaire d'avoir des solutions pour la recherche de tels documents. Une des solutions est la catégorisation des textes.

Ce chapitre présente la définition de la catégorisation des textes ainsi que son processus, il expose aussi quelques applications de la catégorisation des textes, les méthodes d'apprentissage utilisées dans ce domaine et les difficultés qui le caractérise.

II. Définition

F. le processus qui consiste à associer une valeur booléenne à chaque paire (dj, ci) ∈ D × C, où D est l'ensemble des textes et C est l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple (dj, ci) si le texte dj appartient à la classe ci tandis que la valeur F (Faux) lui sera SEBASTIANI définit Formellement dans **[Sebastiani, 2002]** la catégorisation des textes comme associée dans le cas contraire. Le but de la catégorisation des textes est de construire une procédure (modèle, classificateur) notée :

Φ: D × C → {V, F} qui associe une ou plusieurs étiquettes (catégories) à un document dj telle que la décision donnée par cette procédure coïncide le plus possible avec la fonction $\widehat{\Phi}$: D × C → {V, F}, la vraie fonction qui retourne pour chaque vecteur dj une valeur ci.

III. Processus de la catégorisation des textes

Le processus reçoit en entrée un document textuel afin de lui trouver sa catégorie, pour cela plusieurs étapes doivent d'être suivies. D'après **[Jalam, 2003]**, ces étapes sont :

- La représentation des textes
- La Pondération des termes
- La réduction de la taille du vocabulaire
- Choix de classificateur
- Evaluation du modèle

La figure 1.1 résume le processus de catégorisation des textes qui comporte deux phases : l'apprentissage et le classement.

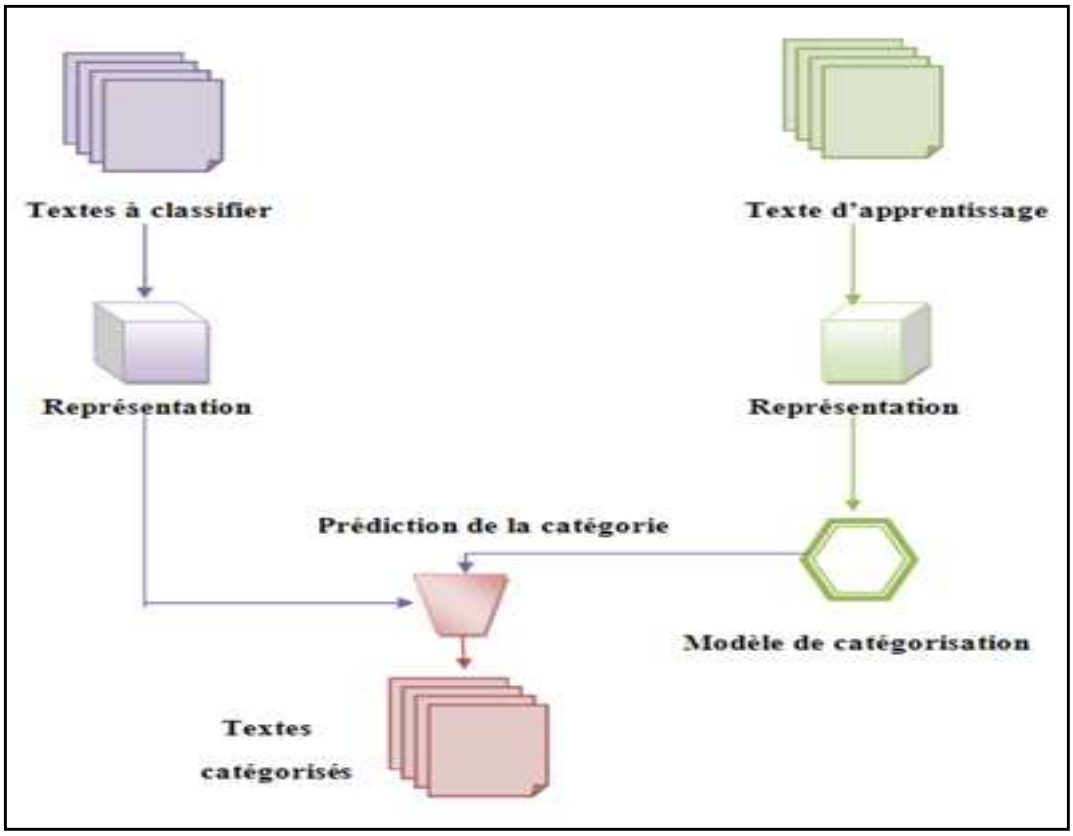

Figure 1.1 : Processus de la catégorisation des textes

[Jalam, 2003]

III.1. Représentation de texte

Afin de bien classifier les textes il est nécessaire d'utiliser une technique de représentation efficace. Les différentes méthodes qui existent pour la représentation des textes sont :

III.1.1. Représentation en sac de mots (bag of words)

Cette méthode consiste à représenter le document sous forme d'un vecteur de mots. Le processus qui permet de convertir le texte d'un document en un ensemble de termes est appelé l'analyse lexicale qui permet de reconnaître les espaces de séparation des mots, les ponctuations, les chiffres,…etc., pour qu'ils seront tous supprimés de la représentation. Cette représentation a comme avantage d'exclure toute analyse grammaticale et toute notion de distance entre les mots, mais présente comme inconvénient la difficulté de délimiter les mots dans certaines langues telles que l'Arabe ou l'Allemand.

III.1.2. Représentation avec les racines lexicales

Cette méthode consiste à remplacer les mots du document par leurs racines lexicales, qui peut être réalisée en utilisant un des algorithmes les plus connus pour la langue anglaise qui est l'algorithme de Porter **[Porter, 1980]** de normalisation de mots qui sert à supprimer les affixes de ces derniers pour obtenir une forme canonique. Cette méthode a comme avantage de regrouper les différentes flexions d'un mot dans une seule composante, et comme inconvénient la perte de sens car la racine extraite peut être commune à des mots se rapportant à des concepts différents.

A titre d'exemple : les mots vol, volant, vole ont la même racine vol mais se rendent à trois notions différentes.

III.1.3. Représentation avec les lemmes

Cette méthode consiste à remplacer les mots du document par leurs lemmes, elle doit utiliser l'analyse grammaticale afin de remplacer les verbes par leurs formes infinitives et les noms par leurs formes au singulier. En effet, Un mot donné peut avoir différentes formes dans un texte, mais leur sens reste le même. Par exemple, les mots vol, volant et vole seront remplacés par leurs lemmes : vol, volant et voler selon le contexte. Cette représentation est simple mais elle peut causer une perte d'informations donnée par le contexte nécessaire à la distinction des lemmes polysémiques (possèdent plusieurs sens) et la présence de synonymes, considérés comme des lemmes différents même s'ils font référence au même concept. **[Ignat, 2007]**

III.1.4. Représentation avec les n-grammes

Cette méthode consiste à représenter le document par des n-grammes. Le n-gramme est une séquence de n caractères consécutifs. Cette technique présente plusieurs avantages. Les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales, indépendante de la langue, les espaces sont pris en considération parce qu'en effet, la non prise en compte de ces derniers introduit du bruit.

III.1.5. Représentation conceptuelle

Cette méthode 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 dans **[Réhel, 2005]** 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.

III.2. Pondération des termes

La pondération des termes permet de mesurer l'importance d'un terme dans un document. Cette importance est souvent calculée à partir de considérations et interprétations statistiques (ou parfois linguistiques). L'objectif est de trouver les termes qui représentent le mieux le contenu d'un document. Les méthodes les plus populaires sont :

III.2.1. Mesure TF (Term Frequency)

Cette mesure est proportionnelle à la fréquence du terme dans le document (pondération locale). Elle peut être utilisée telle quelle ou selon plusieurs déclinaisons (log (TF), présence/absence,. . .).

III.2.2. Mesure TFIDF (Term Frequency Inverse Document Frequency)

Le poids d'un terme *T* dans un document *D* est calculé comme suit :

$$\mathbf{TFIDF(T, D)} = TF(T, D) * \log (N/ DF(T))$$

Avec :

TF(T, D) : la fréquence du terme dans le document,

N : le nombre total de documents de la base documentaire et

DF(T) : le nombre de documents contenant le terme.

III.3. Réduction de la taille du vocabulaire

Vu la taille impressionnante des bases textuelles, il est difficile de prendre l'ensemble de tous les mots comme étant des attributs, en effet cela engendre une perte de mémoire et de temps de calcul.

Plusieurs techniques de réduction existent pour réduire la dimension de vocabulaire qui se divise en deux grandes familles :

- *Sélection d'attributs* : (*feature selection*) prend les attributs (ou mots) d'origine et conserve seulement ceux jugés utiles à la catégorisation, selon une certaine fonction d'évaluation et les autres sont rejetés.
- *Extraction d'attributs* : (*feature extraction*) à partir des attributs de départ, elles créent de nouveaux attributs en faisant soit des regroupements ou des transformations.

III.4. Choix de classificateur

La catégorisation des textes comporte un choix de technique d'apprentissage (classificateur). Parmi les méthodes d'apprentissage les plus utilisées figurent :

❖ **Classificateur bayé-sien naïf**

Comme son nom l'indique, ce classificateur se base sur le théorème de Bayes permettant de calculer les probabilités conditionnelles. Son utilisation lorsqu'il est appliqué à la classification de textes est résumé comme suit: on cherche la classification qui maximise la probabilité d'observer les mots du document. Lors de la phase d'entraînement, le classificateur calcule les probabilités qu'un nouveau document appartienne à telle catégorie à partir de la proportion des documents d'entraînement appartenant à cette catégorie. Il calcule aussi la probabilité qu'un mot donné soit présent dans un texte, sachant que ce texte appartient à telle catégorie. Quand un nouveau document doit être classé, on calcule les probabilités qu'il appartienne à chacune des catégories à l'aide de la règle de Bayes. **[Réhel, 2005]**

La probabilité à estimer est donc : P (cj | a_1, a_2, a_3, ..., a_n) où : c_j est une catégorie et a_i est un descripteur. A l'aide du théorème de Bayes, on obtient :

$$\mathrm{P}(\mathrm{cj}/ai) = \frac{\mathrm{P}(a1, a2, a3, \dots, an/cj)\mathrm{P}(\mathrm{cj})}{P(a1, a2, a3, \dots, an)}$$

Ce classificateur a comme avantage: possibilité en ligne et comme inconvénient: lorsque le modèle est mal spécifié, on aura intérêt à utiliser une méthode discriminative. **[Obozinski, 2010]**

❖ **K-plus proches voisins**

K-plus proches voisins « K-Nearest Neighbour » est une méthode très connue dans le domaine de la catégorisation des textes. Ses performances la situent parmi les meilleures méthodes de catégorisation. L'idée de K-plus proches voisins est de représenter chaque texte dans un espace vectoriel, dont chacun des axes représente un élément textuel (peut être un mot sous sa forme brute ou sous une forme lemmatisée).

K-plus proches voisins est un algorithme de catégorisation dans lequel les classes ne sont pas représentées sous forme de texte. Chaque nouveau texte à traiter sera comparé à l'ensemble des textes du jeu d'apprentissage afin de trouver la catégorie qui lui est la plus proche, soit en moyenne celle qui contient le plus de textes voisins. **[Jaillet & al, 2005]**

L'algorithme de catégorisation de K-plus proches voisins pris de **[Jalam, 2003]**, est le suivant :

Paramètre : le nombre K de voisins

Contexte : un échantillon de L textes classés en C = c1, c2, ..., cn classes

Début

    Pour chaque texte T faire

        Transformer le texte T en vecteur T = (x1, x2, ..., xm),

        Déterminer les K plus proches textes du texte T selon une métrique de distance,

        Combiner les classes de ces K exemples en une classe C.

    Fin pour

Fin

Sortie : le texte T associé à la classe C.

Le choix du paramètre K est primordial pour le bon fonctionnement de cette méthode. Une grande base d'apprentissage permet une plus grande valeur de K, et un K petit est nécessaire pour des petites bases d'apprentissage. **[Jaillet & al, 2005]**

La distance entre un texte et ses voisins se fait via une métrique de distance. Cette métrique peut être comme suit :

- **Mesure Cosinus** qui consiste à calculer le produit scalaire entre deux vecteurs *a* et *b*, que nous divisons par le produit de la norme de ces deux vecteurs. La formule de la mesure Cosinus est alors la suivante :

$$\textbf{Cosinus (a, b)} = \frac{\sum (a \times b)}{\sqrt{\sum a^2 \times \sum b^2}}$$

D'autres mesures ont été proposées dans la littérature, parmi lesquelles on peut citer les mesures de Jaccard et Dice.

- **Mesure de Jaccard**

  La formule de la mesure de Jaccard est alors la suivante :

$$\textbf{J (a, b)} = \frac{\sum (a \times b)}{\sum a^2 + \sum b^2 - \sum a\, b}$$

- **Mesure de Dice**

  La formule de la mesure de Dice est alors la suivante :

$$\textbf{D (a, b)} = 2 \times \frac{\sum (a \times b)}{\sum (a^2 + b^2)}$$

❖ **Méthode de Rocchio**

Cette méthode se base sur la création de profils de catégorie. Le poids des termes est calculé lors de l'apprentissage en fonction des apparitions de ces termes d'une part dans les documents appartenant à la catégorie et d'autre part dans ceux n'y appartenant pas. Le poids X du terme J du profil de la catégorie g est calculé ainsi :

$$Xj = \beta \frac{\sum_{i\, \epsilon bien\ classé} Xi,j}{Ng} + \alpha \frac{\sum_{k \epsilon mal\ classé} Xk,j}{N-Ng}$$

Où : N: le nombre de documents de la collection, Ng: le nombre de documents pré-catégorisés dans la catégorie *g* et Xm,j: le poids du terme j dans le document m. Les paramètres β et α sont fixés à 16 et 4 d'après **[Caropreso, 2001]** et **[Hernandez, 1999]**

Cette méthode est facile à implanter et efficace pour des catégorisations ou un texte ne peut appartenir qu'à une seule catégorie. Mais elle n'est pas très efficace quand un texte peut appartenir à plusieurs catégories et certains documents du corpus d'apprentissage appartenant à une catégorie Ci initialement ne seraient pas classé dans Ci par le classificateur.

❖ **Arbres de décision**

Les arbres de décision sont composés d'une structure hiérarchique en forme d'arbre. Un arbre de décision est un graphe orienté sans cycles, dont les nœuds portent une question, les arcs des réponses et les feuilles des conclusions ou des classes terminales. Un classificateur de texte basé sur la méthode d'arbre de décision est un arbre de nœuds internes qui sont marqués par des termes, les branches qui sortent des nœuds sont des tests sur les termes et les feuilles sont marquées par catégories. **[Abidi, 2011]**

Une méthode pour effectuer l'apprentissage d'un arbre de décision pour une catégorie Ci consiste à vérifier si tous les exemples d'apprentissage ont la même étiquette. Dans le cas contraire, nous sélectionnons un terme Tk, et nous partitionnons l'ensemble d'apprentissage en classes de documents qui ont la même valeur pour Tk, et à la fin on crée les sous-arbres pour chacune de ces classes.

Ce processus est répété récursivement sur les sous-arbres jusqu'à ce que chaque feuille de l'arbre généré de cette façon contienne des exemples d'apprentissage attribués à la même catégorie Ci, qui est alors choisie comme l'étiquette de la feuille. L'étape la plus importante est le choix du terme de pour effectuer la partition.

Toutefois, une telle méthode de construction d'arbre peut faire l'objet de sur-apprentissage, comme certaines branches peuvent être trop spécifiques pour les données d'apprentissage. La plupart des méthodes d'apprentissage des arbres incluent une méthode pour la construction d'arbre. **[Dziczkowski, 2008]**

❖ **Machine à support vectoriel**

Les machines à support vectoriel (Support Vector Machines ou SVM) forment une classe d'algorithmes d'apprentissage qui peuvent s'appliquer à tout problème qui implique un phénomène *F* et qui à partir d'un jeu d'entrées *X*, produit une sortie *Y= F(X). Le* but est de retrouver *F* à partir de l'observation d'un certain nombre de couples entrée/sortie.

Le problème revient à trouver une frontière de décision qui sépare l'espace en deux régions, à trouver l'hyperplan qui classifie correctement les données et qui se trouve le plus loin possible de tous les exemples. On dit qu'on veut maximiser la marge qui veut dire la distance du point le plus proche de l'hyperplan.

Dans le cas de la catégorisation des textes, les entrées sont des documents et les sorties sont des catégories. En considérant un classificateur binaire, on voudra lui faire apprendre l'hyperplan qui sépare les documents appartenant à la catégorie et ceux qui n'en font pas partie.

Les SVM conviennent bien pour la classification de textes parce qu'une dimension élevée ne les affecte pas puisqu'ils se protègent contre le sur-apprentissage. Autrement dit, il affirme que peu d'attributs sont totalement inutiles à la tâche de classification et que les SVM permettent d'éviter une sélection agressive qui aurait comme résultat une perte d'information. On peut se permettre de conserver plus d'attributs. Également, une caractéristique des documents textuels est que lorsqu'ils sont représentés par des vecteurs, une majorité des entrées sont nulles.

Or, les SVM conviennent bien à des vecteurs dits clairsemés. Un autre aspect positif des SVM est qu'aucun ajustement de paramètres manuel n'est requis, car ils ont l'habileté de trouver automatiquement des paramètres adéquats. **[Réhel, 2005]**

- **Réseaux de neurones**

Les réseaux de neurones (Artificial Neural Network) sont généralement optimisés par des méthodes d'apprentissage de type statistique grâce à leur capacité de classification et de généralisation, tels que la classification automatique de codes postaux ou la prise de décision concernant un achat boursier.

Un réseau de neurone est en général composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Chaque couche i est composée de Ni neurones, prenant leurs entrées sur les Ni-1 neurones de la couche précédente. À chaque synapse est associé un poids synaptique, de sorte que les Ni-1 sont multipliés par ce poids, puis additionnés par les neurones de niveau i, ce qui est équivalent à multiplier le vecteur d'entrée par une matrice de transformation. Mettre l'une derrière l'autre, les différentes couches d'un réseau de neurones revient à mettre en cascade plusieurs matrices de transformation et pourrait se ramener à une seule matrice produit des autres, s'il n'y avait à chaque couche, la fonction de sortie qui introduit un non linéarité à chaque étape.

Ceci montre l'importance du choix judicieux d'une bonne fonction de sortie : un réseau de neurones dont les sorties seraient linéaires n'aurait aucun intérêt. **[Zeggane Mokhtar, 2009]**

❖ **Méthode AdaBoost**

Une méthode envisagée pour classifier automatiquement des textes est d'interroger différents classificateurs et de combiner leurs décisions de classification en pondérant chaque classificateur selon sa performance testée sur des exemples de validation semblables aux documents en question.

Il est clair qu'une telle approche est très coûteuse sur le plan informatique étant donné qu'il faut entraîner et faire fonctionner les différents classificateurs puis combiner leurs résultats. Un algorithme AdaBoost permet d'obtenir des résultats intéressants. Son fonctionnement général revient à entraîner un comité de classificateurs en passant par plusieurs itérations, en donnant à chaque étape plus de poids aux exemples incorrectement classifiés à l'itération précédente et en retenant pour les étapes suivantes les classificateurs les plus prometteurs. **[Réhel, 2005]**

Le tableau suivant résume les méthodes d'apprentissage les plus connus, introduit par F.SEBASTIANI dans **[Sebastiani, 2002]**.

Tableau I.1 : Méthodes d'apprentissage de catégorisation de texte [Sebastiani, 2002]

Méthode d'apprentissage	*Type*	*Références*
Word	Non cité	[Yang, 1999]
Drop Bayes Bim Nb	Probabiliste	[Dumais & al, 1998] [Li & Yamanishi, 1999] [Yang & Liu, 1999]
C4.5 Ind	Arbres de décision	[Dumais & al, 1998] [Lewis & Ringuette, 1994]
Swap-1 Ripper Sleeping Experts DL-Esc Charade	Règles de décision	[Apté & al, 1994] [Cohen & Singer, 1999] [Cohen & Singer, 1999] [Li & Yamanishi, 1999] [Moulinier & al, 1996]
LLSF	Régression	[Yang & Liu, 1999]
Balanced Winnow Widrow-HoFF	On line linear	[Dagan & al, 1997] [Lam & HO, 1998]
Rocchio FindSim	Batch linear	[Cohen & Singer, 1999] [Dumais & al, 1998]
CLASSI Nnet	Réseau de neurone	[Ng & al, 1997] [Yang & Liu, 1999]
Gis-W K-NN	Example based	[Lam & Ho, 1998] [Yang & Liu, 1999]
SVMLight	SVM	[Yang & Liu, 1999]
ADABoost	Comité	[Schapire & Singer, 2000]
	Bayé-sien naif	[Dumais & al, 1998]

III.5. Evaluation du modèle

Pour évaluer tout processus de catégorisation, il est nécessaire d'appliquer une méthode d'évaluation. La performance de la catégorisation de textes est souvent mesurée via la précision et le rappel. La précision est défini comme la probabilité conditionnelle et le rappel mesure la largeur de l'apprentissage et correspond à la fraction des documents pertinents, parmi ceux proposés par le classificateur. **[Jalam, 2003]**

$$\textbf{Précision} = A / (A + B) \quad \text{pour } A + B > 0$$

$$\textbf{Rappel} = A / (A + C) \quad \text{pour } A + C > 0$$

Avec :

A : le nombre de documents correctement attribués à la catégorie.

B : le nombre de documents incorrectement attribués à la catégorie.

C : le nombre de documents qui auraient dû lui être attribués mais qui ne l'ont pas été.

Le rappel et la précision donnent deux points de vue différents sur les résultats d'un test. La F-mesure (F-score) a été introduite par Van RIJSBERGEN en 1979, pour combiner les deux mesures en une seule.

$$\textbf{F.mesure}\ (\boldsymbol{\beta}) = \frac{(\beta^2 + 1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

Lorsque $\beta = 1$, la F-mesure est la moyenne harmonique entre la précision et le rappel (précision et rappel sont pondérés de façon égale). **[Grouin & al, 2009]**

IV. Applications de la catégorisation des textes

La catégorisation des textes est utilisée dans de nombreuses applications comme l'identification de la langue, la reconnaissance d'écrivains et la catégorisation de documents multimédia, et bien d'autres.

Une autre forme de catégorisation est l'indexation automatique de textes parce qu'elle consiste à associer à chaque texte, un ou plusieurs termes. L'objectif est de décrire le contenu de ces textes par des mots ou des phrases qui font partie d'un ensemble de vocabulaire vérifié qui peut être vu comme des catégories. Aussi, le filtrage qui consiste à déterminer si un document est pertinent ou non, par exemple : la détection de spam (les courriers indésirables) pour ensuite les supprimer et le routage qui consiste à affecter un document à une ou plusieurs catégories, comme la diffusion sélective d'information.

V. Difficultés particulières de la catégorisation des textes

Le traitement de données textuelles est plus difficile que le traitement des données numérique. Le langage naturel est ambigu, il y a plusieurs façons d'exprimer la même idée (la redondance), ce qui est exprimé possède souvent plusieurs interprétations (l'ambiguïté) et tout n'est pas exprimé dans le discours (l'implicite).

Ajoute à ces particularités, une des difficultés majeures de la catégorisation, il s'agit de la dimension très élevée de l'espace de représentation qui peut prendre plusieurs centaines de milliers pour une collection de textes. En plus, le sur-apprentissage est un problème pouvant survenir dans les méthodes mathématiques et informatiques de catégorisation comme les réseaux de neurones. Il est en général provoqué par un mauvais dimensionnement de la structure utilisée pour la catégorisation.

VI. 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 multilingue et ses types avec les deux étapes supplémentaires pour l'apprentissage et/ou le classement des textes par rapport à la catégorisation des textes monolingue.

Chapitre II : Catégorisation des Textes Multilingue

I. Introduction

Grâce aux progrès et au développement des technologies informatiques, au réseau internet qui relie le monde entier, certains pays finalisent les moyens d'utiliser leurs langues nationales. Ces facteurs aident à la disponibilité de l'information multilingue qui devient indispensable de faciliter la communication d'internet entre différents paramètres de lieu, pays et langues.

Beaucoup de travail est actuellement concentré sur l'anglais puisque c'est la langue principale du web. Pourtant, un besoin se fait sentir pour les autres langues car le web est chaque jour plus multilingue puisque les utilisateurs actuels ne se contentent plus d'accéder aux informations et de les manipuler seulement dans leurs langues maternelles, mais ils tentent de plus en plus de franchir le pas vers les autres langues d'où l'apparition de la catégorisation des textes multilingue.

Plusieurs raisons ont été à l'origine pour les traitements de données multilingues : la disponibilité de plus en plus large des documents mis en réseau et distribués au plan international, le nombre croissant de non-anglophones qui se connectent en ligne, la création de zones de coopération entre des pays (Union Européenne, Forum Asie-Pacifique, etc.), le développement de l'infrastructure de communication et de l'Internet.

Ce chapitre présente les types de catégorisation des textes multilingue. Ensuite, les deux étapes supplémentaires par rapport au schéma classique de catégorisation des textes monolingue. Enfin, nous citons les difficultés de catégorisation des textes multilingue.

II. Définition

La catégorisation des textes multilingue 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. En effet il s'agit de savoir comment catégoriser un document en utilisant des documents d'autres langues.

III. Les types de catégorisation des textes multilingue

La catégorisation des textes multilingue se rapporte à l'attribution des documents basés sur leurs contenus, à une ou plusieurs catégories prédéfinies. La catégorisation des textes multilingue peut être traitée selon différents schémas.

III.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.

- *Traduction des documents étiquetés* : les documents étiquetés sont traduits dans la langue des documents non étiquetés afin d'être utilisé pour catégoriser ces derniers.
- *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.

III.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 d'information n'est faite.

III.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. **[Rigutini & al, 2005]**

IV. Travaux connexes

Le domaine de la catégorisation multilingue des textes étant un domaine récent, les travaux dans ce domaine ne sont pas nombreux par rapport aux autres domaines voisins. En effet, la majorité des travaux proviennent essentiellement de ces domaines voisins et plus particulièrement le domaine de la recherche d'information multilingue.

Les approches proposées par JALAM dans **[Jalam, 2003]** sont parmi les premiers qui abordent le domaine de la catégorisation multilingue. En effet, D'après JALAM trois solutions basées sur la traduction automatique sont proposés et qui sont :

- Le premier, nommé le schéma « trivial » qui représente une extension naïve du schéma de catégorisation monolingue habituel. Il consiste en l'apprentissage de plusieurs modèles (un modèle pour chaque langue).
- Le deuxième est un schéma permettant l'apprentissage d'un seul modèle.
- Le troisième consiste à la traduction de textes de plusieurs ensembles d'apprentissage vers une langue cible pour ensuite apprendre un seul modèle « mixte ».

Une autre approche proposée dans **[Gliozzo & al, 2005]** consiste à résoudre le problème de la catégorisation multilingue par la construction d'un modèle multilingue du domaine à partir d'un corpus comparable, afin de définir par la suite une fonction de similarité générale entre les documents de différentes langues. Cette fonction est utilisée dans un classificateur SVM.

Vu le sucées d'utilisation des ontologies dans la catégorisation monolingue, une autre approche proposée dans **[Guyot & al, 2005]** consiste à utiliser une ontologie multilingue pour la recherche d'information multilingue en écartant l'utilisation des techniques de traduction automatique.

Christopher YANG, Chih-Ping WEI et Huihua SHI ont proposés dans **[Yang & al, 2007]** une approche pour le cas de la catégorisation multilingue par multiple langue qui consiste en trois principales phases :

- *Construction de thésaurus bilingue* en utilisant la technique d'analyse de cooccurrence généralement utilisée dans la recherche d'information par croisement de langue (dite en anglais Cross Language Information Retrieval) et CLTC.
- *Apprentissage de la catégorisation* en tenant en compte non seulement des documents pré classifiés en une langue L1 mais également des documents pré classifiés en une autre langue L2 et en utilisant aussi le thésaurus bilingue construit.
- *Assignation de la catégorie* pour chaque document non classifié dans L1 ou L2 en utilisant le modèle correspondant de catégorisation des textes induit précédemment.

Selon la langue utilisée dans le document non classifié, il nécessite d'employer la méthode respective d'extraction d'attributs pour extraire des attributs à partir du document non classifié. En conclusion, le vecteur de document - attribut est employé pour déterminer une catégorie appropriée sur la base du modèle correspondant de catégorisation des textes.

Une autre approche proposée dans **[Bentaallah & al, 2007]** consiste à utiliser le thésaurus WordNet dont le but de réduire les pertes d'informations causées par l'utilisation des techniques de traduction automatique.

V. Identification de la langue

La détection de la langue dans laquelle le texte à classifier est rédigé est très importante. Elle consiste à attribuer une unité textuelle, supposée monolingue, à une langue. Cette identification devient alors intéressante puisque nous parlons de multilinguisme.

Il existe deux familles d'approches dans l'identification de la langue :

- **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.
- **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.

VI. 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.

Un autre avantage est que les utilisateurs peuvent immédiatement recevoir les documents en leurs langue préférée, ce qui leurs permet de les consulter directement. **[Kadri, 2008]**

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 **[Sahnoun & Haddar, 2009]** trois types de TA qui sont :

- **Traduction mot à mot :** c'est une traduction directe qui ce 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 réussite dans des domaines limitées, mais elle est utilisé seulement pour un couple de langue et pas d'analyses profondes.
- **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 ce 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.
- **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.

VII. 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'ambigüité 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.

VIII. Conclusion

La catégorisation des textes multilingue est un nouveau secteur dans la catégorisation des textes dans laquelle nous devons faire face à deux langues ou plus. Le multilinguisme introduit des contraintes supplémentaires. Il faut adapter le processus habituellement mis en œuvre pour classer les nouveaux textes et certaines techniques à base linguistique utilisées en monolingue, deviennent alors inefficaces.

Dans ce chapitre, nous avons décrits les deux types de catégorisation des textes multilingue à savoir catégorisation des textes par croisement de langues et catégorisation des textes par langues multiples ainsi que les travaux connexes et nous avons conclu par les différents problèmes rencontrés dans ce domaine.

Le chapitre suivant présente la problématique, les expérimentations et l'évaluation de notre travail qui consiste à la représentation conceptuelle pour la catégorisation des textes multilingue en comparant avec une représentation en sac de mots.

Chapitre III : La Représentation Conceptuelle pour la Catégorisation des Textes Multilingue

I. Introduction

Le principe de la catégorisation des textes a été abordé aux chapitres I et II. Nous avons vu en quoi consistait la catégorisation des textes, quels traitements appliqués pour les documents non étiquetés. Aussi, les différents types de la catégorisation multilingue.

Le présent chapitre est organisé de la manière suivante : La section II est dédiée à la problématique qui vise à poser les objectifs et besoins pour la représentation conceptuelle, ainsi que les étapes à suivre pour l'évaluation du processus de catégorisation des textes multilingue. La section III présente de façon détaillée les approches implémentées. En premier, elle illustre une vue générale et schématique des approches suivies puis la description de ses étapes, afin de réaliser notre processus, en utilisant des techniques de recherche d'information, une heuristique à choisir en se basant sur des stratégies de mesure de similarité, combinant en quelque sorte les idées des deux chapitres précédents.

Ces approches dont l'une s'appuie sur des concepts issus d'une base de données lexicographique WordNet pour l'extraction de mapping des mots en synsets qui consiste à découvrir la correspondance sémantique entre les différents termes et la traduction automatique à partir d'une langue source vers une langue cible. Et l'autre s'appuie sur la représentation en sac de mots.

La section IV qui suit dans ce chapitre sera consacrée à la partie expérimentation et évaluation des approches implémentées et nous terminons par une conclusion.

II. Problématique

La problématique dont ce mémoire traite, consiste dans la catégorisation des textes dans des langues différentes selon le même arbre de classification, en représentant les textes par des concepts. C'est l'un des domaines qui tente d'apporter des améliorations et de réduire la tâche de l'humain.

L'objectif à viser c'est de chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories. Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés dans une langue donnée à partir duquel nous évaluons les paramètres du modèle de prédiction pour associer automatiquement une étiquette à tout nouveau texte rédigé dans une langue quelconque.

Notre travail s'intéresse à la représentation conceptuelle, dans laquelle l'unité de vecteur serait un concept (groupe des synonymes), en comparant avec une représentation en sac de mots dont l'unité de vecteur serait un mot. Cette représentation conceptuelle nous permet de voir comment l'intégration d'une ressource externe telle que WordNet admet une amélioration des performances de classification.

Pour cela, nous avons implémenté deux approches qui se différencient dans l'étape de représentation. En effet la première consiste à représenter les documents étiquetés en utilisant WordNet puisqu'ils sont exprimés en L1, et les documents non étiquetés en L2 doivent être traduits afin de pouvoir être traités en utilisant WordNet. Et la deuxième, basée sur la représentation en sac de mots.

Pour les deux approches, nous devons appliquer les méthodes issues de l'apprentissage automatique pour la catégorisation des textes. Nos premiers besoins d'évaluation s'expriment à l'aide d'une présentation générale de la catégorisation des textes selon les étapes suivantes :

a. Représentation des textes dans un format adapté aux algorithmes d'apprentissage qui englobe deux éléments : le choix des termes et leurs pondérations.

b. Choix d'une méthode d'apprentissage pour construire un modèle de prédiction. Il s'agit d'appliquer une méthode qui associe une ou plusieurs catégories à un document non étiqueté. Pour cela, nous choisissons la méthode de K-plus proches voisins (Kppv).

c. Evaluation du modèle afin de s'assurer qu'il est généralisable à d'autres textes.

La section suivante présente une description des approches suivies, en utilisant les trois dernières étapes et le mapping des mots en synsets à travers le WordNet, afin de pouvoir construire une représentation conceptuelle pour la catégorisation multilingue des textes à classer, en utilisant l'algorithme de K-plus proches voisins.

III. Description des approches suivies

- **Première approche**

Comme illustré dans la figure ci-dessous, la première approche utilise la représentation "sac de mots" comme méthode de représentation. Les documents seront représentés par des vecteurs de mots sans utilisation de ressource sémantique.

Une fois les documents représentés, ils seront pondérés afin de donner un poids pour chaque mot dans chaque document. La pondération utilisée est la pondération TF.

Par la suite, nous utilisons la méthode Kppv pour classer les nouveaux documents qui doivent être traduits, représentés et pondérés.

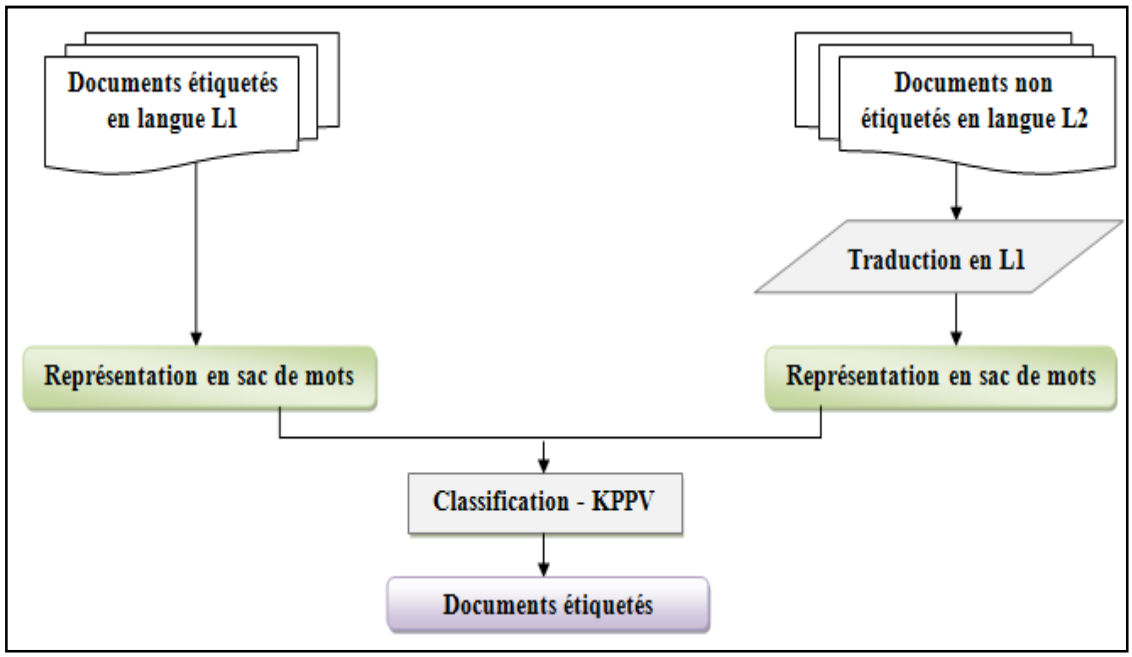

Figure 3.1 : Processus de la première approche

- **Deuxième approche**

Celle-ci est basée sur le WordNet pour traiter les documents étiquetés en langue L1 et la traduction automatique des documents non étiquetés en L2, afin de les exprimer en sac de mots et faire la représentation conceptuelle en utilisant WordNet.

La figure 3.2 illustre cette approche comme suit :

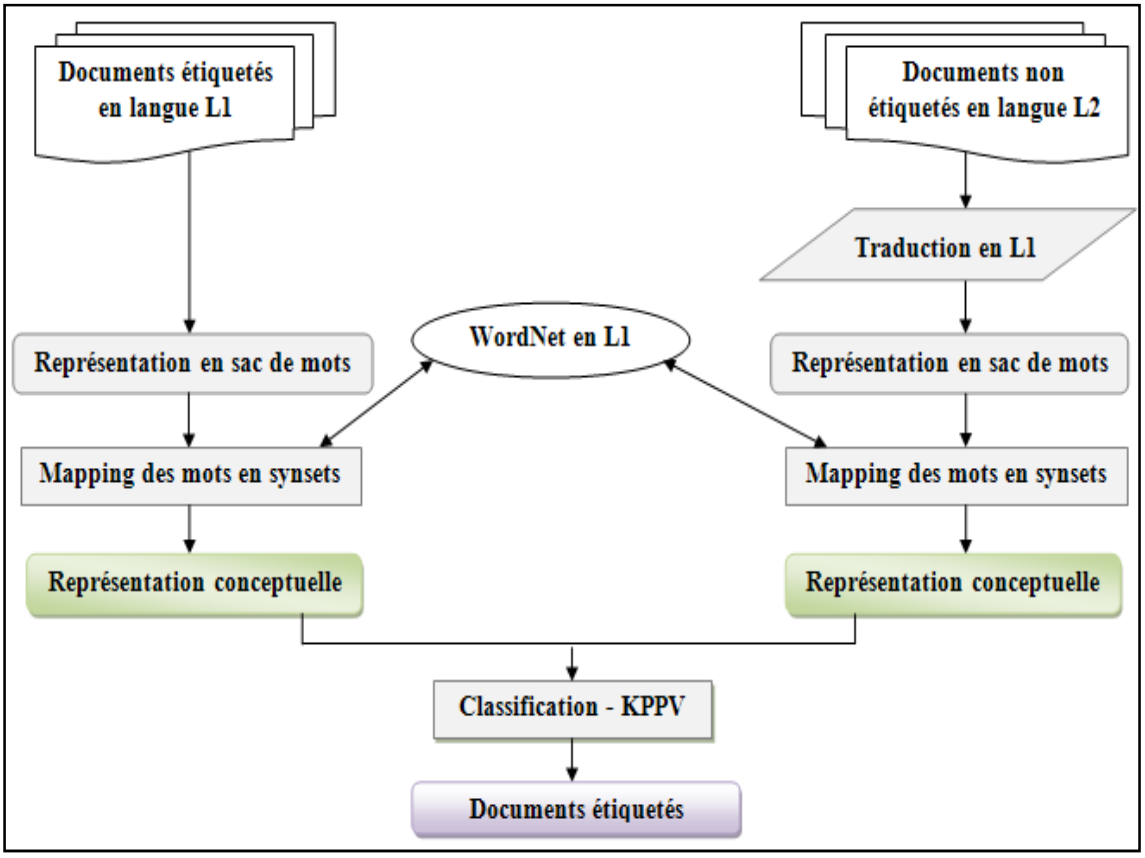

Figure 3.2 : Processus de la deuxième approche

La deuxième approche est différente de la première dans l'étape de représentation. Chacune de ces étapes contiennent un ou plusieurs mécanismes, dans ce qui suit nous allons expliquer chacune d'entre elles.

III.1. Représentation en sac de mots

Etant donné que les documents étiquetés et les documents non étiquetés du corpus, sont exprimés en langage naturel, une série de prétraitements a été mise en œuvre pour extraire l'ensemble des mots. Les textes sont transformés en vecteurs dont chaque composante représente un terme.

Le prétraitement consiste à :

- Filtrer le texte en enlevant les balises HTML,
- Convertir les majuscules en minuscules,
- Enlever les caractères non alphanumériques : les mots sont séparés par des espaces ou des signes de ponctuations, des chiffres.
- Elimination des mots vides.

Dans le cas ou le prétraitement est appliqué sur les documents non étiquetés sachant que ces derniers sont exprimés dans la langue L2 (espagnol), l'élimination des mots vides se fera en fonction de la liste des mots vides de la langue espagnole[1].

Les composants du vecteur sont en fonction de l'occurrence des mots dans le texte. À titre d'exemple, nous présentons dans la Figure 3.3, la transformation d'un texte en vecteur.

The New Zealand government thumbed its nose at public and political opposition on Tuesday by pressing ahead, with the privatisation of a huge plantation forest less than two months before a general election.
It announced the sale of the Forestry Corporation of New Zealand, to a consortium of New Zealand companies, Fletcher Challenge and Brierley Investments and China's China International Trust and Investment Corp for NZ$2.0 billion (US$1.38 billion).

ahead	1	fletcher	1	opposition	1
announced	1	forest	1	plantation	1
brierley	1	forestry	1	pressing	1
challenge	1	general	1	privatisation	1
china's	2	govemment	1	public	1
companies	1	huge	1	sale	1
consortium	1	international	1	thumbed	1
corp	1	investments	2	tuesday	1
corporation	1	months	1	trust	1
election	1	new	3	zealand	3
		nose	1		

Figure 3.3 : Exemple d'un texte anglais et son vecteur associé

[1] La liste des mots vides de la langue espagnole est sur le site:
http://www.ranks.nl/stopwords/spanish.html

III.2. Traduction automatique

Cette étape est utilisée seulement dans les deux approches, dont nous avons traduit les textes rédigés en L2 avant l'étape de représentation en sac de mots.

Le but de la traduction automatique est de produire une interprétation fiable dans la langue cible du texte source, alors que la recherche multilingue vise à trouver assez de similarités entre un document dans une langue source et un document dans une langue cible. La traduction de toute une collection de documents dans une autre langue (celle du corpus d'apprentissage) implique un nombre de tâches inutiles du simple point de vue de la recherche, par exemple l'encodage de l'information linguistique, sémantique et pragmatique. La traduction a était faite par WebTranslator[2] API qui est une bibliothèque pour 14 langues incluant l'anglais, Espagnol, Français, Italien, Deutsch, le grec, le chinois, le japonais, le russe…

III.3. Représentation conceptuelle

La représentation conceptuelle des deux corpus celui de l'étiqueté rédigé dans la langue anglaise et le non étiqueté écrit en espagnol, résulte du traitement du mapping des mots en synsets expliqué ci dessous.

Cette représentation se base sur le formalisme vectoriel pour représenter les documents. Les éléments de cette représentation ne sont plus associés directement à des simples mots mais plutôt à des concepts. Pour cela, il est nécessaire de pouvoir projeter nos termes sur un thesaurus ou une base lexicographique comme Wordnet dans laquelle les mots sont regroupés au sein de groupes de synonymes appelés synsets. Cet outil, ayant pour objectif de représenter des aspects sémantiques d'un lexique. La deuxième approche contient les concepts associés aux mots du document représenté.

Ce type de représentation est utilisé afin de réaliser une catégorisation. Le schéma ci-dessous illustre la représentation matricielle d'un corpus où les lignes représentent les documents du corpus, les colonnes représentent les termes, et l'intersection entre un document Di et un terme Tj représente le nombre d'occurrences du terme Tj dans le document Di.

[2] WebTranslator est disponible sur le site :
http://en.sourceforge.jp/projects/sfnet_webtranslator/downloads/JavaWebTranslator-0.2a/Java%20WebTranslator%200.2a%20(Alpha%20Release)/WebTranslator-bin-0.2a.jar/

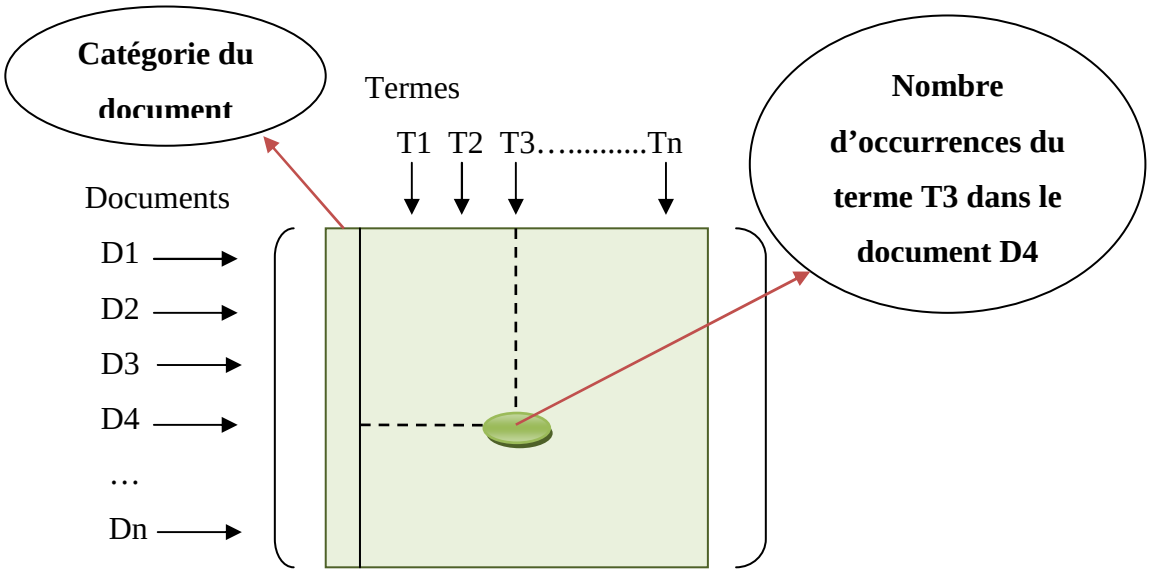

Figure 3.4 : Représentation matricielle d'un corpus

Pour faire la représentation conceptuelle, il est nécessaire d'utiliser le mapping des mots en synsets.

- ***Mapping des mots en synsets***

Après le prétraitement de chaque document et son représentation sous forme d'un vecteur, nous passons à l'étape du mapping des mots en synsets en s'appuyant sur la base lexicographique WordNet.

Le Synset est un ensemble de synonymes qui est la composante importante sur laquelle repose WordNet. Chaque synset indique un sens différent du mot, décrit par une courte définition appelée glossaire. La base lexicographique WordNet renvoie une liste ordonnée de synsets pour chaque mot qui peut ajouter le bruit à la représentation et peut induire une perte d'information. Alors il y aura un problème de la désambigüisation de sens, ce qui n'est pas l'intention de notre étude.

Dans notre cas et pour ne pas avoir ce genre de problème, notre processus consiste à remplacer directement chaque mot par sa première signification en considérant qu'elle est la plus appropriée. A titre d'exemple, le mot « government » a 4 sens et nous prenons juste le premier élément: « authorities, regime -- (the organization that is the governing authority of a political unit; "the government reduced taxes"; "the matter was referred to higher authorities") ».

III.5. Classification K-PPV

Une fois les documents analysés et représentés dans le même espace vectoriel, la méthode de K plus proches voisins ordonne les textes les plus proches sémantiquement au document à classer pour les regrouper par catégorie. Ce choix a été fait pour la simplicité d'utilisation de cet algorithme et sa fréquente utilisation dans le domaine de la catégorisation des textes.

Lors de l'étiquetage d'un nouveau document à classer dans une ou plusieurs catégories, il sera comparé aux documents étiquetés à l'aide d'une mesure de similarité. Ses K plus proches voisins sont alors considérés, en observant les catégories des textes déjà étiquetés, et celle qui revient le plus parmi les catégories de ces voisins, est assignée au document à classer.

Une des caractéristiques fondamentales de ce type de classificateur est l'utilisation d'une mesure de similarité entre les documents. Ces derniers étant représentés sous la forme vectorielle, donc comme des points dans un espace à n dimensions.

La mesure de similarité utilisée est la mesure cosinus qui est préférable en catégorisation de textes pour plusieurs raisons :

- Elle s'étend aux vecteurs pondérés et est devenue la mesure standard des vecteurs pondérés dans le domaine de la recherche d'information.
- Elle résout certains problèmes essentiels à d'autres mesures telles que le produit scalaire comme la favorisation des vecteurs longs, discrimination des vecteurs dont la différence entre les longueurs est significative.

Rappelons que la mesure de similarité cosinus entre deux documents a et b, est calculé comme suit :

$$\textbf{Cosinus (a, b)} = \frac{\sum(\mathrm{Pt(a)} \times \mathrm{Pt(b)})}{\sqrt{\sum \mathrm{Pt(a)}^2 \times \sum \mathrm{Pt(b)}^2}}$$

Ou : Pt(a) est le poids du terme t dans le document a et Pt(b) est le poids du terme t dans le document b.

La valeur de K est un paramètre à déterminer lors de l'utilisation de ce type de classificateur.

La figure 3.5 décrit l'algorithme de K-PPV.

Pour chaque nouveau document faire :

- Représenter ce document en vecteur de termes,
- Déterminer les K plus proches documents au document à classer selon la mesure de similarité cosinus,
- Calculer la présence de chaque catégorie dans les k documents sélectionnés,
- Affecter la catégorie majoritaire au nouveau document.

Figure 3.5 : Algorithme de K plus proches voisins

IV. Expérimentation et évaluation

Afin d'évaluer et valider la contribution présentée dans ce mémoire, une phase d'expérimentation s'avère indispensable. Cette phase a pour objectif d'étudier les performances de nos approches implémentées. En outre, ceci nous permettra aussi d'identifier les contraintes et les insuffisances de nos approches.

Une présentation de l'environnement de développement qui va supporter notre application ainsi que les différentes ressources utilisés sont décrites dans un premier lieu dans cette section. La suite sera consacrée à l'évaluation des résultats.

IV.1. Technologies et outils de développement

Nos expérimentations ont été développée sur une machine possédant les caractéristiques suivantes : un processeur Intel ® Dual-Core CPU T4400, 2.20 GHz et une mémoire de 2GO. L'ensemble est piloté par le système d'exploitation Windows 7. Les outils et langages utilisés pour la manipulation des données ainsi que l'implémentation sont décrits comme suit.

IV.1.1. Langage JAVA

Notre choix pour le langage de programmation s'est porté sur le langage JAVA, et cela parce qu'il est un langage orienté objet simple ce qui réduit les risques d'incohérence et il possède une riche bibliothèque de classes comprenant des fonctions diverses telles que les fonctions standards, le système de gestion de fichiers ainsi que beaucoup de fonctionnalités qui peuvent être utilisées pour développer des applications diverses. Il existe une multitude de bibliothèques développées et fournies pour être utilisées en JAVA. Les API (Application Programming Interface) des autres langages autres que JAVA ne sont pas finalisées et doivent encore être mises à jour.

IV.1.2. Environnement de développement

L'environnement de développement utilisé, est NetBeans 7.0 car il possède de nombreux points forts qui sont à l'origine de son énorme succès dont les principaux sont :

- Un environnement de développement intégré (EDI).
- En plus de JAVA, NetBeans[3] permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).
- Les principaux modules de base pour NetBeans concernent le langage de programmation JAVA. Les modules agissent sur des fichiers qui sont inclus dans l'espace de travail (appelé workspace). Ce dernier regroupe les projets qui contiennent une ou plusieurs arborescences de fichiers.
- La construction incrémentale des projets JAVA grâce à son propre compilateur, qui permet en plus de compiler le code même avec des erreurs, de générer des messages d'erreurs personnalisés, de sélectionner la cible, …

IV.1.3. WordNet

Pour la mise en correspondance entre les deux corpus, nous avons utilisé WordNet[4] de version **2.0** qui est une base de données lexicographique. Le choix de WordNet été cause de diverses raisons :

- C'est la base la plus riche et la plus générale qui contient tous les domaines,
- Il utilise la langue anglaise qui est la langue la plus utilisée dans le monde. Des versions de ce dernier existent pour d'autres langues.

La structure du Wordnet repose sur des ensembles de synonymes appelés synset (synonym set en anglais). Chaque synset représente alors un sens, un concept de la langue anglaise. Chacun d'eux contient tous les mots synonymes pouvant exprimer le sens auquel il fait référence. Les liens sémantiques ne relient alors pas les mots entre eux mais les synsets auxquels les mots sont affectés.

[3] Site : www.Netbeans.org
[4] Lien : http://wordnet.princeton.org

Le tableau III.1 ci-dessous montre la structure de WordNet d'anglais noté EWN en nombre de mots, nombre de synsets et nombre de sens, globalement et par catégorie grammaticale. La plupart sont des noms (74.6%), le reste étant constitué par des adjectifs (14.6%), des verbes (7.6%) et des adverbes (3.2%). La polysémie (nombre de sens par mot) se manifeste dans Wordnet par le fait qu'il y a des mots qui peuvent appartenir à plusieurs synsets (146350 formes traitées / 111223 synsets).

Tableau III.1 : Caractéristiques du nombre de mots et de concepts dans WordNet

Catégorie	*Mots*	*Concepts*	*Total Paires Mots-Sens*
Nom	109195	75804	134716
Verbe	11088	13214	24169
Adjectif	21460	18576	31184
Adverbe	4607	3629	5748
Total	146350	111223	195817

IV.1.4. JWNL API

JWNL[5] (Java WordNet Library) est une API Java pour avoir accès au dictionnaire relationnel WordNet dans des formats multiples, aussi bien que la découverte des relations hiérarchiques et de traitement morphologique. Elle est compatible avec des versions WordNet 2.0 à 3.0 et est une mise en œuvre Java complète. L'API courant est JWNL 1.3. JWNL 1.4 est dans le développement.

IV.2. Corpus utilisé

Le corpus d'ILO (International Labour Organisation) est une collection de documents à texte intégral, chacune a une catégorie prise de l'organisation international du travail. ILO a une base de données trilingue contenant des conventions d'ILO et des recommandations, l'information de ratification, commentaires du comité d'experts et comité de la liberté d'association, de représentations, de plaintes, d'interprétations, d'aperçus généraux, et de nombreux documents relatifs. Les langues concernées sont l'anglais, l'espagnol et le Français.

[5] JWNL est disponible sur : http://sourceforge.net/projects/jwordnet/

Notre base de test est basée sur des codes de catégories choisis avec un nombre de documents (en anglais et en espagnol) par catégorie représenté comme suit :

Tableau III.2 : Les six catégories choisies du corpus multilingue d'ILO

Code de catégorie	*Description de la catégorie*	*Anglais*	*Espagnol*
1	Lois spéciales du secteur économique	108	121
2	Les conditions d'emploi	394	86
3	Les conditions de travail	299	71
4	Développement économique et social	22	23
5	Emploi	410	448
6	Relations de travail	276	278

IV.3. Evaluation

Afin de pouvoir montrer le rôle de l'utilisation de WordNet dans la catégorisation des textes multilingue, nous avons examiné les approches sur un corpus de test extrait du corpus d'ILO décrit ci-dessus. Un point important à clarifier à propos de notre méthodologie d'évaluation est le choix de classificateur dont l'un d'entre eux a été mis à l'œuvre dans le cadre de notre travail, à savoir le classificateur des K plus proches voisins (Kppv). Les grandes lignes de cet algorithme ont déjà été présentées au chapitreI.

Les paramètres d'évaluation standard sont : la précision (P) et le rappel (R). Pour ce faire, nous évaluons nos approches expliquées précédemment.

A. Pour la première approche qui se base sur la représentation en sac de mots, les résultats obtenus sont comme suit :

Tableau III.3 : Précision et Rappel pour la première approche

Catégorie	K=1		K=3		K=5	
	P	R	P	R	P	R
Lois spéciales du secteur économique	0.8333	0.8696	0.8	0.8275	0.7583	0.8425
Les conditions d'emploi	0.6512	0.8116	0.5581	0.7058	0.6511	0.6666
Les conditions de travail	0.7887	0.6087	0.6901	0.4579	0.7042	0.3968
Développement économique et social	0.8261	1.0	0.5652	1.0	0.3913	1.0
Emploi	0.9241	0.9718	0.8839	0.9658	0.875	0.9824
Relations de travail	0.9494	0.8651	0.9422	0.8392	0.9314	0.8628

B. Pour la deuxième approche conceptuelle qui utilise WordNet pour la catégorisation des textes non étiquetés, les résultats obtenus sont comme suit :

Tableau III.4 : Précision et Rappel pour la deuxième approche

Catégorie	K=1		K=3		K=5	
	P	R	P	R	P	R
Lois spéciales du secteur économique	0.8916	0.8699	0.7333	0.8888	0.6833	0.9111
Les conditions d'emploi	0.7209	0.7469	0.6162	0.6794	0.6046	0.6117
Les conditions de travail	0.8873	0.525	0.8732	0.3583	0.8450	0.2752
Développement économique et social	0.8695	1.0	0.4782	1.0	0.3478	1.0
Emploi	0.9017	0.9782	0.8459	0.9594	0.7991	0.9781
Relations de travail	0.9205	0.9586	0.9133	0.9405	0.8808	0.9457

Après l'analyse des résultats obtenus à la suite de cette comparaison entre ces deux approches implémentées, les graphes suivants, ont été réalisés comme suit :

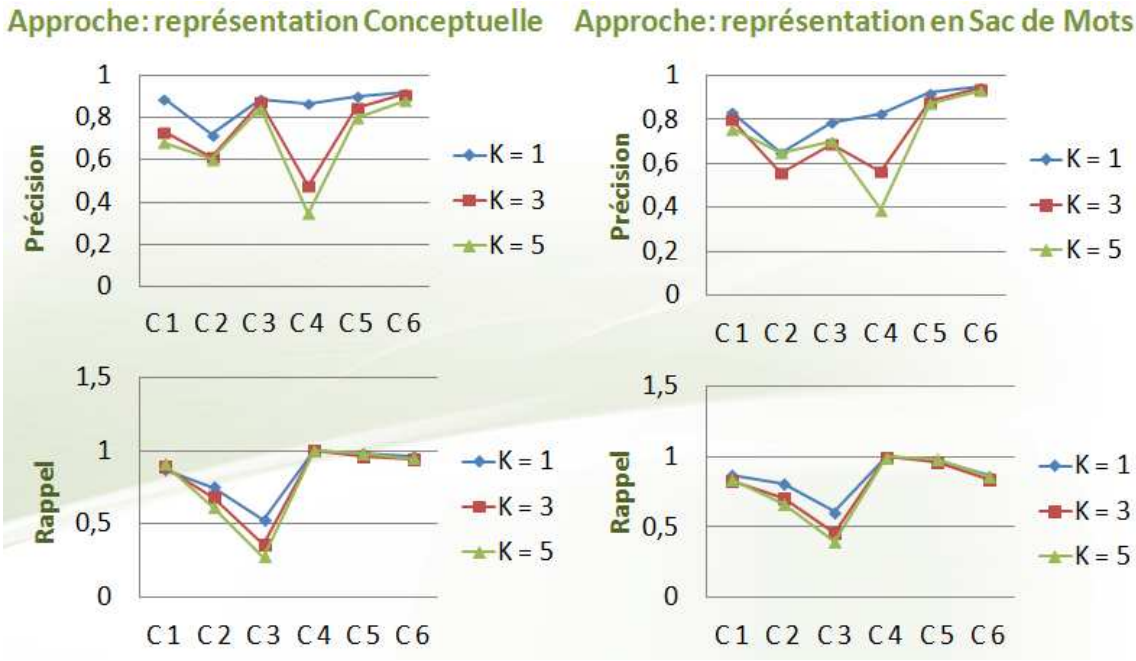

Figure 3.6 : les graphes réalisés des deux approches

V. Discussion

Dans la section précédente les expérimentations ont été évaluées pour la comparaison des deux approches implémentées pour catégoriser les documents non étiquetés.

Les deux tableaux ci-dessus, montrent respectivement les résultats de catégorisation des textes multilingue lorsque les documents sont représentés en sac de mots (approche 1) et en concepts (approche 2) sur le corpus d'ILO. Sur cette collection et dans la majorité des cas, la précision et le rappel de la deuxième approche est légèrement supérieure à la première approche.

Mais pour bien interpréter nos résultats on s'appuie sur les deux graphes de la précision en constatons que l'approche qui s'appuie sur la représentation conceptuelle est plus significative que celle qui s'appuie sur la représentation en sac de mots. L'approche implémentée est très bien adaptée au contexte de la catégorisation du corpus puisque les mesures d'évaluation sont uniformes (généralement presque égaux). Ceci est important face à la masse des documents à classer, ce qui permet d'obtenir une classification meilleure, mais pour plus dont le cas où le paramètre de Kppv soit égal à 1 (k=1), malgré la petite différence par rapport aux deux autres cas (k=3) et (k=5).

Les bons résultats obtenus sont dues à un modèle basé sur une représentation conceptuelle des textes qui possède une information supplémentaire par rapport à la représentation en sac de mots.

Nous observons que l'approche implémentée a un avantage, c'est la connaissance de tous les lemmes d'un texte grâce au WordNet. La prise en compte des concepts permet d'améliorer encore les résultats.

VI. Conclusion

Ce chapitre a été consacré à la description et la mise en œuvre des approches implémentées, qui avaient comme intérêt de présenter un processus de catégorisation des textes multilingue en utilisant une représentation conceptuelle basée sur la base lexicographique WordNet. La collection de test d'ILO est la ressource à partir de laquelle nous avons pu constituer des ensembles d'exemples de référence pour l'apprentissage automatique dans les deux langues rédigés en anglais et en espagnol. La description de chaque étape de notre processus été comme suit :

- pour la représentation des documents, le choix été porté sur la traditionnelle représentation sacs de mots.
- La traduction automatique des textes dans la langue source vers la langue cible.
- Le mapping des mots en synsets en utilisant la base lexicographique WordNet.
- La représentation conceptuelle qui est le résultat de mapping des mots en synsets.
- Et la classification en utilisant la méthode KPPV.

Conclusion Générale

Nos travaux développés dans ce mémoire s'inscrivent dans le cadre de la représentation conceptuelle pour la catégorisation multilingue des textes. Rappelons que le but de la catégorisation est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu.

Notre mémoire s'articule en trois chapitres, le premier chapitre introduit un cadre général dont nous avons présenté les points cruciaux du domaine de la catégorisation des textes monolingue : le processus et les principales méthodes d'apprentissage ayant fait leurs preuves dans ce domaine, aussi, les difficultés liées à la catégorisation des textes. Dans le second chapitre, nous avons décrit les types de la catégorisation des textes multilingue. Nous avons ensuite mis l'accent sur les différentes difficultés particulaires à la catégorisation des textes multilingue. Dans le troisième chapitre, nous avons proposé une approche permettant d'associer automatiquement une étiquette à tout nouveau texte rédigé dans une langue L2 en se basant sur un ensemble de textes préalablement étiquetés dans une langue donnée L1, dit ensemble d'apprentissage en s'appuyant sur les étapes du processus de catégorisation des textes monolingue expliqué au chapitre I.

La représentation conceptuelle dans laquelle l'unité de vecteur serait un concept (groupe des synonymes appelé synsets), nous a permis de voir comment l'intégration d'une ressource externe Wordnet a permis l'amélioration de la performance de notre classificateur. Les éléments de cette représentation ne sont plus associés directement à des simples mots mais plutôt à des concepts.

Malheureusement, le temps est court et il a été nécessaire de fixer certains paramètres pour en étudier d'autres plus en profondeur. Évidemment, il aurait été intéressant d'observer le comportement de notre approche sur plus de deux corpus de textes et sur plus d'un classificateur. Notre perspective dans un premier temps, est de consolider la démarche implémentée en évaluant sur d'autres collections, puis élargir notre domaine vers d'autres langues avec plus de corpus de textes et de travailler avec la dernière version WordNet 3.0 si ça donne de meilleurs résultats.

Références Bibliographies

[Abidi, 2011] Karima ABIDI, « La catégorisation de texte Multilingue », Mémoire de magistère, Ecole supérieur d'Informatique, Algérie, 2010-2011.

[Bel & al, 2003] Nuria Bel, Cornelis H.A. Koster et Marta Villegas, «Cross-Lingual Text Categorization», Université de Barcelone, Espagne, Université de Nijmegen, Pays-Bas, 2003.

[Bentaallah & al, 2007] Bentaallah, M.A., Malki, M.: WordNet based Multilingual Text Categorization. Journal of Computer Science, Vol 6 (2007)

[Caropreso, 2001] Caropreso Maria Fernanda, Stan Matwin , Fabrizio Sebastiani, 2000 « Statistical Phrases in Automated Text Categorization » Department of Computer Science of the

[Dziczkowski, 2008] Grzegorz DZICZKOWSKI, « Analyse des sentiments : système autonome d'exploration des opinions exprimées dans le critiques cinématographiques », Thèse de doctorat, Paris, 4 Décembre 2008.

[Gliozzo & al, 2005] Gliozzo, A., Strapparava, C. « Cross Language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora». Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan, USA, pages 9–16, 2005.

[Grouin & al, 2009] Cyril Grouin, Martine Hurault-Plantet, Patrick Paroubek, Jean-Baptiste Berthelin, « DEFT'07 : une campagne d'évaluation en fouille d'opinion », LIMSI–CNRS, Orsay, France, 2009.

[Guyot & al, 2005] Guyot, J., Radhouani, S., and Falquet, G.: « Ontology based multilingual information retrieval». In CLEF Workhop, Working Notes in Multilingual Textual Document Retrieval Track. Vienna, Austria, 2005.

[Hernandez, 1999] Nathalie HERNANDEZ, « Etude de l'utilisation des syntagmes nominaux pour la catégorisation automatique de documents », Institut de Recherche en Informatique de Toulouse, France, 1999.

[Ignat, 2007] Camelia Ignat, « Représentation de textes a l'aide d'étiquettes sémantiques dans le cadre de la classification automatique », European Commission, IPSC, Strasbourg, France, 2007.

[Jaillet, 2004] Simon JAILLET, « Catégorisation automatique de documents » LIRMM UMR 5506, 161 rue Ada, 34392 Montpellier Cedex 5 France

[Jaillet & al, 2005] Simon JAILLET, Maguelonne TEISSEIRE, Jacques CHAUCHE, Violaine PRINCE, « Classification automatique de documents, Le coefficient des deux écarts », Université Montpellier2, France, 2005.

[Jalam, 2003] Radwan JALAM, « Apprentissage automatique et catégorisation de textes multilingues », Thèse de doctorat, Université Lumière Lyon 2, France, Juin 2003.

[Kadri, 2008] Kadri Youssef, «Recherche d'information translinguistique sur les Documents en Arabe », thèse en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en Informatique, Université Montpellier, France, septembre 2008.

[Mallak, 2011] Ihab Mallak**, «** De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information », l'Université Toulouse III - Paul Sabatier, Discipline ou spécialité **:** Informatique et applications, 2011.

[Obozinski, 2010] Guillaume OBOZINSKI, « Introduction aux modèles graphiques », cours, Décembre 2010-2011.

[Porter, 1980] M. F. Porter, «An algorithm for sufix stripping », Program, pp 130–137, Morgan Kaufmann Publishers Inc, 1980.

[Ralaivola, 2006] Liva Ralaivola, « Modèles de représentation, sélection d'attributs, classification, catégorisation », présentation PPT, Université de Provence, France, 19 décembre 2006.

[Réhel, 2005] Simon RÉHEL, « Catégorisation automatique de textes et Cooccurrence de mots provenant de documents non étiquetés », Mémoire, Université Laval Québec, Canada, Janvier 2005.

[Rigutini & al, 2005] Leonardo Rigutini, Marco Maggini et Bing Liu, « An EM based training algorithm for Cross-Language Text Categorization», Université de di Siena, Italie, Université Illinois à Chicago, USA, 2005.

[Sahnoun & Haddar, 2009] Hadjer Sahnoun, Kais Haddar, « Étude comparative des techniques de la traduction automatique et leurs expérimentations sur les entités nommées avec NOOJ », Laboratoire MIRACL, FSS, 2009.

[Sebastiani, 2002] FABRIZIO SEBASTIANI, «Machine Learning in Automated Text Categorization», Conseil recherche National, Italie, Mars 2002.

[Yang & al, 2007] Christopher C. Yang , Chih-Ping Wei , Huihua Shi, «Feature Reinforcement Approach to Poly-Lingual Text Categorization», Institut de gestion de technologie, national Tsing Hua Université de Taiwan, Département de gestion de l'information Université chinoise de Hong Kong, Shatin, N.T., Hong Kong, 2007.

[Zeggane Mokhtar, 2009] Yasmine Hanane Zeggane Mokhtar, « Algorithmes d'apprentissage pour la classification de documents », Mémoire de licence en ligne, Université de Mostaganem-Algérie, 2009.

Résumé

Ce mémoire s'inscrive dans la problématique générale liée à la représentation conceptuelle pour la catégorisation de textes multilingues. Le but est de représenter des documents et des catégories à l'aide d'un même formalisme, qui se repose sur une représentation vectorielle des documents qui à son tour, n'a plus axée sur des mots mais sur une représentation plus sémantique de ceux-ci. L'objectif est d'associer automatiquement une étiquette à tout nouveau texte rédigé dans la langue espagnole en se basant sur un ensemble de textes préalablement étiquetés dans la langue anglaise. Cette représentation conceptuelle s'appuie sur des concepts issus de la base de données lexicographique WordNet et l'expérimentation est effectuée sur un corpus extrait du corpus d'ILO.

Mots clés : représentation conceptuelle, représentation en sac de mots, catégorisation de textes, multilingue, WordNet, mapping.

Abstract

This memory falls under the general problems related to the conceptual representation for the multilingual text categorization. The purpose is to represent documents and categories using the same formalism, which rests on a vectorial representation of the documents which in its turn, did not center any more on words but on a more semantic representation of those. The objective is automatically to associate a label all new text written in the Spanish language while being based on a whole of texts labelled beforehand in the English language. This conceptual representation is based on concepts resulting from the lexicographical data base WordNet and the experimentation is carried out on a corpus extracted the corpus of ILO.

Key words: representation conceptual, representation out of bag of words, categorization of texts, multilingual, WordNet, mapping.

ملخص

هذه المذكرة تتحدث عن المشكل العام المتعلق بتمثيل المفاهيم لتصنيف النصوص متعددة اللغات. الغرض هو تمثيل الوثائق والفئات باستخدام نفس الشكلية، التي تقوم على تمثيل متجه من الوثائق التي بدورها لم تركز أكثر على الكلمات ولكن على التمثيل الدلالي لها. والهدف من ذلك هو ربط تلقائيا النص الجديد مع فئة مكتوب باللغة الإسبانية، استنادا إلى مجموعة من النصوص مصنفة سابقا باللغة الإنجليزية. هذا التمثيل يستند على المفاهيم النظرية من قاعدة البيانات المعجمية WordNet و يتم تنفيذ التجربة على قاعدة وثائقية مستخرجة من منظمة العمل الدولية ILO.

الكلمات الرئيسية: تمثيل المفاهيم, تمثيل الكلمات, متعددة اللغات, تصنيف النصوص, mapping, WordNet.