وزارة التعليم العالي والبحث العلمي

BADJI-MOKHTAR UNIVERSITY–ANNABA-

UNIVERSITÉ BADJI MOKHTAR–ANNABA-

جامعة باجي مختار
-عنابة-

Faculté : Sciences de l'ingénieur – Année 2014 –

Département : Informatique

THÈSE

Présentation en vue de l'obtention du diplôme de doctorat

Identification d'opinions dans les textes arabes en utilisant les ontologies

Option : Texte, Parole et Imagerie

Par

Lazhar FAREK

DIRECTEUR DE THESE : Pr. Tlili-Guiassa Yamina

DEVANT LE JURY

PRESIDENT : Mme. Ghoualmi-Zine Nacira	Pr. Université Badji-Mokhtar - Annaba
EXAMINATEURS :	
Pr. Laskri Mohmed Tayeb	Pr. Université Badji Mokhtar – Annaba
Pr. Laouar Reda	Pr. Université Larbi Tébessi – Tébessa
Dr. Lafifi Yacine	MC. Université 08 Mai 1945 – Guelma
Dr. Redjimi Mohamed	MC. Université 20 Août 1955 – Skikda

Dédicace

Je dédie ce travail :

À tous les membres de ma famille,

À ma femme,

À mes collègues de travail,

À tous ceux qui, de près ou de loin, ont soutenu mon travail par leurs encouragements et conseils.

Remerciements

Merci Allah, le Tout Puissant.

Je tiens à exprimer ma profonde reconnaissance à la directrice de cette thèse, Mme. Tlili-Guiassa Yamina, pour m'avoir dirigée, guidée, conseillée, pendant tout le déroulement de ce travail, dont j'espère être à la hauteur de ses attentes.

Mes remerciements vont également à ma femme Dr. A. Benaidja pour son encouragement et soutien.

Je tiens à remercier aussi tous les membres du jury qui me font l'honneur d'examiner mon travail.

ملخص

في هذه الأطروحة، نقدم طريقة تعتمد على الاستكشاف الأنطولوجي لتحديد الآراء الواردة في النصوص العربية، مع العلم أن الرأي هو وجهة نظر، تعبير عن المشاعر أو تقييم صريح أو ضمني تجاه كائن ما أو أحد خصائصه .

الآراء الصريحة يمكن استخراجها بواسطة الإسقاط المباشر للمفاهيم الأنطولوجية على النص. إلا أن الآراء الضمنية تحتاج إلى استكشاف عميق للطبقة الدلالية للأنطولوجيا، واستغلال العلاقات بين المفاهيم، الأفراد والصفات .

الكائن الذي يتم تعبير الرأي تجاهه، يمكن أن يتكون من عدة كائنات فرعية، الكائن الفرعي نفسه يمكن أن يتكون من أجسام أخرى فرعية، وهكذا، مع العلم أن الرأي يمكن التعبير عنه تجاه كائن أو أحد خصائصه. وبالتالي جاءت فكرة تصور المجال المدروس باستخدام الأنطولوجيات .

لربط الميزات المستخرجة مع تعابير الآراء، تستند طريقتنا على استخدام الارتباطات النحوية لاستخراج الأزواج ميزة–رأي، من خلال اعتماد مجموعة من القواعد اللغوية .

نعتمد في تصنيف الآراء الى إيجابية وسلبية على استخدام ال (SVM)كأسلوب للتلقين الذاتي .

وأدت النتائج المتحصل عليها الى اكتشاف العوامل التي أثرت على أداء النظام والتي ستكون موضوعات بحث لتحسينات ممكنة في أبحاثنا المستقبلية.

الكلمات الرئيسية: التنقيب في الرأي، تحديد الهوية، الأنطولوجيا، التصنيف، النص، العربية، النحوية التبعية، الرأي الصريح، الرأي الضمني.

Résumé

Dans ce manuscrit, nous présentons une approche basée sur une exploration ontologique pour identifier les opinions exprimées dans les textes arabes, sachant qu'une opinion est un point de vue, une émotion ou une évaluation exprimée explicitement ou implicitement sur un objet quelconque ou sur l'une de ses caractéristiques.

Les opinions explicites peuvent être extraites par projection directe des concepts ontologiques sur le texte. Cependant, les opinions implicites ont besoin d'une exploration profonde de la couche sémantique de l'ontologie, en exploitant les relations entre les concepts, les individus et les attributs.

Un objet sur lequel une opinion est exprimée, peut être composé de plusieurs sous-objets, chaque sous-objet lui-même peut être composé d'autres sous-objets, et ainsi de suite, sachant qu'une opinion peut être exprimée sur un objet ou une de ses propriétés. De ce fait, vient l'idée de conceptualiser le domaine étudié en utilisant les ontologies.

En vue d'associer les caractéristiques extraites avec leurs mots d'opinion, notre approche repose sur l'utilisation des dépendances grammaticales pour extraire l'ensemble des couples caractéristique-opinion, en combinant une liste de règles linguistiques.

La tâche de classification d'opinions en positive et négative est guidée par les Machines à Vecteurs de Support (*ang. Support Vector Machines SVM*) comme une technique d'apprentissage supervisé.

Les résultats obtenus nous ont mené à découvrir les facteurs qui ont influencé la performance de notre système, qui seront sujets d'éventuelles améliorations dans nos futurs travaux de recherche.

Mots-clés : fouille d'opinion, identification, ontologie, langue arabe, texte, classification, dépendance grammaticale, opinion explicite, opinion implicite.

Abstract

In this manuscript, we present an approach based on an ontological exploration to identify opinions expressed in Arabic texts, knowing that an opinion is a point of view, an emotion or an assessment explicitly or implicitly expressed on an object or on one of its characteristics.

Explicit opinions can be extracted by direct projection of the ontological concepts on the text. However, implicit opinions need a deep exploration of the semantic ontology layer, exploiting the relationships between concepts, individuals and attributes.

An object on which an opinion is expressed, can be composed of several sub-objects, each sub-object itself can be composed of other sub-objects, and so on, knowing that an opinion can be expressed on an object or one of its properties. Thus comes the idea of conceptualizing the studied domain using ontologies.

To associate the extracted features with their opinion expressions, our approach is based on the use of grammatical dependencies to extract feature-opinion pairs, by combining a list of linguistic rules.

The task of classifying opinions into positive and negative is guided by Support Vector Machines (SVM) as a supervised learning technique.

The obtained results led us to discover the factors that influenced the performance of our system to be subjects to possible improvements in our future researches.

Keywords: opinion-mining, identification, ontology, Arabic, text, classification, grammatical dependency, explicit opinion, implicit opinion.

Liste de Figures

Liste des Tableaux

Table des matières

Chapitre 1

Introduction

1.1. Problématique de recherche

Cette thèse présente une contribution à deux domaines : la fouille de données d'opinion (*ang. Opinion-Mining*), et l'ingénierie des connaissances (IC), dont l'objectif est de concevoir et de développer un système capable d'identifier les opinions dans les textes et plus précisément dans les textes arabes de différentes sources : forums de discussion, sites de e-commerce, blogs, journaux, etc., dans un cadre ontologique conceptualisant les connaissances du domaine étudié.

L'Opinion-Mining (OM) ou l'analyse des sentiments (*ang. Sentiment Analysis*), également appelé analyse de subjectivité (*ang. Subjectivity Analysis*), est une discipline en plein essor, réunit plusieurs disciplines : Traitement automatique des langages naturels (TALN), recherche d'information (RI), et la linguistique, dont l'objectif principal est d'extraire des opinions à partir des textes et de classifier ces opinions selon leurs orientations sémantiques en deux catégories principales : positive et négative. La tâche d'identification d'opinions à partir des textes nécessite une analyse lexicale et syntaxique profonde, surtout dans notre cas, où la langue traitée est l'arabe, cette dernière, qui se caractérise par une morphologie complexe, présente l'un des grands challenges que nous devons faire face.

L'ingénierie des connaissances propose des concepts, des méthodes et des techniques permettant d'acquérir, de modéliser, de formaliser, et d'opérationnaliser des connaissances mobilisées dans certains domaines. La représentation des connaissances est un système définissant une série de classes et une série de propriétés qui relient les classes. Dans le domaine du Web sémantique par exemple, un concept correspond à une classe et une relation correspond à une propriété. Ces classes et les relations entre eux forment ce que l'on appelle

ontologie informatique, permettant de représenter précisément un corpus de connaissances sous une forme utilisable par la machine. Les ontologies informatiques représentent un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques et/ou des relations de composition et d'héritage (au sens objet).

La quantité énorme des informations disponibles notamment sur le Web rend aujourd'hui les systèmes actuels chaque jour plus préoccupés par la manière de gérer la surcharge d'information, en s'assurant que l'utilisateur aura l'accès aux meilleures sources avec le moindre effort. Ces dernières années, une attention particulière a été donnée pour résoudre ce problème. Le secteur de e-commerce est l'un des plus influencé par la quantité de données produites par les clients sous forme de commentaires et critiques sur les produits commercialisés, cette augmentation a été remarquée surtout après l'apparence du Web 2.0.

Actuellement, les ontologies constituent un enjeu stratégique dans la représentation et la modélisation des connaissances. Récemment, elles ont été introduites pour formaliser les connaissances dans les systèmes experts. Elles définissent les primitives indispensables pour leurs représentations, ainsi que leurs sémantiques dans un contexte particulier [14].

Les opinions des clients représentent une source d'information qui ne devraient pas être maltraitée ou ignorée par la communauté de recherche. Ce travail souligne la nécessité de développer une approche capable de profiter pleinement de ces données, dont l'objectif est d'exploiter les sources d'opinions en vue de faciliter la prise de décision plus rapidement avec le moindre coût.

En Opinion-Mining, les travaux de recherche peuvent être catégorisés en deux principaux sous-thèmes : la classification d'opinions au niveau de la caractéristique et la classification d'opinion au niveau du document. La différence entre ces deux niveaux est expliquée dans [139] comme suit : la classification d'un document en positive ou négative est basée sur le sentiment global exprimé par le détenteur de l'opinion (*ang. Opinion holder*). Cependant, l'objectif de la classification au niveau de la caractéristique est de produire un résumé plus ou moins détaillé sur les caractéristiques de l'objet sur lequel une opinion a été exprimée. En effet, l'opinion globale prévue exprimée dans un document dépend sur les polarités positives et négatives (orientations sémantiques) des expressions d'opinion utilisées dans le document.

Dans un commentaire ou une critique, les opinions peuvent être exprimées de manière explicite comme dans les exemples suivants : « Un bon travail » (1), « Une mauvaise habitude » (2), tandis que dans d'autres, il est difficile de localiser l'entité sur laquelle l'opinion est exprimée comme dans les exemples suivants : « Ahmed a une bonne note » (3), « Le mauvais service nous a obligé de quitter le complexe touristique » (4). Dans les exemples (1) et (2), il est clair de constater que les mots d'opinion « bon » et « mauvais » sont respectivement exprimés sur les deux entités « travail » et « habitude » qui donnent respectivement une opinion positive envers « travail » et une opinion négative envers « habitude ». Dans l'exemple (3), deux entités sont présentes « Ahmed » et « note » et un seul mot d'opinion « bon », il est facile de trouver en utilisant les outils de TALN que l'adjectif « bon » est lié à l'entité « note », si une opinion positive est exprimée envers « note », intuitivement, nous pouvons déduire que l'opinion positive est exprimée implicitement sur l'entité « Ahmed » parce que c'est lui qui a obtenu la « note ». La même observation pour l'exemple (4) où l'opinion négative exprimée sur l'entité « service » en utilisant l'adjectif « mauvais », une opinion est implicitement exprimée sur l'entité « complexe touristique », parce que les deux entités « complexe touristique » et « service » sont sémantiquement dépendantes.

Dans ce manuscrit, nous présentons une approche d'identification d'opinion, basée sur une exploration ontologique des textes arabes. Cette approche vise à étudier le rôle des ontologies et leurs contributions pour extraire les caractéristiques (*ang. features*) sur lesquelles des opinions ont été exprimées. Les ontologies sont utilisées pour conceptualiser les connaissances du domaine étudié, et pour résoudre partiellement l'ambiguïté sémantique des concepts inter-domaines. Pour identifier les expressions d'opinion, nous utilisons un lexique de sentiments, où chaque mot de ce lexique peut être un adjectif, un verbe ou un adverbe éventuellement combiné avec certains modificateurs tels que la négation et la confirmation, et pour chaque mot de ce lexique, une polarité positive ou négative est associée, avec des taux exprimant le degré de positivité et négativité.

1.2. Motivation

Du point de vue des clients, donner une considération aux opinions des autres avant d'acheter un produit est un comportement commun, même avant l'existence de l'Internet. Dans l'ère numérique, la différence est que le client a un accès à des milliers d'opinions, ce qui améliore grandement la prise de décision. Fondamentalement, les clients veulent trouver le meilleur

produit avec le prix le plus bas. En d'autres termes, ils cherchent des produits qui satisfont leurs besoins avec des prix dont ils sont capables de payer.

Il est important de souligner que le but d'analyser les opinions des autres vient de leur caractère neutre, qui ne sont habituellement pas liés à une organisation ou une entreprise parce qu'elles représentent la voix des consommateurs ordinaires, et qu'elles diffèrent grandement des annonces (les annonces sont généralement biaisées et ont une tendance à favoriser le produit, en soulignant les aspects positifs et en dissimulant les négatifs) [20].

Du point de vue du commerce électronique (*ang. e-commerce*), recevoir les commentaires des consommateurs peut grandement améliorer ses stratégies afin d'augmenter les bénéfices du secteur. Par exemple, une boutique en ligne peut placer des annonces en mesurant le niveau de satisfaction des consommateurs pour un produit donné.

Il est fréquent de trouver des produits avec des milliers d'opinions, donc, il pourrait être une tâche difficile pour un client d'analyser chacune d'eux. En outre, il pourrait être un travail très fastidieux de trouver des opinions sur quelques caractéristiques d'un produit, généralement, lorsqu'il s'agit d'une exigence d'un client expérimenté.

Une différence importante rend les techniques actuelles de classification pas assez efficaces pour décrire les informations représentées par les opinions. Cette différence est principalement due à la nature de l'information textuelle dans le monde. Ces informations sont soient des faits ou des opinions. Les systèmes de recherche actuels sont axés sur les faits (par exemple, des mécanismes de classement utilisés par les moteurs de recherche). Un fait est habituellement égal à tous les mêmes autres faits. Cependant, une opinion est une croyance ou un jugement sur un sujet ou un objet. Par conséquent, une opinion exprimée sur un objet est généralement différente des autres opinions exprimées sur le même objet. Dans ce sens, un mécanisme de résumé est indispensable pour rendre les opinions facilement exploitables.

Créer des systèmes capables de traiter les informations subjectives, exige effectivement de surmonter un certain nombre de nouveaux challenges [21]. Ces challenges sont présentés dans les travaux de Pang et Lee (2008) [21], où les auteurs, ont montré un exemple concret d'un moteur de recherche des opinions et des critiques. D'après les auteurs, une telle application permettait de combler un besoin d'information important.

D'après les auteurs, le développement d'une application complète de recherche d'opinions et de critiques pourrait impliquer de maîtriser les problèmes suivants :

- Si l'application est intégrée dans un moteur de recherche polyvalent, alors on aurait besoin pour déterminer si l'utilisateur voulait chercher des informations factuelles ou subjectives. Cela peut et ne pas être un problème difficile en soi : peut-être des requêtes avec des indicateurs comme « opinion », « critique », peuvent indiquer s'il s'agit d'une recherche subjective, mais en général la classification des requêtes est un problème difficile qui se pose à son tour. En effet, il était un sujet d'étude dans [22].

- Difficulté de déterminer les documents pertinents pour une requête orientée-opinion, ainsi, la difficulté de déterminer les portions de documents contenant les informations subjectives. Parfois, cela est relativement facile, comme dans les textes récupérés à partir des sites où les critiques et les commentaires des utilisateurs sont présentés dans un format organisé, contrairement aux blogs par exemple qui contiennent moins d'informations subjectives.

- Une fois les documents cibles sont récupérés, on est encore devant un autre problème qui est l'identification de l'opinion globale exprimée dans le document et/ou les opinions spécifiques exprimées envers des caractéristiques particulières ou envers un concept des items ou des sujets en question.

- Certains sites font ce genre d'extraction plus facile, par exemple, il y a des sites qui spécifient un ensemble prédéfini de caractéristiques à commenter, tandis que les textes libres peuvent être difficiles à analyser par les machines. De même, ils peuvent poser un challenge supplémentaire, par exemple, si les citations sont incluses dans un article de journal, il faut prendre soin d'attribuer les points de vue exprimés dans chaque citation à l'entité convenable.

- Enfin, le système doit présenter l'information subjective qu'elle a recueillie sous forme de résumé compréhensible. Cela peut impliquer une partie ou l'ensemble des actions suivantes :

    a) Agrégation des « votes » qui peuvent être enregistrés sur différentes échelles (par exemple, un utilisateur utilise un système d'étoiles, mais d'autres utilisateurs utilisent des notes en lettres) ;

b) Choix sélectif de certaines opinions ;
c) Représentation des points de désaccords et les points de consensus ;
d) Identification des détenteurs de l'opinion ;

Noter que cela pourrait plus approprié de produire une représentation visuelle des opinions plutôt qu'une représentation textuelle. Cette dernière est habituellement utilisée dans les résumés thématiques multi-documents.

1.3. Contexte de travail

Le premier axe de notre travail est la construction d'une ontologie qui va cueillir les concepts du domaine étudié, elle va donner une représentation formelle des concepts ainsi que des différentes relations qui relient ces derniers. En effet, la construction des ontologies souligne une difficulté même pour un spécialiste de domaine [16], il existe trois approches pour concevoir une ontologie : l'approche ascendante (Up-Down) [17], l'approche descendante (Top-Down) [19] et l'approche mixte (Middle-Out) [18]. Nous allons détailler ces approches dans le chapitre 4 du présent manuscrit. Dans notre travail, nous avons utilisé une approche (Middle-Out) en commençant par l'identification des concepts les plus importants pour trouver les concepts les plus génériques et les plus spécifiques dont on aura besoin.

Le deuxième axe de notre recherche est la définition d'une méthode de construction d'une base lexicale arabe à partir des textes du domaine étudié, permettant d'appréhender de gros volumes de données qui seront utilisées dans la phase d'identification d'opinions. Ces connaissances sont des mots ou des expressions à caractère purement subjectif qui reflètent des sentiments, des opinions, des émotions ou des états psychologiques.

Le lexique comporte des milliers de mots de sentiments, des émotions et des états psychologiques, tels que des noms comme حب (amour) ou خوف (peur), des verbes comme يحب (aimer), يرعب (effrayer), et des adjectifs comme محب (amoureux) ou حسود (jaloux). Dans ce contexte nous proposons une classification dans laquelle ces mots sont en plusieurs classes homogènes, chaque classe est nommée par le sentiment ou l'état psychologique décrit, comme la classe خوف (peur), qui contient les mots relatifs à un sentiment de peur (خوف - peur, خشية - crainte, ذعر - frayeur, يرعب - effrayer, مرعب - effrayant, etc.). Nous distinguons deux catégories de mots :

a) Les mots de polarité négative qui décrivent un sentiment ou un état psychologique plutôt désagréable, comme la peur ou la colère.

b) Les mots de polarité positive qui décrivent un sentiment ou un état psychologique plutôt agréable, comme l'amour ou l'amusement.

Le troisième axe est la conception d'une approche permettant l'identification d'opinions à partir des textes, en se basant sur l'ontologie du domaine construite dans la première phase de notre travail et le lexique de sentiments, construit dans la deuxième phase, puis la classification des opinions identifiées en positive ou négative, tout en utilisant des techniques symboliques pour la fouille de la subjectivité et les unités élémentaires d'opinion.

1.4. Organisation du manuscrit

Y compris ce chapitre, ce travail est divisé en six (6) chapitres :

- **Chapitre 2 : Fouille de données d'opinion**

À travers ce chapitre, nous présentons les différents concepts utilisés en fouille de données d'opinion : subjectivité, polarité, intensité d'opinion, ainsi, la définition des différents concepts comme : opinion, sentiment, émotion, opinions implicites et explicites et le lien entre eux, etc.

Dans ce chapitre, nous présentons aussi le problème de classification d'opinions et les différentes techniques et algorithmes utilisés. Nous présentons aussi, les applications de la fouille d'opinions, et sa contribution efficace dans la prise de décision.

- **Chapitre 3 : Travaux connexes**

Les travaux connexes que nous présentons dans ce chapitre, dans le domaine de fouille de données d'opinions nous mènent à découvrir les points faibles et les points forts de chaque approche utilisée, les difficultés rencontrées, en vue de donner plus de motivation envers notre approche.

- **Chapitre 4 : Ingénierie ontologique**

Nous présentons dans ce chapitre un aperçu général sur les ontologies, leurs différents types, leurs caractéristiques, le rôle et l'avantage d'utilisation des ontologies, nous présentons aussi les démarches et les différents outils utilisés pour construire une ontologie.

Dans ce chapitre, nous donnons une importance aux ontologies de domaine et leurs contributions dans la phase de conceptualisation et représentation des concepts du domaine étudié.

- **Chapitre 5 : ONTOMART : Notre approche proposée**

Dans le cadre de ce chapitre, nous proposons une approche basée-ontologies, écourtée ONTOMART (**ONT**ology-based **O**pinion **M**ing for **AR**abic **T**exts), pour l'identification des opinions à partir des textes arabes, nous présentons l'architecture générale de notre approche en détaillant chaque module à part, en partant de la sélection des données jusqu'à la phase de classification des opinons identifiées.

- **Chapitre 6 : Implémentation et évaluation**

C'est dans ce chapitre, que nous présentons l'implémentation et l'évaluation de notre approche, en concluant par un résumé regroupant les difficultés rencontrées, les points faibles et les points forts de notre approche.

Enfin, une conclusion donnant un aperçu synthétique sur notre approche et un ensemble de perspectives sur les facteurs qui ont influencé la performance de notre système, et qui seront sujets d'éventuels futurs travaux de recherche.

Chapitre 2

Fouille de données d'opinion

2.1. Introduction

Dans leur livre intitulé « *Opinion Mining and Sentiment Analysis* » [21], Pang et Lee (2008) exposent les intérêts socio-économiques d'un moteur de recherche hypothétique qui répondrait à la requête « Que pensent les gens de ... ? ».

« Que pensent les autres ? » a toujours été un élément d'information important pour la plupart d'entre nous pendant le processus de prise de décision. Avant l'apparence des réseaux sociaux et les sites de commerce électroniques sur le web, beaucoup d'entre nous demande à ses amis ou sa famille de recommander un produit ou un service avant la prise de décision. Mais avec le développement du Web 2.0, la prise de décision est devenue plus ou moins facile qu'avant, l'information est devenu à la portée de tout le monde où les utilisateurs disposent de nombreuses et nouvelles facilités (forums, blogs, groupes de discussions, etc.) pour exprimer leurs opinions sur des différents sujets, produits ou services.

La quantité énorme d'informations portant des opinions et des critiques disponibles pose un problème du coût et du temps pour l'explorer, d'où il est devenu indispensable de proposer de nouvelles approches afin d'extraire automatiquement les opinions exprimées. Pour répondre à ce besoin, de nombreux travaux de recherche sont apparus. Ils sont issus de différents domaines : fouille de données, aide à la décision, modélisation des connaissances, traitement automatique des langues ou linguistique.

Ce chapitre donne un aperçu général sur le problème de fouille d'opinions, en présentant les différents concepts utilisés, les axes de recherche, et le problème de classification d'opinions.

Des concepts assez compliqués sont exposés dans ce chapitre, comme les opinions implicites et explicites exprimées sur des caractéristiques (*ang. features*) implicites et explicites, illustrés par des exemples.

2.2. Complexité de la notion d'opinion

Selon la définition citée dans [93], les opinions sont généralement des expressions subjectives qui décrivent les sentiments, les évaluations des gens vers des entités, des événements et leurs propriétés.

La différence entre « fait » et « opinion » est qu'un fait est quelque chose qui est empiriquement vrai et peut-être étayé par des preuves, tandis qu'une opinion est une croyance qui peut ou peut ne pas être sauvegardée avec un certain type de preuves. Une opinion est peut aussi être définie comme une déclaration subjective qui peut être le résultat d'une émotion ou une interprétation individuelle d'un fait. Par exemple, les différences biologiques entre mâles et femelles sont un fait, une préférence pour un sexe sur l'autre est une opinion.

Pour démontrer la complexité de la notion d'opinion, nous allons nous baser sur un ensemble de critiques[1] exprimées par des clients ont déjà visité un restaurant à la ville de Montréal (Canada). Les exemples sont les suivants, pris sous forme de captures écran :

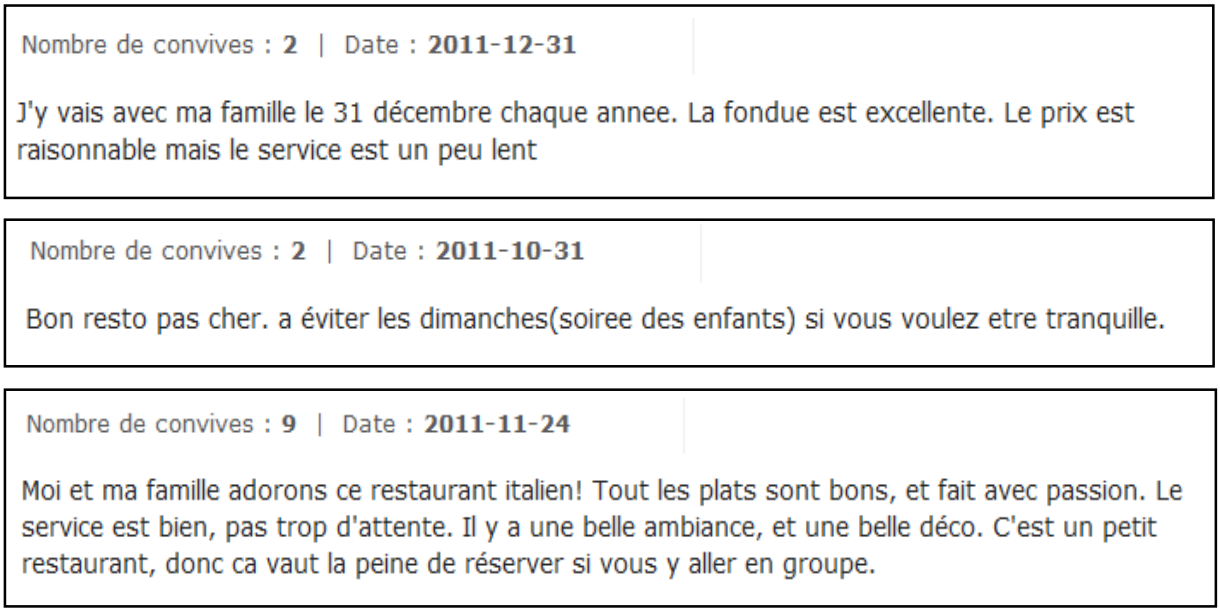


Figure 2.1 Ensemble de commentaires collectés du site web *restomontreal.ca*

[1] http://www.restomontreal.ca/restaurant-reviews/

Dans la première critique, il y a trois phrases, la première phrase *«J'y vais avec ma famille chaque année »* est objective, car elle ne porte aucun mot porteur d'opinion, tandis que les deux dernières phrases *« La fondue est excellente »* et *« Le prix est raisonnable mais le service est un peu lent »* sont subjectives à cause de la présence des mots et d'expressions porteurs d'opinions : *excellente*, *raisonnable* et *un peu lent*. La présence de deux polarités opposées dans la dernière phrase « *le prix est raisonnable mais le service est un peu lent* » qui rend difficile à extraire l'orientation globale dans cette phrase.

Dans la deuxième critique, dans la première phrase « *Bon resto pas cher* », il est facile de déduire que l'orientation sémantique globale est positive parce que l'auteur a utilisé que des mots et des expressions positives « bon » et « pas cher ». Dans la deuxième phrase « *à éviter les dimanches (soirée des enfants) si vous voulez être tranquille* » : le client donne un conseil aux autres clients d'éviter ce restaurant les dimanches pour qu'ils soient tranquilles. Bien que le verbe « éviter » ait une polarité négative, la phrase porte une opinion positive, parce que l'auteur a spécifié que le bruit causé par la soirée des enfants est seulement les dimanches. Ce genre de phrases nécessite une analyse profonde pour extraire correctement leurs orientations globales.

Dans la troisième critique, « *Moi et ma famille adorons ce restaurant italien ! Tous les plats sont bons, et fait avec passion. Le service est bien, pas trop d'attente. Il y a une belle ambiance, et une belle déco. C'est un petit restaurant, donc ça vaut la peine de réserver si vous y aller en groupe.* ». L'orientation sémantique globale est claire, toutes les expressions et les mots porteurs d'opinions sont positifs : *adorer, bon, pas trop d'attente, belle ambiance, belle déco*, sont des porteurs d'opinions positives, mais, comment nous pouvons savoir que la phrase « *C'est un petit restaurant* » porte une opinion positive, sachant que seule porteur d'opinion dans cette phrase est l'adjectif « *petit* » ? Comme l'auteur du commentaire n'a pas utilisé des modificateurs qui changent la direction de l'opinion comme *mais, pourtant, tandis que*, etc., nous pouvions facilement déduire que « *petit* » porte une polarité positive en appliquant la pseudo-règle linguistique des conjonctions intra-phrase (nous détaillons les règles de conjonction dans le chapitre 5 du présent manuscrit).

En général, les sentiments et la subjectivité sont très sensibles au contexte [52,53] et dépendent du domaine. La dépendance du domaine est en partie une conséquence des changements de vocabulaire, par exemple, la même expression peut indiquer différents

sentiments dans différents domaines. L'exemple suivant, cité dans [139], illustre la dépendance de domaine : Dans le domaine électronique, « long » peut avoir une polarité positive, comme dans le commentaire « the battery life of Camera X is long». Tandis que dans le domaine informatique, il peut avoir une polarité négative, comme dans « Program X takes a long time to complete ». Dans [139], les auteurs ont mentionné aussi qu'il n'est pas facile de construire des lexiques de sentiments pour tous les domaines d'intérêt.

De plus sur Internet, chacun utilise son propre vocabulaire, ce qui rend la tâche plus difficile même s'il s'agit du même domaine. En plus, il est très difficile d'affecter correctement le poids pour des phrases de la critique. Très souvent nous avons une description très positive d'un restaurant, avec la meilleure nourriture, le bon accueil, la propreté, etc., mais une phrase comme « *Malgré tout ça j'ai quitté le restaurant avant de terminer mon repas* » peut changer toute l'opinion en négative bien que toutes les autres phrases soient positives.

Ces exemples montrent qu'il est encore impossible d'arriver à un cas idéal de notation des sentiments dans un texte écrit par les divers utilisateurs. Car cela ne respecte aucune règle et il est impossible de prévoir tous les cas possibles, en plus très souvent la même phrase peut être considérée comme positive pour une personne et négative pour une autre.

2.3. Facteurs de difficulté de la fouille d'opinion

La fouille d'opinions est un problème de traitement automatique de textes. Dans cette section, nous distinguons ce problème avec la fouille de textes « classique » qui ne s'intéresse pas à l'opinion portée par les textes. Dans [47], l'auteur explique que le succès vienne à l'exploitation de plus en plus poussée de l'apprentissage artificiel dans le traitement de l'information, notamment la catégorisation de textes. Une autre explication est l'avènement de l'internet social et les intérêts socio-économique qu'il soulève.

Comme nous l'avons mentionné dans l'introduction de ce manuscrit, ce domaine est connu sous les noms de *opinion-mining*, *sentiment analysis* ou encore *subjectivity analysis*. Il s'intéresse au traitement automatique des opinions, des sentiments et de la subjectivité dans les textes. À titre de rappel, les informations textuelles sont, en général, soit des faits, soit des opinions. Tandis qu'un fait est une expression objective, une opinion est subjective, et décrit

les sentiments ou l'évaluation d'une personne, à propos d'un objet d'intérêt ou de ses propriétés.

Dans [46], l'auteur décrit la fouille d'opinion comme un sous-domaine de la fouille de textes qui consiste à analyser des textes afin d'en extraire des informations liées aux opinions et sentiments. L'une des tâches de la fouille d'opinions, appelée *classification d'opinions*, a pour objectif de classer les textes suivant l'opinion qu'ils expriment. Cette classification peut se faire sur deux classes (positif ou négatif), sur trois classes (positif, négatif ou neutre) ou sur plus de classes encore.

Dans [21] et d'après les auteurs, la fouille d'opinion peut être divisée en trois sous-domaines :

- *L'identification des textes d'opinion*, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de textes selon qu'ils sont objectifs ou subjectifs ;

- *Le résumé d'opinions*, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions contenant les opinions), chiffré (pourcentage, note), graphique (histogramme) ou encore image (thermomètre, étoiles, pouce levé ou baissé...) ;

- *La classification d'opinions*, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

L'opinion-mining est certainement un domaine encore jeune mais son exploitation présente de nombreux problèmes techniques et malgré les améliorations depuis les débuts des recherches, ils n'ont toujours pas été résolus de manière convaincante.

Parmi les facteurs qui rendent la fouille d'opinion difficile, nous trouvons :

1) Hétérogénéité des données

Les données textuelles d'opinion sont hétérogènes. D'une part, les opinions sont parfois dans des critiques formatées, parfois présentées dans un format libre. Dans les sites d'évaluation tels que Epinions[2] et Amazon[3], les critiques se présentent dans un format qui présente les avantages et les inconvénients de manière très synthétiques, comme indiqué dans la figure 2.2.

D'autre part, dans un format libre, le langage utilisé pour exprimer une opinion rend parfois plus difficile l'extraction des opinions sur les blogs ou les forums que dans un article de journal en ligne [137], en effet un langage plus familier, des phrases grammaticalement incorrectes ou des expressions locales empêchent l'analyse correcte de ces opinions.

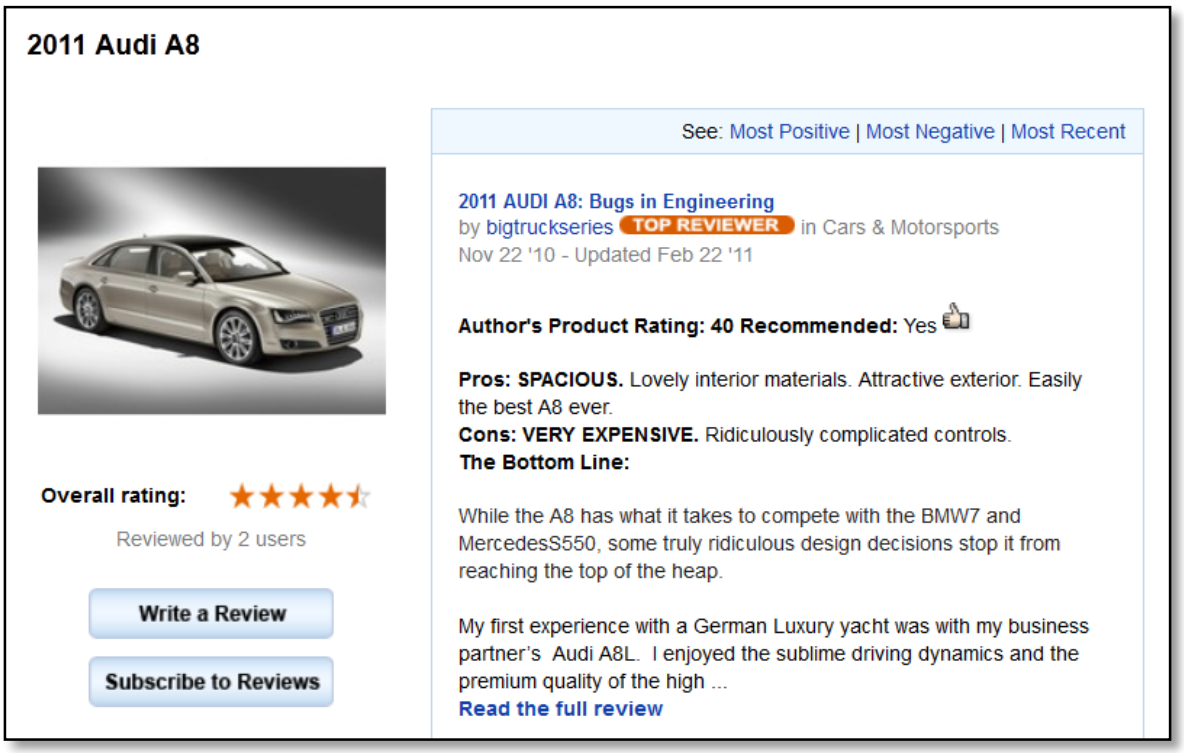


Figure 2.2 Exemple d'évaluation retrouvé sur *Epinions.com* dans un format de type Avantages et Inconvénients (Pros and Cons)

Même dans un article ou un commentaire exprimé dans un langage soutenu, une faute d'orthographe ne va pas permettre la reconnaissance du mot et peut engendrer des résultats erronés surtout si la faute se situe sur un mot particulièrement porteur d'opinion.

[2] www.epinions.com
[3] www.amazon.com

Figure 2.3 Ensemble de commentaires sur une vidéo, retrouvés sur *Youtube*[4]

Le sarcasme[5], l'ironie[6], le cynisme ou le langage figuré comme la métaphore qui influence le sens d'un texte sont complètement ignorés. Les figures 2.3 et 2.4 illustrent un ensemble de commentaires exprimés dans un langage de sarcasme et d'ironie.

Figure 2.4 Ensemble de commentaires retrouvés sur *Facebook*[7], exprimés en sarcasme

[4] www.youtube.com

[5] Le **sarcasme** désigne une moquerie ironique, une raillerie tournant en dérision une personne ou une situation.

[6] L'**ironie** désigne un décalage entre le discours et la réalité, entre deux réalités ou plus généralement entre deux perspectives, qui produit de l'incongruité.

2) Dépendance du contexte

En général, les sentiments et la subjectivité sont très sensibles au contexte. Les techniques existantes utilisent des mots d'opinion tels que « *grand* », « *étonnant* », « *pauvre* », « bon » ,« *mauvais* », etc., pour identifier l'orientation d'une opinion sur une caractéristique d'un produit. Bien que les orientations de ces mots soient évidentes, les orientations de nombreux autres mots dépendent du contexte.

Par exemple, le mot « *longue* » peut indiquer une opinion positive ou une opinion négative sur une caractéristique d'un produit dépendant de la caractéristique elle-même, comme indiqué dans l'exemple suivant :

« La durée de vie de la batterie de mon ordinateur portable est *longue*, c'est *magnifique* ».

« Mon ordinateur portable prend une *longue* durée pour démarrer, c'est un problème vraiment *gênant* ».

Dans la première phrase, l'adjectif « *longue* » porte une polarité positive alors que dans la deuxième porte une polarité négative.

Pour résoudre ce problème, plusieurs règles linguistiques ont été proposées dans [133] (ces règles sont expliquées dans le chapitre 5 de ce présent manuscrit). L'approche proposée essaie de déduire les orientations d'opinions sur une caractéristique du produit en utilisant le contexte. Les auteurs ont présenté aussi une fonction d'agrégation des opinions multiples dans une phrase. Les résultats des expériences montrent que les règles et la fonction d'agrégation sont très utiles.

3) Dépendance du domaine

La dépendance du domaine est en partie une conséquence des changements de vocabulaire, par exemple, la même expression peut indiquer différents sentiments dans différents domaines [21,48].

[7] www.facebook.com

4) Ambiguïté subjective

La principale difficulté qu'apporte le langage est le caractère ambigu des mots. L'ambiguïté, en fouille d'opinion, se situe plus au niveau de la subjectivité, qu'au niveau sémantique. Dans [45], par exemple, selon son contexte, le mot *grand* peut tantôt être factuel (par exemple, pour désigner la taille d'un individu), tantôt exprime une opinion (par exemple, la taille d'un mobile).

2.4. Terminologie

L'analyse des sentiments ou la fouille d'opinions est le traitement automatique des opinions, sentiments et émotions exprimées dans les textes. Nous utilisons le segment du texte suivant, il s'agit d'un commentaire sur une « voiture » pour comprendre les différents concepts utilisés (un nombre est associé à chaque phrase pour en faciliter la consultation) :

« (1) Hier, j'ai acheté une voiture (2) C'était vraiment cher (3) mais, elle était jolie (4) elle, m'a plait beaucoup (5) les chaises étaient confortables (6) Cependant, mon frère a dégouté ce modèle »

La question est : Qu'est-ce-que nous voulions fouiller ou extraire de ce texte ?

La première chose que nous pouvions noter est qu'il y a beaucoup d'opinions dans ce commentaire. Les phrases (3), (4) et (5) expriment des opinions positives, tandis que les phrases (2) et (6) expriment des opinions ou émotions négatives. Ensuite, nous remarquons que les opinions sont exprimées envers certaines caractéristiques ou objets. Les opinions dans les phrases (2), (3) et (5) sont sur la voiture comme un tout, et l'opinion exprimée dans la phrase (5) était sur les chaises.

Finalement, nous pouvions aussi noter que le détenteur (holder) de l'opinion dans les phrases (2), (3) et (4) est l'auteur de commentaire « moi », mais dans la phrase (6) c'est « mon frère »

Avec cet exemple, nous définissons d'une façon formelle l'analyse des sentiments et la fouille d'opinions. Généralement, les opinions peuvent être exprimées envers n'importe quoi : un produit, un service, un individu, une organisation, un événement, ou un sujet [20,23].

Notons qu'un objet peut avoir un ensemble de composants (parties) et un ensemble d'attributs (propriétés), et chaque composant peut avoir à son tour, ses propres sous-composants et son

ensemble d'attributs, et ainsi de suite. Ainsi, un objet peut-être décomposé hiérarchiquement selon la relation « est une partie de ».

1) Objet

Un *objet O* est une entité qui peut être un produit, une personne, un événement, une organisation ou un sujet. Il est associé à une paire *O(T, A)*, où *T* est une hiérarchie de *composants* (ou *parties*), *sous-composants*, et ainsi de suite, et *A* est un ensemble d'*attributs* de *O*. Chaque composant a son propre ensemble de sous-composants et attributs [21, 22, 30].

Exemple : Une marque particulière de téléphone cellulaire est un objet. Il a un ensemble de composants, par exemple, la batterie et l'écran, il a aussi un ensemble d'attributs, par exemple, la qualité de son, la taille et le poids. Le sous-composant « batterie » a aussi son ensemble d'attributs, par exemple, la durée de vie de la batterie, et la taille de la batterie, etc.

En se basant sur cette définition, un objet peut être représenté comme un arbre, une hiérarchie ou une taxonomie. La racine de l'arbre est l'objet lui-même. Chaque nœud non racine est un composant ou un sous-composant de l'objet. Chaque lien est une relation de type « est une partie de ». Chaque nœud est également associé à un ensemble d'attributs ou de propriétés. Une opinion peut être exprimée sur un nœud ou un attribut du nœud.

Suite à l'exemple précédent de la section 2.4, on peut exprimer une opinion sur le téléphone cellulaire lui-même (le nœud racine), par exemple, « *Je n'aime pas ce téléphone* », ou sur l'un de ses attributs, par exemple, « *La qualité de son de ce téléphone est mauvaise* ». De même, on peut aussi exprimer une opinion sur l'un des composants du téléphone ou tout autre attribut des composants.

En pratique, il est probablement trop complexe d'utiliser une représentation hiérarchique d'un objet et des opinions sur l'objet. Ainsi, pour des raisons de simplification et pour éviter toute éventuelle confusion, il est conseillé d'utiliser le terme *caractéristique* ou *trait* pour représenter à la fois les *composants* et les *attributs* [30].

Dans cette simplification, l'objet lui-même peut aussi être vu comme une *caractéristique* (mais une caractéristique spéciale), qui est la racine de l'arbre original. Un commentaire exprimé envers l'objet lui-même est appelée une opinion générale sur l'objet (par exemple, « J'aime ma voiture »). Une opinion sur une caractéristique spécifique est appelée opinion

spécifique sur une caractéristique de l'objet, par exemple, « *L'écran tactile de l'iPhone est vraiment souple* », où « *écran tactile* » est une caractéristique (composant) de l'iPhone.

L'utilisation du terme *caractéristique* est assez commune dans le domaine des produits parce que les gens utilisent souvent le terme *caractéristique du produit*. Toutefois, lorsque les objets sont des événements et des sujets, le terme caractéristique peut ne pas sembler naturel. En effet, dans certains autres domaines, les chercheurs utilisent également le terme *thème* [24] ou *aspect* [25, 26] pour signifier *caractéristique*.

2) Opinion exprimée sur une caractéristique

Une opinion exprimée sur une caractéristique *c* d'un objet *O*, évaluée dans un document *d*, est un groupe de phrases consécutives en *d*, qui exprime une opinion positive ou négative sur *c* [21].

Il est possible qu'une séquence de phrases (au moins une) dans un document véhiculant une opinion exprime une opinion sur un objet ou une caractéristique de l'objet. Il est également possible qu'une seule phrase exprime des opinions sur plus d'une caractéristique, par exemple :

« *La qualité de son de ce téléphone est bonne, mais la vie de la batterie est courte.* »

La plupart des recherches actuelles portent sur les phrases, c'est-à-dire, chaque passage est composé d'une seule phrase [21].

3) Caractéristique explicite et caractéristique implicite

Dans [93], si une caractéristique *c* ou l'un de ses synonymes apparaît dans une phrase *p*, *c* est appelée *caractéristique explicite* dans *p*. Si ni *c*, ni aucun de ses synonymes apparaissent dans *p*, mais *c*, est implicite, alors *c*, est appelée *caractéristique implicite* dans *p*.

Exemple :

« *La durée de vie de la batterie* » dans la phrase suivante est une caractéristique explicite :

« *La durée de vie de la batterie de ce téléphone est trop courte* »

« *La taille* » est une caractéristique implicite dans la phrase suivante comme il n'apparaît pas dans la phrase, mais il est implicite : « *Ce téléphone est trop grand* ».

Ici, « *grand* », qui n'est pas un synonyme de « *taille* », est appelé *indicateur de caractéristique*. Beaucoup d'indicateurs de caractéristiques sont des adjectifs et des adverbes. Certains adjectifs et adverbes sont généraux et peuvent être utilisés pour modifier quoique ce soit, par exemple, *bon*, *mauvais*, et *grand*, mais beaucoup d'indicateurs indiquent les caractéristiques qu'ils sont susceptibles de modifier, par exemple, *belle* (apparence).

4) **Détenteur de l'opinion** *(ang. Opinion holder)*

Le détenteur d'une opinion est la personne ou l'organisation qui exprime l'opinion.

Les détenteurs d'opinions sont également appelés sources d'opinion [27]. Dans le cas des critiques sur produits, les détenteurs d'opinions sont généralement les auteurs des commentaires postés. Les détenteurs d'opinions sont plus importants dans les articles des journaux parce qu'ils expriment souvent d'une manière explicite l'état de la personne ou l'organisation qui détient une opinion particulière [28, 29, 24]. Par exemple, le détenteur de l'opinion dans la phrase « *Ahmed a exprimé son désaccord sur le projet* » est « *Ahmed*».

5) **Opinion**

Une opinion sur une caractéristique *c*, est un point de vue positif ou négatif, une attitude, une émotion ou une évaluation sur *c* d'un détenteur d'opinion [23, 30].

6) **Orientation d'une opinion**

L'orientation d'une opinion sur une caractéristique *c* indique si l'opinion est positive, négative ou neutre [21,23].

L'orientation de l'opinion est également connue sous le nom *l'orientation du sentiment*, la *polarité de l'opinion*, ou *l'orientation sémantique* [21, 30].

7) **Modèle d'un objet**

Un objet *O* est représenté par un ensemble fini de caractéristiques, $F = \{f_1, f_2, ..., f_n\}$, ce qui inclut l'objet lui-même comme une caractéristique spéciale. Chaque caractéristique $f_i \in F$ peut

être exprimée avec un quelconque ensemble fini de mots ou de phrases $W_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$, qui sont des synonymes de la caractéristique [30], ou indiqués par un quelconque ensemble fini d'indicateurs de caractéristiques $Ii = \{i_{i1}, i_{i2}, ..., i_{iq}\}$ de la caractéristique.

8) Modèle d'un document d'opinions

Un document général d'opinions *d* contient des opinions sur un ensemble d'objets $\{O_1, O_2, ..., O_Q\}$ à partir d'un ensemble de détenteurs d'opinions $\{h_1, h_2, ..., h_p\}$. Les opinions sur chaque objet o_j sont exprimées sur un sous-ensemble de caractéristiques de F_j *de* l'objet O_j. Une opinion peut être l'un des deux types suivants :

- **L'opinion directe**

Une opinion directe est un quintuple $(O_j, c_{jk}, oo_{ijkl}, h_i, t_l)$, où O_j est un objet, c_{jk} est une caractéristique de l'objet O_j, oo_{ijkl} est l'orientation ou la polarité de l'opinion sur la caractéristique c_{jk} de l'objet O_j, h_i est le détenteur de l'opinion et t_l est le temps où l'opinion est exprimée par h_i. L'orientation de l'opinion oo_{ijkl} peut être positive, négative ou neutre (ou mesurée sur une échelle plus fine pour exprimer des opinions différentes de différentes densités (forces) [34]. Pour la caractéristique c_{jk} commentée par le détenteur d'opinion h_i, il / elle choisit un mot ou une expression de l'ensemble correspondant à des synonyme W_{jk}, ou un mot ou une phrase de l'ensemble des indicateurs de la caractéristique I_{jk} pour décrire la caractéristique, puis exprime une opinion positive, négative ou neutre sur la caractéristique.

- **L'opinion comparative**

Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs objets. Une opinion comparative est généralement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe, mais pas toujours.

Ce modèle de texte d'opinions couvre l'essentiel mais pas toutes les informations intéressantes ou tous les cas possibles. Par exemple, il ne couvre pas la situation décrite dans la phrase suivante : « *Le viseur et les lentilles de cette caméra sont trop proches* », qui exprime une opinion négative sur *la distance* entre les deux composantes.

Dans les opinions directes, il y a en fait deux principaux sous-types :

- Dans le premier sous-type, les opinions sont exprimées directement sur un objet ou sur les caractéristiques de l'objet, par exemple, « *La qualité de son de ce téléphone est bonne.* »

- Dans le deuxième sous-type, les opinions sont exprimées sur un objet en fonction de ses effets sur certains autres objets. Ce sous-type se produit souvent dans le domaine médical où les patients expriment des opinions sur les médicaments ou décrivent leurs effets secondaires. Par exemple, la phrase « *Après avoir pris ce médicament, mon genou gauche me sentait bien* », décrit un effet souhaitable de médicament sur le genou, et implique donc une opinion positive sur le médicament.

9) Emotions

Les émotions sont nos sentiments subjectifs et les pensées [23,30]. Les émotions ont été étudiées dans de nombreux domaines, par exemple, la psychologie, la philosophie, la sociologie, la biologie, etc. En se basant sur [36], les gens ont 6 types d'émotions primaires, i.e., *l'amour, la joie, la surprise, la colère, la tristesse et la peur*, qui peuvent être divisées en de nombreuses émotions secondaires et tertiaires. Chaque émotion peut aussi avoir des intensités différentes. Les forces d'opinions sont étroitement liées à l'intensité de certaines émotions, par exemple, la joie et la colère. Toutefois, les concepts d'émotions et d'opinions ne sont pas équivalents, bien qu'ils aient une grande intersection.

10) Subjectivité

Soit l'exemple suivant :

« (1) Samedi dernier, j'ai acheté un téléphone Nokia et mon collègue a acheté un téléphone Motorola. (2) Nous nous sommes appelés quand nous sommes arrivés à la maison. (3) Le son sur mon téléphone n'était pas si clair, pire que mon ancien téléphone. (4) La caméra était bonne. (5) Mon collègue était très heureux avec son téléphone. (6) Je voulais un téléphone avec une bonne qualité de son. (7) Donc mon achat a été une vraie déception. (8) hier, j'ai retourné le téléphone. »

L’exemple ci-dessus a révélé un autre problème appelé *la subjectivité*. C'est, dans un document, quelques phrases expriment des opinions et d'autres pas. Par exemple, les phrases (1), (2), (6) et (8) n'expriment aucune opinion. Le problème de la subjectivité a été largement étudié dans [37, 38, 39, 40].

11) Phrase subjective

Une phrase objective exprime quelques informations factuelles sur le monde, alors qu'une phrase subjective exprime des sentiments ou des croyances personnelles.

Par exemple, dans l'exemple cité dans la sous-section précédente, les phrases (1), (2) et (8) sont des phrases objectives, tandis que toutes les autres phrases sont des phrases subjectives. Les expressions subjectives prennent plusieurs formes, par exemple, *les opinions, les allégations, les désirs, les croyances, les soupçons et des spéculations* [41, 42]. Ainsi, une phrase subjective peut ne pas contenir une opinion. Par exemple, la phrase (6) de l'exemple de la sous-section précédente, est subjective, mais elle n'exprime pas une opinion positive ou négative sur n'importe quel téléphone spécifique. De même, nous devons également noter que pas toutes les phrases objectives ne contiennent aucune opinion, la deuxième phrase de l'exemple cité dans la sous-section 12 montre ceci.

12) Opinion explicite et opinion implicite

Une opinion explicite sur une caractéristique *c*, est une opinion exprimée explicitement sur *c* dans une phrase subjective. Une opinion implicite sur la caractéristique *c*, est une opinion sur *c* implicite dans une phrase objective.

Exemple :

La phrase suivante exprime une opinion positive explicite :

« *La qualité de son de ce téléphone est incroyable.* »

La phrase suivante exprime une opinion négative implicite :

« *L’afficheur s'est cassé en deux jours.* »

Bien que cette phrase énonce un fait objectif, il indique implicitement une opinion négative sur l'afficheur. En fait, la phrase (8) dans l'exemple de la section précédente, peut aussi être

déclarée à impliquer une opinion négative. En général, les phrases objectives qui impliquent des opinions positives ou négatives indiquent souvent les raisons de l'opinion.

13) Phrase d'opinions

Une phrase d'opinions est une phrase qui exprime explicitement ou implicitement des opinions positives ou négatives. Elle peut être une phrase subjective ou objective.

Comme nous pouvons le voir, les concepts *phrase d'opinions* et *phrase subjective* ne sont pas les mêmes, bien que les phrases d'opinions sont souvent un sous-ensemble de phrases subjectives. Les approches pour les identifier sont similaires. Ainsi, pour la simplicité de présentation, nous utilisons les deux termes indifféremment. La tâche de déterminer si une phrase est objective ou subjective est appelé *classification de subjectivité*.

2.5. Exemple d'application de la fouille d'opinions

Dans la section 1.2 du chapitre 1, nous avons cité l'exemple de Pang et Lee (2008), qui est une application de recherche d'opinions sur le web. Des exemples similaires à l'exemple de Pang et Lee ont été mis en pratique. Dans cette section, nous cherchons à montrer en exemple : Les mêmes capacités qu'un moteur de recherche orienté-opinion doive avoir, peuvent aussi servir comme une base pour la création des sites web pour l'agrégation des opinions des internautes. Parmi ces sites, nous citons le célèbre site Epinions.com, ce dernier intègre un moteur de recherche orienté-opinion.

Figure 2.5 Page d'accueil du site *Epinions.com*

Ce site offre à ses utilisateurs de poster leurs commentaires sur des produit et services, comme il permet d'effectuer des recherches d'opinions et de lire les commentaires des autres utilisateurs pour prendre une décision fondée sur les recommandations et les conseils postés.

2.6. Conclusion

Ce chapitre présente brièvement le domaine de la fouille de données d'opinion, ses problématiques, et la terminologie utilisée.

Les facteurs qui rendent difficile la fouille d'opinion comme l'hétérogénéité des données, la dépendance du domaine et du contexte, ainsi l'ambiguïté subjective, offrent un challenge de traitement automatique et des approches qui s'adaptent au niveau de granularité pour gagner en précision.

La classification des opinions est l'un des sujets le plus largement étudié dans le domaine de fouille d'opinions, dont une grande importance a été accordée à l'étude des techniques de classification qui seront présentées dans le chapitre suivant du présent manuscrit.

Chapitre 3

Travaux connexes

3.1. Introduction

L'analyse des textes en matière d'étude des sentiments ou d'opinions n'est pas récente [88,89]. La première apparition du terme Opinion-Mining, signifiant littéralement, fouille d'opinion, était en 2003 [90]. Depuis cette année-là, le domaine est devenu très actif avec différents challenges qui se sont créés et des centaines d'articles sur le sujet ont été publiés.

Dans [21], et d'après l'auteur, la fouille d'opinion peut être divisée en trois sous-domaines :

- l'identification des textes d'opinion, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de textes selon qu'ils sont objectifs ou subjectifs ;

- le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte ;

- la classification d'opinions, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

Ce sont clairement l'identification des textes d'opinion et la classification d'opinions qui sont les plus étudiées dans la littérature et sont également sujets à différents challenges [91,92].

Dans ce chapitre, nous nous intéresserons en premier axe à la tâche d'identification et de classification d'opinions, et en deuxième axe, aux travaux de recherche qui focalisent sur la langue arabe, dont beaucoup de travaux ont été effectués sur le sujet, et trois grandes catégories de méthodes peuvent être mises en avant : les approches basées sur la linguistique, les approches basées sur l'apprentissage automatique et les approches hybrides.

3.2. État de l'art sur la classification d'opinions

La fouille de textes est l'extraction de connaissances dans des textes en langage naturel.
La classification d'opinions est une tâche particulière de la fouille de textes et répond à un problème de traitement automatique de textes qui s'intéresse au traitement des opinions ou des sentiments et de la subjectivité dans les textes. Cette tâche peut être le sujet le plus largement étudié [37, 38, 39, 40, 41, 42], qui classifie un document d'opinions (par exemple, un commentaire sur un produit) pour définir s'il exprime une opinion positive ou négative.

La tâche est aussi connue sous le nom de *classification de sentiments au niveau des documents (ang. Document-level sentiment classification)*, car il considère l'ensemble des documents comme l'unité d'information de base [21]. Naturellement, elle peut également être appliquée à des phrases individuelles.

Dans cette section, nous décrivons les objectifs de la classification d'opinions, ainsi que ses intérêts, et les trois grandes approches de classification utilisées en fouille d'opinion : les approches linguistiques basées sur les lexiques, les approches statistiques basées sur l'apprentissage automatique et les approches hybrides fondées sur les deux premières.

— Objectifs et définition

Dans [93], et d'après l'auteur, l'opinion peut être définie comme l'expression des sentiments d'une personne envers une entité. L'opinion est subjective et peut être décrite avec certains attributs. L'attribut d'opinion le plus étudié est la polarité (positive, négative, et éventuellement neutre) qui définit si l'opinion est favorable ou défavorable. D'autres attributs sont l'intensité de l'opinion et le degré de subjectivité.

La classification de polarité d'opinion consiste donc à déterminer si un texte exprime des opinions positives ou négatives, quel que soit le sujet du texte ou les caractéristiques qui y sont discutées. Cette tâche porte sur des textes subjectifs, c'est à dire des textes contenant des opinions sur un sujet ou un objet tel que un film, une voiture, un téléphone portable ou encore une personnalité. Une définition plus formelle citée dans [93], de la classification d'opinion est donnée comme suit :

Soit C un ensemble de classes ordonnées représentant chacune un degré d'opinion et D un ensemble de textes subjectifs. La classification d'opinions consiste alors à trouver, pour tout $d \in D$*, les couples (d, c) tel que* $c \in C$*.*

Ce type de classification se fait généralement sur deux classes, "contient une opinion positive" ou "contient une opinion négative". On parle alors de classification binaire. Elle peut également se faire sur trois classes en ajoutant une classe de neutralité destinée aux textes porteurs des opinions neutres. On peut également envisager une classification sur un nombre de classes supérieur à trois afin de mieux préciser l'intensité de l'opinion, mais cela reste assez rare dans le domaine.

— **Complexité et caractéristiques de la tâche**

La classification d'opinions pourrait se comparer à une classification de textes comme les autres : « étant donné les mots présents dans le texte, j'en déduis une classe ». Par exemple, pour la classification par thème, il suffit de trouver un certain nombre de mots en lien, par exemple, au sport, à la politique ou encore au cinéma pour découvrir la classe du document. Seulement, en ce qui concerne les opinions, elles sont rarement exprimées par un seul mot. De plus, les discours contenant des opinions sont par définition des discours subjectifs et l'interprétation de la subjectivité est une tâche délicate. En effet, les opinions sont généralement exprimées par des verbes et des adjectifs qualificatifs qui possèdent une certaine hiérarchie suivant le degré de l'opinion qu'ils expriment. De plus, il existe la possibilité d'ajouter des adverbes afin de modifier la « force » de l'adjectif ou du verbe.

Outre les problèmes d'intensité d'opinion liés aux adjectifs et verbes, les opinions et les sentiments peuvent être exprimés de plusieurs façons. Ils peuvent par exemple être exprimés à l'aide des termes relevant de l'émotion. Ces termes liés à l'émotion telle que « triste », ou

« effrayant » sont difficiles à interpréter du point de vue de l'opinion si l'on ne connaît pas les goûts de l'auteur. Les goûts de l'auteur ont donc une importance mais le sujet également.

Dans le cas d'un film par exemple, le commentaire « Ce film est à mourir de rire » peut exprimer un avis très positif si l'objet visé est une comédie mais pas nécessairement s'il s'agit d'un film d'horreur ou d'un drame. Les avis peuvent également être exprimés sans employer un seul mot directement lié à l'opinion comme dans le commentaire « courrez vite, acheter le livre » [94]. Le traitement des négations semble également être une grosse problématique de la classification d'opinion [21].

3.2.1. Les approches linguistiques

La principale tâche dans cette approche est la conception de lexiques (ou dictionnaires) d'opinions. L'objectif de ces lexiques est de répertorier le plus de mots possible qui sont porteurs d'opinions. Ces mots permettent ensuite de classer les textes en deux (positif et négatif) ou trois catégories (positif, négatif et neutre).

A titre d'exemple, dans [33], les auteurs décrivent un système appelé *Opinion Observer*, qui permet de comparer des produits concurrents en utilisant les commentaires écrits par les internautes. Pour cela, ils ont une liste prédéfinie de termes désignant des caractéristiques de produits. Lorsque l'une de ces caractéristiques est présente dans un texte, le système extrait les adjectifs proches dans la phrase. Ces adjectifs sont ensuite comparés aux adjectifs présents dans leur lexique d'opinions et ainsi, une polarité est attribuée à la caractéristique du produit.

a) Construction des lexiques d'opinion

Pour construire des lexiques d'opinions (sentiments), trois méthodes peuvent être appliquées [95,96] :

- La méthode manuelle ;
- La méthode basée sur les corpus ;
- La méthode basée sur les dictionnaires.

La méthode manuelle, qui consiste à remplir le lexique de mots d'opinions sans l'aide d'outils particuliers, demande un effort important en terme du temps. La sélection des mots porteurs d'opinion et le choix de leur polarité se font donc uniquement par l'expertise humaine.

On peut supposer qu'une part non négligeable de subjectivité peut alors entrer en jeu et peut entraîner certaines erreurs de classification. Quelle que soit la méthode de construction des lexiques, une première étape de classification manuelle est nécessaire.

En effet, une première liste de mots et d'expressions doit être identifiée. Ces mots sont appelés les germes (*ang. seed words*). Cet ensemble de mots est ensuite utilisé afin de découvrir, répertorier et classer d'autres mots et expressions porteurs d'opinions.

L'une des solutions permettant d'agrémenter cet ensemble de mots est donc l'utilisation de corpus de textes. Dans [97], l'auteur propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions présents dans la liste de germes. C'est ce qu'on appelle, chercher les "cooccurrences" de mots. Un mot apparaissant plus souvent à côté de mots positifs sera donc classé dans la catégorie positif et inversement.

Dans [98], les auteurs ont proposé une méthode similaire, mis à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé. D'autres méthodes [99,100] utilisent également cette hypothèse dans le but de compléter les lexiques d'opinion : deux mots ou groupes de mots ayant un fort taux d'apparitions communes sont supposés posséder une forte proximité sémantique.

Une autre méthode basée sur le corpus permettant de compléter le lexique d'opinions, consiste à utiliser les conjonctions de coordination présentes entre un mot déjà classé et un mot non classé [101,102]. Par exemple, si la conjonction AND sépare un mot classé positif dans le lexique d'opinion et un mot non classé, alors le mot non classé sera considéré comme étant positif. A l'inverse, si la conjonction BUT sépare un mot classé positif et un mot non classé, alors le mot non classé sera considéré comme étant négatif. Les conjonctions utilisées sont les suivantes : AND, OR, BUT, EITHER-OR, et NEITHER-NOR.

La dernière solution la plus utilisée est une méthode basée sur les dictionnaires. Elle consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet [12,103] afin de déterminer l'orientation sémantique de nouveaux mots. Par exemple, Liu [31,32] utilisent ce dictionnaire afin de prédire l'orientation sémantique des adjectifs. Dans WordNet, les mots sont organisés sous forme d'arbres, comme indiqué dans la figure suivante :

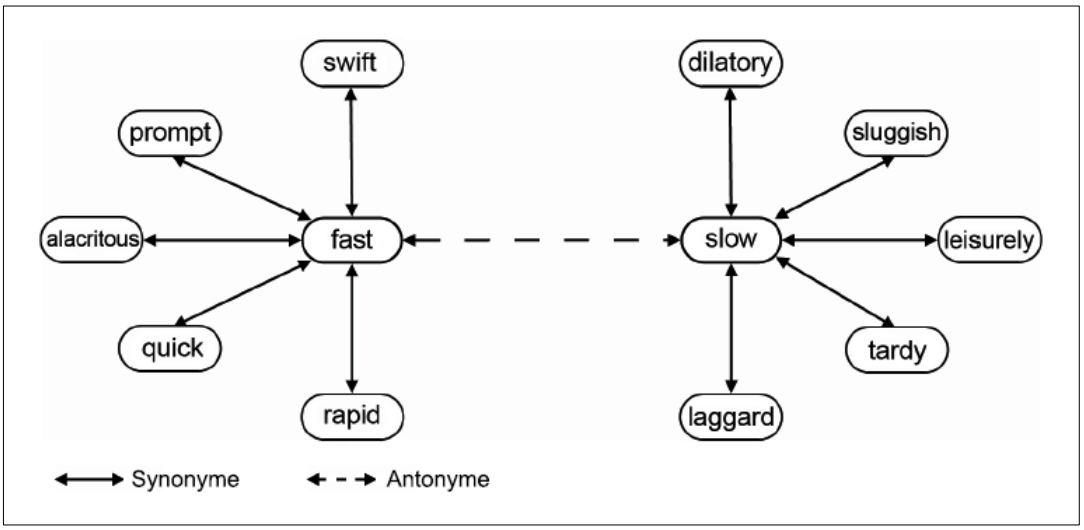


Figure 3.1 Exemple d'arbre de synonymes et d'antonymes présents dans WordNet

Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation sémantique connue. Cette méthode peut toutefois entrainer un certain nombre d'erreurs car un grand nombre de mots peuvent avoir différentes significations.

Des analyses plus poussées sont également faites. Afin de mesurer plus précisément la force de l'opinion exprimée dans une phrase, un moyen utilisé est l'extraction des adverbes associés aux adjectifs. Pour ce faire, Benamara et al. [120] proposent une classification des adverbes en cinq catégories : les adverbes d'affirmation, de doute, de forte intensité, de faible intensité et les adverbes de négation et minimiseurs. Un système d'attribution de points en fonction de la catégorie de l'adverbe permet de calculer la force exprimée par le couple adverbe-adjectif. Le tableau 3.1 présente quelques exemples de mots classés dans les catégories définies par Benamara et al. [120].

Classe	Mots
Adverbes d'affirmation	Absolutely, entirely, fully, certainly, fairly, exactly, totally, enough...
Adverbes de doute	Even, possibly, roughly, apparently, seemingly...
Adverbes de forte intensité	So, really, very, pretty, highly, extremely, much, well, too, quite...
Adverbes de faible intensité	Only, a little, almost, a bit, little, rather, nearly, barely, scarcely, weakly, slightly...
Négation et minimiseurs	Not, never, less, no...

Tableau 3.1 Exemple de catégories d'adverbes

Toutes ces catégories d'adverbes ne sont pas toujours prises en compte car elles n'ont pas la même importance au niveau de la prédiction de note. Les négations paraissent logiquement être des termes importants à détecter, en plus des adjectifs et des verbes, car ils permettent d'inverser la polarité d'une phrase. Pour traiter cet aspect, Das et Chen [121] proposent par exemple d'ajouter des mots dans le dictionnaire d'opinion comme "like-NOT" qui sont utilisés lors de la détection d'une couple like-négation. Les négations peuvent alors être *not, don't, didn't* ou *never* pour le cas de l'anglais et *ne pas* pour le français. Mais le problème de la détection de la négation reste un problème très ouvert, les méthodes existantes n'étant pas réellement convaincantes.

En effet, l'expression de la négation peut être faite sans l'utilisation de ce type de mots. Les expressions telles que « *je suis contre* » ou encore « *je m'oppose* » peuvent également permettre d'inverser la polarité du reste de la phrase. La difficulté est également due aux différentes façons d'utiliser la négation comme le sarcasme ou l'ironie. L'interprétation de la négation nécessite alors une analyse syntaxique qui est un traitement très coûteux en temps de calcul et pas forcément très efficace suivant la qualité du texte analysé.

b) Classification des textes grâce aux lexiques

Une fois que les mots porteurs d'opinion sont répertoriés dans les lexiques, la dernière étape consiste à déterminer la polarité d'une phrase à l'aide de ces mots. La solution la plus simple consiste à compter le nombre de mots positifs et le nombre de mots négatifs présents. S'il y a une majorité de termes positifs, la phrase est déclarée positive. A l'inverse, si les mots négatifs sont les plus nombreux, la phrase est déclarée négative. Les phrases possédant autant

de mots négatifs que de mots positifs peuvent être déclarées neutres [98], ou encore, la polarité de la phrase peut dépendre du dernier mot d'opinion parcouru [31]. On peut encore extraire plusieurs opinions dans une même phrase et les associer aux caractéristiques discutées [104].

3.2.2. Les approches basées sur l'apprentissage automatique

Dans ce type d'approches, les mots sont généralement considérés comme des variables équivalentes. L'aspect sémantique n'est alors pas pris en compte. Les méthodes les plus utilisées pour la classification d'opinions sont les méthodes de classification supervisée. Ce type de méthodes consiste à construire un modèle de classification à l'aide d'exemples. Des exemples sont des données dont on connait déjà la classe. On parle dans ce cas de données classées ou étiquetées. Voici une définition plus formelle du problème de la classification supervisée :

Soit X un ensemble de données, Y un ensemble d'étiquettes (ou classes) et D un ensemble des représentations des données. Soit $d : X \rightarrow D$*, une fonction qui associe à chaque donnée* $x \in X$ *une représentation* $d(x) \in D$ *et* $S \subset D \times Y$ *un ensemble de données étiquetées* $(d(x), y)$ *avec* $y \in Y$*. La classification supervisée consiste à construire un classifieur en s'appuyant sur l'ensemble S, qui permette de prédire la classe de toute nouvelle donnée* $x \in X$*, représentée par* $d(x) \in D$*.*

D'après cette définition, quatre éléments distincts entrent en jeu dans la classification supervisée :

– La représentation des données (l'ensemble D) ;

– Les étiquettes ou classes de prédiction (l'ensemble Y) ;

– Les données étiquetées, qui constituent le corpus d'apprentissage (l'ensemble S) ;

– Le classificateur ou le prédicteur.

a) Le corpus d'apprentissage

Tout d'abord, il est nécessaire de posséder des exemples (données étiquetées) afin de construire le "corpus d'apprentissage". Ce corpus ayant un impact direct sur l'apprentissage du modèle et par conséquent, sur les résultats de la classification, il est nécessaire que les exemples soient les plus représentatifs possibles de l'ensemble des données. En classification d'opinions, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'´étiquetage des exemples est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage. Cependant, les données présentes sur les sites Web 2.0, qui peuvent souvent être étiquetées par les auteurs eux-mêmes ou par d'autres internautes, permettent aujourd'hui de traiter des corpus beaucoup plus conséquents.

b) Les classes de prédiction

Concernant le choix des classes, il est généralement imposé par le corpus d'apprentissage utilisé et la tâche visée. On parle de classification binaire lorsque le nombre de classes |Y| est égal à 2. Il peut naturellement être supérieur, mais il ne faut pas oublier qu'un nombre de classe élevé augmente le taux d'erreurs. En classification d'opinions, le nombre de classes de prédiction choisi est généralement de deux (aime, n'aime pas) ou trois (aime, avis neutre, n'aime pas) [91,105].

c) La représentation

Les documents textuels sont des suites de caractères. Afin d'en permettre l'exploitation, il est nécessaire de représenter les textes numériquement. En fouille de textes, les documents sont généralement représentés sous leur forme vectorielle dite en "sac de mots". Les mots sont alors considérés comme des objets distincts non ordonnés. Cette représentation consiste tout d'abord à sélectionner des variables représentant les dimensions d'un vecteur puis à représenter chaque document sur ce vecteur. En classification de texte, la représentation implique donc une étape préliminaire appelée segmentation, qui consiste à découper le texte afin de sélectionner les variables. Une fois les variables acquises, plusieurs documents peuvent être représentés sur le même espace vectoriel. On obtient alors une matrice dite de Salton [106]. Nous présentons tout d'abord en quoi consiste exactement la segmentation, puis

nous énumérons les types de représentations les plus utilisées en recherche d'information qui sont également celles utilisées en classification d'opinions.

— **La segmentation**

Un document est un ensemble de caractères. La segmentation consiste à découper cet espace de caractères afin d'obtenir un espace de variables. Cet espace de variables est appelé vocabulaire et différents choix concernant sa construction sont possibles suivant le délimiteur choisi. On peut considérer les marques de ponctuation, afin de découper le document en phrases, ou encore les espaces associés à la ponctuation afin d'obtenir des mots. Ce sont d'ailleurs les mots qui sont les variables les plus utilisées en classification d'opinion, mais d'autres choix existent. On peut par exemple limiter la taille des variables à un nombre *n* de caractères, on parle alors de n-grammes de lettres. Les n-grammes peuvent également être formés de mots. Ces n-grammes de mots peuvent être construits selon leur ordre dans la phrase, afin de conserver un sens sémantique. Par exemple, dans la phrase « je n'aime pas ce film », le bigrammes (ou 2-grammes) « aime pas » sera considéré. On peut également construire les n-grammes de mots selon qu'ils apparaissent dans la même phrase par exemple. Une fois le vocabulaire sélectionné, différents choix sont possibles concernant la représentation du document sur le vecteur.

— **La représentation binaire**

Cette représentation est la moins coûteuse en temps de calcul. Elle consiste à indiquer, pour un document, quels mots du vocabulaire sont présents (valeur égale à 1) et quels mots sont absents (valeur égale à 0) [108, 107,109].

— **La représentation fréquentielle**

Cette représentation est une extension de la représentation binaire qui prend en compte le nombre d'occurrences des variables dans chaque document. Un texte est donc représenté par un vecteur dont chaque composante correspond à la fréquence des variables dans le texte [110].

— **La représentation fréquentielle normalisée**

Cette représentation consiste à normaliser les vecteurs de représentation des textes par la longueur des textes. C'est-à-dire que les fréquences des variables obtenues avec la représentation fréquentielle sont remplacées par la proportion des variables dans chaque document. La proportion s'obtient en divisant la fréquence de la variable par la taille du document.

— **La représentation TF-IDF**

La mesure statistique TF-IDF permet l'évaluation d'une variable dans un document à la fois par sa fréquence dans le document concerné, mais également par sa présence dans tous les autres documents du corpus.

La valeur TF (Term Frequency) correspond à la fréquence de la variable *v* dans le document *d* normalisée par la taille de *d*. La valeur IDF (Inverse Document Frequency) mesure l'importance de la variable *v* dans l'ensemble du corpus en calculant le logarithme de l'inverse du nombre de documents contenant *v* [111,112].

d) Le classificateur

Un classificateur est une procédure qui, à l'aide d'un ensemble d'exemples, produit une prédiction de la classe de toute donnée. Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinions.
On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, les règles de décision ainsi que des méthodes combinant différents classificateurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les Machine à Vecteurs de Support (SVM) [107, 108, 109, 110, 111, 113] et les classificateurs Naïfs Bayésiens (NB) [98, 109, 110].

— **Les Machines à Vecteurs de Support**

Les Machines à Vecteurs de Support, appelées encore Séparateurs à Vaste Marge, sont des classificateurs binaires très populaires en classification de textes. Considérons tout d'abord le cas où les données sont linéairement séparables. Nommons positif et négatif les deux classes

y ∈ Y. Si le problème est linéairement séparable, les individus positifs sont séparables des individus négatifs par un hyperplan H. Notons H+ l'hyperplan parallèle à H qui contient l'individu positif le plus proche de H, respectivement H− pour l'individu négatif. Une machine à vecteurs de support linéaire recherche alors l'hyperplan qui sépare les données de manière à ce que la distance entre H+ et H− soit la plus grande possible.

Cet écart entre les deux hyperplans H+ et H− est appelé la « marge ». Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple.

Dans le cas où les données ne sont pas linéairement séparables, l'idée des SVM est de changer l'espace de représentation. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace, généralement plus grand, appelé « espace de re-description ».

Dans le cas d'une classification multi-classes, plusieurs méthodes sont possibles, la plus connue étant le principe du Un-contre-tous. Il consiste à apprendre tout d'abord un modèle à l'aide de la première classe face à toutes les autres, puis de la deuxième face à toutes les autres et ainsi de suite. On obtient ainsi plusieurs classificateurs que l'on applique tous lors de la classification d'une nouvelle donnée. La classe attribuée à cette donnée est alors celle obtenant le meilleur score.

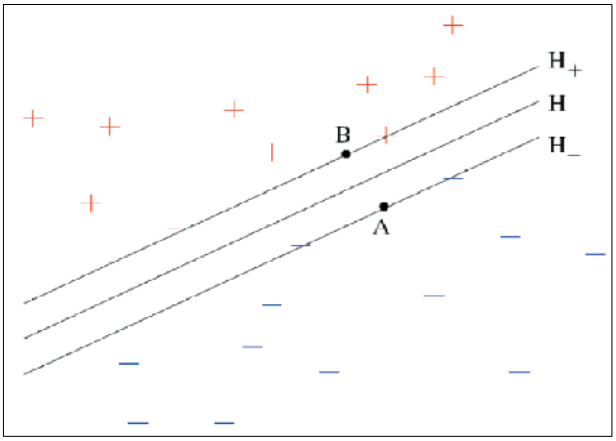


Figure 3.2 Exemple d'hyperplan (H) séparant les individus appartenant à la classe (+) et ceux appartenant à la classe (–)

— **Les classifeurs Naïfs Bayésiens**

Le principe d'un classificateur naïf bayésien consiste à maximiser la probabilité Pr(y|d), soit la probabilité d'occurrence de la classe de prédiction y connaissant la représentation de la nouvelle donnée x (on suppose donc ici d = d(x) = (d1, d2, ..., dn)), et ce pour toutes les classes y ∈ Y et toutes les composantes qui interviennent dans la définition de l'espace de représentation D. Pour cela, on fait appel la règle de Bayes.

Règle de Bayes. Soient A et B deux évènements. La règle de Bayes dit alors que la probabilité de l'évènement A sachant l'évènement B (Pr(A|B)) peut se calculer à l'aide des probabilités des évènements A et B (Pr(A) et Pr(B)) et connaissant la probabilité de l'´évènement B sachant l'évènement A (Pr(B|A)) par la formule suivante :

$$Pr(A|B) = Pr(B|A)\, Pr(A) / Pr(B)$$

Application à la classification. En appliquant la règle de Bayes à la problématique de la classification, on obtient l'équation suivante :

$$Pr(y|d) = Pr(d|y)\, Pr(y) / Pr(d)$$

Les probabilités de l'expression de droite doivent être estimées, à l'aide du corpus d'apprentissage S, afin de calculer la quantité qui nous intéresse, soit P(y|d) :

– Pr(y) est la probabilité d'observer la classe y ;

– Pr(d) est la probabilité d'observer la représentation d ;

– Pr(d|y), la vraisemblance de l'évènement « observer la représentation d » si s ∈ S est de classe y. Ce terme est plus difficile à estimer que le précédent.

En pratique, on ne s'intéresse qu'au numérateur, le numérateur ne dépendant pas de y. Concernant Pr(d|y), l'hypothèse habituellement faite dans ce type de classifieurs est que toutes les composantes d_i sont indépendantes, ce qui permet de calculer facilement la probabilité globale d'une classe connaissant une donnée. Cette non-dépendance des composantes correspond à « l'hypothèse de Bayes naïve ». On considère donc que Pr(d|y) =

$\prod_i Pr(d_i|y)$. Maximiser Pr(y|d) revient donc à maximiser $(\prod_i Pr(d_i|y))Pr(y)$.

Les Pr(di|y) sont évalués par les fréquences observées dans les exemples de l'ensemble S.

3.2.3. Les approches hybrides

Plusieurs types d'hybridation sont présents dans la littérature. Dans tous les cas, elles utilisent des éléments décrits dans les deux approches précédentes. On peut distinguer parmi ces approches trois méthodes distinctes :

– La linguistique au service de l'apprentissage automatique ;

– L'apprentissage automatique au service de la linguistique ;

– Une fusion a posteriori des résultats des deux approches.

a) La linguistique au service de l'apprentissage automatique

Une approche hybride consiste à utiliser les outils linguistiques afin de préparer le corpus avant de classer les textes à l'aide de l'apprentissage supervisé. Wilson et al. [113] préparent les données à l'aide d'outils de TALN afin de sélectionner un vocabulaire d'opinions. Ces mots présélectionnés sont ensuite utilisés comme vecteurs de représentation des textes pour les outils d'apprentissage supervisé. Trois algorithmes d'apprentissage sont comparés : BoostTexter [114], Ripper [115] et SVM^{light} [116]. Nigam et Hurst [117] utilisent des techniques provenant du Traitement Automatique des Langues afin de détecter dans les textes les mots et expressions porteurs d'opinions et ajoutent des marques dans le texte (traits grammaticaux et + ou - pour opinion positive et opinion négative). Ils utilisent ensuite l'apprentissage automatique pour classer les textes selon leur opinion générale.

b) L'apprentissage automatique au service de la linguistique

Une autre façon de combiner les méthodes est d'utiliser les techniques d'apprentissage automatique dans le but de construire les dictionnaires d'opinions nécessaires à l'approche linguistique.

Hatzivassiloglou et McKeown [101] présentent une méthode ayant pour objectif de définir l'orientation sémantique des adjectifs pour la construction du dictionnaire d'opinions. Ils extraient tout d'abord tous les adjectifs du corpus à l'aide d'un analyseur syntaxique, puis utilisent un algorithme de clustering afin de classer les adjectifs selon leur polarité.

Riloff et Wiebe [118] combinent les deux approches afin de répertorier les expressions porteuses d'opinion qui, selon eux, sont plus riches que des mots pris individuellement.

Turney et Littman [94] utilisent une approche statistique pour classer un plus grand nombre de types de mots selon leur polarité : adjectifs, verbes, noms...

c) Une fusion a posteriori des résultats des deux approches

Une dernière façon d'utiliser conjointement les approches basées sur la linguistique et celles basées sur l'apprentissage automatique est de construire plusieurs types de classificateurs afin de combiner leurs résultats, soit par des systèmes de vote, soit par un algorithme d'apprentissage [119]. Dans le cas d'un système de vote, on attribue généralement des poids suivant les performances des classificateurs utilisés. Si les résultats diffèrent, on conserve alors le résultat du classificateur ayant le poids le plus fort. Dans le cas de l'utilisation d'un algorithme d'apprentissage, les poids sont calculés automatiquement selon l'indice de confiance attribué par chaque classificateur à ses propres résultats.

3.2.4. Les différentes évaluations utilisées

La classification d'opinions est une tâche de classification dite supervisée car les classes sont déterminées à l'avance.

		Classes réelles	
		Classe 1	Classe 2
Classes prédites	Classe 1	A	C
	Classe 2	B	D

Tableau 3.2 Exemple d'une matrice de confusion pour une classification binaire

Comme toute tâche de classification supervisé, les résultats obtenus peuvent alors être représentés sous la forme d'une matrice de confusion (Tableau 3.2).

Les réponses du classificateur sont comptabilisées dans chaque ligne de la matrice et les réponses attendues sont comptabilisées dans les colonnes. À partir de cette matrice, plusieurs évaluations peuvent être calculées.

a) Le taux d'erreurs

Le taux d'erreurs représente le pourcentage de bonnes prédictions en rapport au nombre total de prédictions. La formule correspondant au calcul de ce taux est la suivante :

TE = Objets bien classés / Nombre total d'objets

En appliquant la formule à l'exemple précédent, on obtient donc :

TE = (A + D) / (A + B + C + D)

b) F-score

Le F-score est la mesure la plus utilisée en classification d'opinions. Son évaluation permet de tenir compte à la fois de la précision ainsi que du rappel. Il se mesure à l'aide de la formule suivante :

F-score = 2 × (Précision × Rappel) / (Précision + Rappel)

c) Précision

$$Précision_i = \frac{Objets\ correctements\ attribués\ \text{à}\ la\ classe\ i}{Nombre\ d'objets\ attribués\ \text{à}\ la\ classe\ i}$$

Afin de calculer la précision totale, on effectue la moyenne des précisions de chaque classe :

$$Précision = \frac{\sum_{i=0}^{n} Précision_i}{n}$$

En appliquant cette formule à notre exemple, on obtient :

$$Précision = \frac{\left(\frac{A}{A+B}\right) + \left(\frac{D}{B+D}\right)}{2}$$

d) Rappel

Le rappel est défini par le nombre de documents bien classés en rapport au nombre de documents pertinents contenus dans le corpus. Il s'oppose au silence, le silence représentant les documents pertinents non trouvés. Il se calcule de la façon suivante :

$$Rappel_i = \frac{Documents\ correctements\ attribués\ \text{à}\ la\ classe\ i}{Nombre\ de\ documents\ attribués\ \text{à}\ la\ classe\ i}$$

Afin de calculer la précision totale, on effectue la moyenne des précisions de chaque classe :

$$Rappel = \frac{\sum_{i=0}^{n} Rappel_i}{n}$$

En appliquant cette formule à notre exemple, on obtient :

$$Rappel = \frac{\left(\frac{A}{A+B}\right) + \left(\frac{D}{B+D}\right)}{2}$$

3.3. Travaux connexes

Contrairement à la langue anglaise, la langue arabe n'a pas attiré beaucoup d'attention des chercheurs dans le domaine d'opinion-mining. Plusieurs raisons peuvent être à l'origine de l'absence de travaux de recherche : l'arabe est syntaxiquement et morphologiquement complexe qui provoque une ambiguïté lexicale et sémantique élevée, surtout quand les voyelles ne sont pas employés, un manque remarquable de ressources linguistiques pour l'arabe comme les lexiques de sentiments et les ressources subjectives sur le Web pour la création des corpus. Dans cette section, nous présentons quelques travaux connexes focalisent

sur l'utilisation des ontologies ou taxonomies comme modèles de représentation des connaissances, et ceux qui s'intéressent à l'extraction d'opinions dans la langue arabe.

3.3.1. Utilisation des règles linguistiques

Parmi les travaux qui dépendent de l'extraction simples des caractéristiques, nous citons le travail de Hu et Lui [104], les auteurs ont utilisé un ensemble de règles linguistiques pour extraire les caractéristiques des produits à partir des commentaires d'utilisateurs (noms et groupes-nominaux). Seules les expressions d'opinion (les adjectifs seulement) qui sont proches de caractéristiques dans les textes sont extraites. Enfin, un résumé pour chaque caractéristique a été produit pour afficher pour chaque caractéristique l'ensemble des expressions d'opinion positives et négatives et le nombre total pour les deux catégories.

Afin d'améliorer la phase d'extraction de caractéristiques, Popescu et Etzioni [122] suggèrent dans leur système appelé OMINE, d'extraire uniquement les groupes nominaux dont la fréquence est supérieure à un seuil déterminé expérimentalement en utilisant le calcul de la PMI (Point-wise Mutual Information) entre chacun de ces noms et les expressions de métonymie associées au produit.

La principale limitation de ces approches, est qu'il y a un grand nombre de caractéristiques extraites mais il y a un manque d'organisation. Ainsi, les concepts sémantiquement similaires ne sont pas regroupés (par exemple, « accueil » et « réception »), et les relations possibles entre les caractéristiques d'un objet ne sont pas reconnus (par exemple, « café » est une « boisson »). En outre, l'analyse de polarité (positive, négative ou neutre) du document se fait en attribuant la polarité dominante des mots d'opinion qu'il contient (en général des adjectifs), indépendamment des polarités associées individuellement à chaque caractéristique.

3.3.2. Utilisation des modèles de représentation des connaissances

Pour la représentation des connaissances du domaine étudié, nous distinguons deux sous-familles d'approches, celles qui utilisent les taxonomies et celles qui utilisent les ontologies :

— **Utilisation des taxonomies**

Les approches qui utilisent les taxonomies pour la représentation des connaissances, ne cherchent pas une liste simple de concepts de domaine, mais plutôt une liste hiérarchisée de concepts.

Nous rappelons que la taxonomie est une liste de termes organisés hiérarchiquement par une relation de spécialisation de type « est une sorte de ». Dans [123], les auteurs utilisent des taxonomies prédéfinis et des mesures de similarité sémantique pour extraire automatiquement les caractéristiques d'un produit et de calculer la distance entre les concepts prédéfinis dans la taxonomie.

Dans leur système appelé PULSE [124], les auteurs analysent une grande quantité de textes contenus dans une base de données. Une taxonomie contenant les marques et les modèles de voitures, est automatiquement extraite de la base de données. Couplé avec une technique de classification, les phrases correspondantes à chaque feuille de la taxonomie sont extraites. A la fin du processus, un résumé qui peut être plus ou moins détaillé est produit.

Le système décrit dans [125], extrait des informations sur les services offerts aux clients sur le net, et utilise une fonction d'agrégation pour calculer l'orientation sémantique des sentiments exprimés sur tous les concepts, puis il produit un résumé. L'extraction automatique combine une méthode dynamique, où les différents aspects des services sont les noms les plus communs, et une méthode statique, où une taxonomie regroupant les concepts considérés comme les plus pertinents par l'utilisateur est utilisée pour annoter manuellement les phrases. Les résultats ont également montré que l'utilisation d'une hiérarchie améliore considérablement la qualité du processus d'extraction des caractéristiques.

— **Utilisation des ontologies**

Ces travaux ont pour but d'organiser les caractéristiques en utilisant un modèle plus élaboré de représentation : l'ontologie, contrairement à la taxonomie, ne se limite pas aux relations hiérarchiques entre les concepts, mais peut décrire d'autres types de relations paradigmatiques comme la synonymie ou des relations plus complexes telles que la relation de composition et les relations spatiales.

Généralement, les caractéristiques extraites à partir des textes correspondent exclusivement à des termes contenus dans l'ontologie. La phase d'extraction de caractéristiques est guidée par une ontologie de domaine, construite manuellement [126], ou semi automatiquement [127, 128], qui est ensuite enrichi par un processus automatique d'extraction des termes qui correspondent à l'identification de nouvelles caractéristiques.

Pour extraire les termes, l'auteur dans [127] utilise un modèle d'extraction couplé avec un extracteur terminologique entraîné sur un ensemble de caractéristiques liées à un ensemble de produits et dans certaines critiques, l'identification se fait manuellement. Les caractéristiques similaires sont regroupées à l'aide des mesures de similarité sémantique.

Le système OMINE décrit dans [128], propose un mécanisme pour l'enrichissement de l'ontologie en utilisant un glossaire de domaine qui comprend des termes spécifiques, tels que des mots de jargon, d'abréviations et d'acronymes.

Dans [126], les auteurs ont proposé d'enrichir l'ontologie utilisée, par l'ajout d'un ensemble de concepts en utilisant une méthode basée sur un corpus : les phrases contenant des concepts conjointement liés, sont reconnus et extraits. Ce processus est répété de manière itérative jusqu'à ce qu'aucun nouvel concept ne soit trouvé.

Les ontologies ont également été utilisées pour soutenir la fouille de polarité. Par exemple, dans [129], où une approche a construit manuellement une ontologie pour les critiques de films et l'ont incorporé dans la tâche de classification de polarité, qui a amélioré considérablement les performances de leur approche.

Les ontologies ont joué un rôle important dans la tâche d'extraction de caractéristiques, construite manuellement [126], ou semi-automatiquement [127,128], dotée d'un processus d'extraction automatique des termes, correspondant à une nouvelle identification des caractéristiques. Les caractéristiques similaires sont regroupées en utilisant des mesures de similarité sémantique.

Dans [129], les auteurs ont construit manuellement une ontologie pour les critiques de films, intégrée ensuite dans la tâche de classification de polarité, ce qui a amélioré sensiblement les performances de leur approche.

Cadilhac et al, (2010) [134], ont affirmé que l'utilisation d'une hiérarchie de caractéristiques améliore les performances des fonctionnalités des systèmes d'identification d'opinions au nivea
u des caractéristiques.

3.4. La fouille d'opinion en langue arabe

Dans cette section, nous présentons quelques travaux de recherche en fouille d'opinion qui se focalise sur la langue arabe et qui peuvent donner plus de motivation à notre travail de recherche :

Dans [140], les auteurs ont proposé une approche pour la classifications des commentaires Facebook, écrits dans la langue de l'argot arabe, basée sur l'utilisation d'un classificateur SVM, leur approche utilise un lexique construit manuellement regroupant les mots de sentiment de l'argot arabe, le lexique a été utilisé pour extraire les mots d'opinion dans les commentaires, ce qui a amélioré le résultat de classification par rapport à l'utilisation d'un lexique de sentiments classique.

L'approche proposée dans [141], centrée autour de la création d'un lexique de sentiments pour la langue arabe en traduisant les lexiques disponibles pour l'anglais vers l'arabe.

Le travail de recherche présenté dans [142], a porté sur l'extraction du détenteur de l'opinion (*ang. Opinion holder*). Les chercheurs ont combiné des techniques d'apprentissage automatique et des règles linguistiques pour identifier les détenteurs d'opinion. Le corpus utilisé a été créé manuellement à partir des sites Web de la presse arabe.

L'auteur a présenté dans [143] une approche pour extraire les opinions dans les documents arabe, en combinant plusieurs techniques. Les résultats ont montré que les performances de classification s'améliorent lorsque de nombreux classificateurs collaborent.

3.5. Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur la classification d'opinions, qui nous intéresse particulièrement dans notre travail de recherche.

Nous avons dressé l'état de l'art des méthodes existantes : les méthodes basées sur la construction de lexiques d'opinions, les méthodes basées sur l'apprentissage automatique et les méthodes hybrides qui mêlent les deux premières citées.

Les méthodes basées sur les lexiques offrent une analyse fine mais ont besoin toujours d'une intervention humaine que les méthodes statistiques et nécessitent des connaissances sur le langage. Les méthodes hybrides permettent de trouver un compromis : minimisation de l'intervention humaine tout en permettant d'analyser les opinions de manière précise.

Les travaux connexes présentés dans ce chapitre, qui utilisent des modèles de représentation des connaissances, ont montré que les performances des systèmes de classification s'améliorent considérablement lors de la combinaison d'une ontologie ou une taxonomie.

Dans notre étude, nous croyons que l'utilisation d'une ontologie pour conceptualiser les connaissances du domaine étudié peut faciliter l'extraction des caractéristiques explicites et implicites sur lesquelles des opinions ont été exprimées, d'où une amélioration de la performance du système de classification. Le chapitre suivant sera dédié à étudier les techniques et les outils utilisés en ingénierie ontologique.

Chapitre 4

Ingénierie ontologique

4.1. Introduction

L'ingénierie ontologique est reconnue aujourd'hui dans différents domaines de recherche : Ingénierie des connaissances, Intelligence artificielle, gestion des connaissances, linguistique, systèmes d'information, recherche et extraction d'informations, etc. Cette discipline se situe au carrefour de plusieurs disciplines auxquelles elle emprunte des concepts, notamment : l'ingénierie des connaissances, l'ontologie et la linguistique.

L'ingénierie ontologique est une méthodologie qui nous donne la logique du design d'une base de connaissances, le cœur d'une conceptualisation du monde-cible, des contraintes sémantiques de ces concepts ainsi que des théories et des technologies permettant l'accumulation de connaissances qui est indispensable pour le traitement des connaissances dans le monde réel [51].

Une ontologie peut être vue comme une conceptualisation du monde réel ainsi qu'une base solide sur laquelle une base de connaissances soit partageable, dont le partage et la réutilisation sont les concepts-clés d'une ontologie.

En premier axe de ce chapitre, nous présentons un état de l'art sur l'ingénierie des connaissances dont l'ingénierie ontologique est un sous-champ. En deuxième axe, nous présentons la discipline naissante - Ingénierie ontologique - qui est concernée par la construction d'ontologies, en donnant un bilan sur les méthodes de construction des ontologies, les outils, et les langages de spécification et environnements de développement disponibles pour la construction des ontologies.

4.2. L'ingénierie des connaissances

L'ingénierie des connaissances abrégée en IC ou KM (Knowledge Management en anglais) est définie dans [57] comme suit :

« l'étude des concepts, méthodes et techniques permettant de modéliser et/ou acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches ne se formalisant a priori pas ou peu ».

Dans [50], l'IC a pris la place du domaine de l'Acquisition des Connaissances (AC) à partir des années 80. L'évolution des procédés liés à l'acquisition de connaissances peut s'analyser à travers l'évolution de ces deux domaines de recherche.

Dans cette section, nous présentons le domaine de l'IC, en expliquant les notions et les techniques d'acquisition, modélisation et présentation des connaissances qui s'articulent autour de notre sujet d'étude.

4.2.1. La notion de connaissance

Parmi les définitions issues de la littérature dans le domaine de l'IC, nous citons quelques-unes :

« Une connaissance est la capacité d'exercer une action pour atteindre un but. » [56].

« La connaissance est l'information organisée qui est applicable à la résolution de problèmes. » [58].

« La connaissance inclut des restrictions implicites et explicites entre objets ainsi que des opérations et des relations, qui permettent de définir des heuristiques générales et spécifiques comme les procédés d'inférences liées à la situation à modéliser. » [59].

La connaissance est définie dans un cadre bien précis et prend sa signification dans le contexte de son utilisation. On ne peut pas parler de connaissance a priori. On ne peut parler de connaissance qu'à partir du moment où l'information manipulée par le système prend un sens pour l'utilisateur, c'est-à-dire qu'il peut établir un lien avec cette information et celle qu'il possède déjà [57].

L'information, constituée des données, devient connaissance à partir du moment où elle sert de fondement à une inférence[1], au déclenchement d'un processus [60].

4.2.2. Représentation de la connaissance

Le processus d'ingénierie des connaissances définit des étapes pour organiser la connaissance au sein de représentations formelles [9]. Un modèle conceptuel de la connaissance est ensuite traduit en une représentation qui pourra être manipulée par les systèmes informatiques [49].

Représenter la connaissance a pour objectif de modéliser la connaissance en omettant certains détails non significatifs pour en permettre une meilleure manipulation [61]. La représentation de la connaissance s'appuie alors sur des représentations au niveau conceptuel pouvant modéliser la « structure cognitive » d'un domaine [62].

Les ontologies sont des exemples de telles représentations. Elles permettent de représenter pour un domaine de connaissance donné, les concepts, les relations entre les concepts, ainsi que la sémantique de ces relations. Dans ce qui suit, nous présentons les ontologies comme outils de représentation formelles des connaissances, leurs types, et leurs avantages d'utilisation.

4.2.3. Modèles de représentation de la connaissance

Une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine [63,64].

Les ontologies sont utilisées dans de nombreux domaines. Les domaines recensés dans [65] sont :

— L'ingénierie des connaissances ;

— La modélisation qualitative ;

— L'ingénierie des langages ;

[1] Une inférence est définie dans [61] comme *« une façon générique de désigner l'ensemble des mécanismes par lesquels des entrées (perceptives ou non) sont combinées à des connaissances préalables afin d'obtenir des comportements élaborés »*.

— La conception de bases de données ;

— La recherche d'information ;

— L'extraction d'information ;

— La gestion et l'organisation de connaissances.

Depuis, grâce à l'essor du Web, elles sont utilisées dans le domaine de l'e-commerce et sont au cœur du Web Sémantique [66], future version du Web actuel. Un des plus grands projets reposant sur l'utilisation des ontologies consiste à ajouter au Web une véritable couche de connaissance permettant des recherches d'information au niveau sémantique et non plus au simple niveau lexical et/ou syntaxique. A terme, il est prévu que des applications déployées sur l'Internet pourront mener des raisonnements utilisant les connaissances stockées sur la toile [65].

Derrière l'utilisation d'ontologies dans ces différents domaines, se cachent en fin de compte plusieurs représentations de connaissances. Ces représentations peuvent être distinguées suivant deux axes : la nature de la connaissance représentée dans l'ontologie et le degré d'engagement sémantique qui a motivé la formalisation de l'ontologie. Le premier axe fait en particulier référence au type de connaissances représentées (génériques, de domaines ou liées à la tâche). Le second axe fait en particulier référence au niveau sémantique des connaissances que l'ontologie représente (ressource terminologique versus ressource conceptuelle). Nous présentons ces deux aspects : nature des connaissances et engagement sémantique dans les sections suivantes [50].

4.2.4. Connaissances du domaine

Dans [49], les connaissances du domaine d'un système à base de connaissances (SBC) sont définies comme les connaissances relatives au domaine de l'application et nécessaires pour que les méthodes de raisonnement puissent s'exécuter. Les travaux qui ont porté sur ces connaissances depuis une décennie ont eu un double impact : d'une part, ils ont montré l'intérêt de distinguer les connaissances du domaine selon leur nature et de raisonner sur des modèles de connaissances multiples en exploitant les spécificités de chacun d'eux ; d'autre part, ils ont montré l'intérêt de disposer de modèles de connaissances structurés, exprimés à

l'aide de langages ayant une sémantique bien définie et exprimés à différents niveaux de granularité.

D'après [50], la connaissance associée à un domaine peut être représentée de façon plus formelle au travers d'une ontologie. Pour un utilisateur, accéder par une ontologie à la connaissance à partir de laquelle l'information d'un corpus a été indexée peut lui permettre de spécifier son besoin et les lacunes de sa connaissance par rapport à l'information qui lui est disponible. D'autre part, la représentation des granules d'information à partir d'une ontologie peut définir un vocabulaire contrôlé (termes et concepts) à partir duquel l'utilisateur spécifiera son besoin. La description du besoin correspond, dans ce cas-là, aux caractéristiques des granules car elles ont été indexées à partir des mêmes ressources.

4.3. Les ontologies

Nous présentons dans cette section, la notion d'ontologies, leurs types et méthodes de construction, ainsi leurs caractéristiques et constituants.

4.3.1. Présentation

La rapidité de l'évolution de la masse d'informations dans tous les domaines a généré un besoin d'organisation et de structuration des contenus. Il est extrêmement difficile pour les humains d'interpréter l'information peu ou pas structurée complètement disponible sur Internet.

Les ontologies permettent aux humains et aux machines de partager les connaissances du domaine et de coopérer ensemble [84]. On les utilise en général pour permettre aux machines de raisonner et d'interpréter des informations, ainsi que d'améliorer la pertinence des recherches. Actuellement, les ontologies constituent un enjeu stratégique dans la représentation et la modélisation des connaissances.

En philosophie, l'ontologie a été définie comme l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe.

Le terme ontologie est repris en informatique et en science de l'information, où une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations. Elle est employée pour raisonner à propos des objets du domaine concerné, dont l'objectif

premier est de modéliser un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire.

Les ontologies sont employées dans l'intelligence artificielle, le Web sémantique, le génie logiciel, l'informatique ou encore l'architecture comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :

— *Individus* : les objets de base ;

— *Classes* : ensembles, collections, ou types d'objets ;

— *Attributs* : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager ;

— *Relations* : les liens que les objets peuvent avoir entre eux ;

— *Evénements* : changements subis par des attributs ou des relations.

4.3.2. Définitions

Une des premières définitions de l'ontologie communément admise en Intelligence Artificielle a été énoncée par Gruber [1] comme la *« spécification explicite d'une conceptualisation »*. Cette définition de l'ontologie a ensuite été affinée dans [2] comme *« spécification formelle et explicite d'une conceptualisation partagée »* :

— *Formelle* → l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel ;
— *Explicite* → la définition explicite des concepts utilisés et des contraintes de leur utilisation ;
— *Conceptualisation* → le modèle abstrait d'un phénomène du monde réel par identification des concepts-clés de ce phénomène ;
— *Partagée* → l'ontologie n'est pas la propriété d'un individu, mais elle représente un consensus accepté par une communauté d'utilisateurs.

En clair, une ontologie fournit les moyens d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés sémantiques dans un langage de

représentation des connaissances formel favorisant le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage [3].

Dans [5], l'auteur, définit plus formellement une ontologie comme étant :

— Un ensemble de concepts ;

— Un ensemble de relations entre ces concepts ;

— Un ensemble d'axiomes (transitivité, réflexivité, symétrie des relations…)

4.3.3. Caractéristiques d'une ontologie

Les ontologies possèdent les caractéristiques fondamentales suivantes [68] :

— *Les ontologies sont formelles*

Ceci signifie qu'elles sont exprimées dans une langue qui a une syntaxe clairement définie, et base mathématique pour leur signification. Comme, les concepts sont exprimés formellement, ils peuvent être traités par des programmes informatiques.

Les « concepts » ou les « objets » qui existent dans des techniques de modélisation traditionnelles (schéma relationnel et UML, par exemple) sont seulement semi-formels. Elles ne peuvent donc pas être manipulées automatiquement par des logiciels sans un effort considérable (et coûteux) de programmation de manière à souligner leurs significations [8].

— *Les ontologies sont lisibles par les humains*

Ceci signifie qu'elles peuvent être développées, partagées, et comprises non seulement par des programmes informatiques, mais aussi par les communautés d'experts du domaine ainsi que des utilisateurs potentiels.

— *Les ontologies sont vastes*

Elles sont conçues dans le but d'inclure toute la signification appropriée des concepts liés à un domaine et simplement celle requise pour une application particulière. Cela veut dire que, si toute la signification des concepts est capturée par une ontologie, elle peut être comprise, modifiée, et contrôlée par n'importe quel expert de domaine.

— *Les ontologies sont partageables*

Elles sont construites sur la base de bibliothèques communes de concepts fondamentaux et sont utilisables à travers de multiples domaines d'application. Ceci facilite la combinaison des ontologies développées séparément pour permettre la communication entre les systèmes d'information, qui doivent partager des informations basées sur des concepts communs.

4.3.4. Constituants d'une ontologie

Les ontologies sont, à l'heure actuelle, au cœur des travaux menés en ingénierie des connaissances. Les ontologies permettent de représenter les connaissances et les manipuler automatiquement, tout en gardant leur sémantique. Une ontologie ne peut être construite que dans le cadre d'un domaine précis de la connaissance [69,70].

Les connaissances sont définies à travers des concepts. Les liens entre concepts sont appelés relations. Afin de relier les concepts, l'ontologie se présente sous forme d'une organisation hiérarchique des concepts. Nous détaillons ci-après ces deux éléments :

a) Les concepts

Un concept peut représenter un objet matériel, une notion ou une idée. C'est une représentation de l'esprit qui abrège et résume une multiplicité d'objets empiriques, ou mentaux par abstraction et généralisation des traits communs identifiables. Un concept peut être divisé en trois parties : un terme (ou plusieurs), une notion et un ensemble d'objets. Le terme est un élément lexical qui permet d'exprimer le concept en langue naturelle, il peut admettre des synonymes. La notion également appelée intension du concept, contient la sémantique du concept, exprimée en termes de propriétés et d'attributs, et de contraintes [67].

L'ensemble d'objets, appelé extension du concept, regroupe les objets manipulés à travers le concept ; ces objets sont appelés instances du concept. Les concepts sont organisés en taxonomie au sein d'un réseau de concepts, et peuvent être structurés hiérarchiquement.

Le concept d'ontologie[2] est une notion qu'il n'est pas toujours facile de caractériser. En effet, il est utilisé dans différents contextes, i.e. La philosophie, la linguistique, l'intelligence artificielle (IA)…, et chacun en donne une définition particulière.

Les *concepts* (aussi appelés « classes ») représentent les objets, abstraits ou concrets, réels ou fictifs, élémentaires ou composites, du monde réel [50].

Dans [72], l'auteur a décrit le concept et ses constituants comme suit :

Un concept est composé de trois parties : un ou plusieurs *termes*, une *notion* et un *ensemble d'objets*. La notion correspond à la sémantique du concept, elle est définie à travers ses propriétés et ses attributs.

La notion est appelée *intention* du concept. L'ensemble d'objets correspond aux objets définis par le concept, est appelé *extension* du concept ; les objets sont les *instances* du concept.

Le ou les *termes* permettent de désigner le concept. Ces termes sont aussi appelés *labels* de concept. Par exemple, le terme « lapin » renvoie à un animal possédant de longues oreilles et une queue et à l'ensemble des objets ayant cette description. Afin que les concepts soient reconnus de façon non ambiguë par la machine, il est souhaitable qu'un concept soit identifié à partir de plusieurs termes, ce qui permet de gérer la synonymie et de les désambiguïser les uns par rapport aux autres.

b) Les relations

Les relations représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine. Elles établissent des liens sémantiques binaires, organisables hiérarchiquement [7]. Par exemple, dans le domaine de cinématographie, les concepts « Personnalité » et « Film » peuvent être reliés entre eux par la relation sémantique « réalise (Personnalité, Film) » dans laquelle « Personnalité » est le domaine et « Film » la portée (ou « range » en anglais).

[2] Charles Guillemot dans sa page perso : Les ontologies : kesaco ?
http://charles-guillemot.info/2010/03/10/ontologie-kesaco/

Les instances des concepts peuvent être reliées entre eux par des relations au sein d'une ontologie. Une relation est définie comme une notion de lien entre des entités, exprimée souvent par un terme ou par un symbole littéral ou autres. Les relations sont caractérisées par un terme ou plusieurs termes, et une signature qui précise le nombre d'instances de concepts que la relation lie, leurs types et l'ordre des concepts, c'est-à-dire la façon dont la relation doit être lue.

Les relations traduisent les associations pertinentes existant entre les concepts présents dans le segment analysé de la réalité. Ces relations incluent les associations suivantes : Sous-classe de : généralisation ; spécialisation ; Partie de : agrégation ou composition ; Associée à ; Instance de,…etc. Ces relations nous permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres.

Une relation sémantique R représente un type d'interaction entre les concepts d'un domaine c_1, c_2, .., c_n. Elle se définit formellement à partir d'un produit de n concepts : R : $c_1 \times c_2 \times \ldots \times c_n$; « subsume », « est un phénomène lié à » sont des exemples de relations binaires [50].

Les relations les plus courantes dans la littérature sont les relations d'équivalence, taxonomiques, patronymiques, de dépendance, topologique, causale, fonctionnelle, chronologique [71].

Les *relations* représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine. Elles établissent des liens sémantiques binaires, organisables hiérarchiquement [7].

Les relations *taxonomiques* ou de *subsomption*, à priori, vont permettre de construire des hiérarchies strictes entre les concepts. Pour cela, dans les différents travaux sur les ontologies, seule la relation *is a* (*est un*) est généralement citée de taxonomique ou de subsomption.

Cependant, dans certains cas particuliers d'ontologies, d'autres relations peuvent être taxonomiques, c'est-à-dire le cas de la relation de composition *part of* (*partie de*) quand elle définit une hiérarchie stricte [85].

c) Les fonctions

Les fonctions constituent des cas particuliers des relations, dans laquelle un élément de la relation, le nième (extrant) est défini en fonction des n-1 éléments précédents (intrants). Nous

notons néanmoins que ce constituant d'ontologie est rarement évoqué dans la description d'ontologies. Nous pensons que cela est dû au fait que les relations « fonctionnelles » sont plutôt présentes dans certains domaines scientifiques comme la physique, la chimie,… [85].

d) Les axiomes (ou règles d'inférences)

Elles permettent de définir certaines propriétés de relations sous forme d'assertions, acceptées comme vraies, à propos des abstractions du domaine traduites par l'ontologie. Les axiomes spécifient la façon d'utilisation des primitives terminologiques du domaine (i.e. les concepts et les relations). Ces axiomes sont spécifiques aux ontologies et les distinguent des thésaurus, que ne représentent que des terminologies alors que les ontologies des connaissances sont au sens large.

Certains axiomes se retrouvent dans de nombreuses ontologies et/ou sont communs à de nombreuses primitives. On appelle ici ces axiomes particuliers des schémas d'axiome. Ces schémas d'axiomes peuvent être [85] :

- Les propriétés algébriques d'une relation (symétrie, réflexivité, transitivité) ;
- La propriété de subsomption entre concepts ou entre relations (relation *is a*) ;
- La cardinalité d'une relation ;
- …etc.

Certains schémas d'axiomes sont intégrés dans les formalismes de représentation des connaissances pour décrire des ontologies. Par exemple la relation *is a* (*est un*) apparait bien dans les formalismes de types Entité-Relation et Graphes Conceptuels.

e) Les instances

Elles constituent, la définition extensionnelle de l'ontologie ; ces objets véhiculent les connaissances (statiques, factuelles) à propos du domaine du problème [85].

Les instances de concepts (aussi nommés individus) ne font pas à proprement parler partie de l'ontologie, mais plutôt de la base de connaissances [86]. En effet, ces dernières permettent de stocker les instances des concepts, mais aussi les instances de relations et les valeurs des propriétés en fonction des contraintes imposées par l'ontologie. Dans le monde de l'ingénierie des connaissances, par référence à la logique de description, on parle aussi de la

Terminological-Box (ou T-Box) pour l'ontologie et de la Assertion-Box (ou A-Box) pour la base de connaissance [87].

f) Les attributs

Les attributs correspondent à des caractéristiques, des spécificités particulières, attachées à un concept et qui permettent de le définir de manière unique dans le domaine [7]. Leurs valeurs sont littérales, i.e. de type primitif, comme une chaîne de caractères ou un nombre entier. Par exemple, un concept « Personne » peut avoir les attributs suivants : un « numéro de sécurité sociale », une « date de naissance », une « adresse », etc.

Dans la Figure 4.1, le concept « Digitale » est une sous-classe de « Caméra ». Les concepts « Lentille » et « Caméra » sont reliés entre eux par la relation sémantique « est_un_composant_de (Lentille, Caméra) » dans laquelle « Lentille » est le domaine et « Caméra » la portée (ou « range » en anglais).

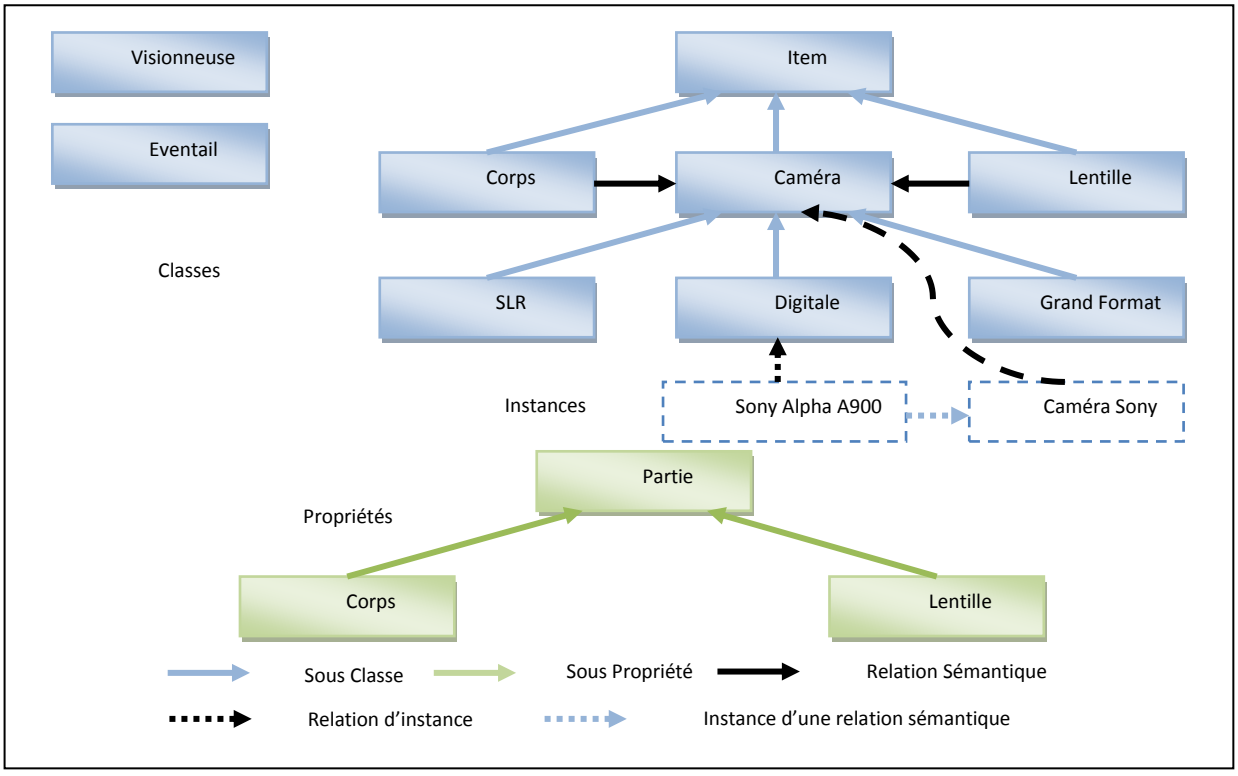


Figure 4.1 Représentation graphique de quelques concepts d'une ontologie

Dans l'exemple de la Figure 4.1, « Sony Alpha A900 » est une instance du concept « Digitale » et une relation sémantique « est_un » est instanciée entre cette instance et celle du concept « Caméra », i.e. «Caméra Sony». La base de connaissances contiendra donc les informations Digitale (Sony Alpha A900), Caméra (Sony) et est_un (Sony Alpha A900, Caméra Sony). L'action de définir et d'instancier une base de connaissances a été récemment appelée « peuplement d'ontologie ».

4.3.5. Rôle des ontologies

Les ontologies sont surtout utilisées pour la représentation de connaissance et l'application de raisonnement sur ces connaissances. Cependant, une ontologie possède des caractéristiques qui, au-delà de cette représentation, favorisent la réutilisation et le partage de données. Déjà en 1991, Gruber insistait sur le rôle que pouvaient tenir les ontologies pour favoriser la modularité et la réutilisabilité dans les systèmes informatiques [1]. Gruber souligne les difficultés techniques occasionnées par la conception d'ontologies communes. Ces idées ont été beaucoup approfondies et développées dans [11]. Pour lui les systèmes à base de connaissance mettent en place des techniques d'interopérabilité basées sur la communication et les opérations à partir de représentations formelles de la connaissance. Ils peuvent souvent être comparés à des agents qui négocient et échangent des connaissances.

Le partage et l'échange de données entre agents exigent le respect de certaines propriétés [10]. Pour l'auteur le rôle-clef d'une ontologie en extraction d'information est d'établir l'accord entre le descripteur recherché et les données.

Dans [4], une ontologie permet de définir les mots d'un langage naturel, les prédicats utilisés dans les calculs de prédicats, les types de concepts et de relations des graphes conceptuels, les classes d'un langage orienté objet ou les champs des tables d'une base de données relationnelle. Or la plupart de ces méthodologies sont connues et utilisées parce qu'elles favorisent l'échange et la réutilisation des connaissances.

Les ontologies servent à la représentation des données échangées dans un domaine particulier, afin de faciliter la communication interne au système informatique et externe entre les différents acteurs du domaine. Leur utilisation peut varier de la représentation des données à la recherche d'informations. Les ontologies ont été employées dans divers domaines et pour différents objectifs. Elles ont également porté leurs fruits au sein des systèmes à base de

connaissances et du Web sémantique. Nous énumérons dans ce qui suit, les principaux rôles que peuvent jouer les ontologies [73] :

— *La communication*

Il existe trois types de communication dans un projet : communication homme-homme, homme-système ou entre les différents modules du système. Ces trois types possèdent tous des caractéristiques particulières qui engendrent certains problèmes auxquels les ontologies peuvent apporter des solutions.

La communication entre humains pose surtout des problèmes quand les acteurs de cette communication ne sont pas du même domaine et ne parlent donc pas nécessairement le même langage. La réutilisation, le partage de connaissance et d'ontologies, suppose que plusieurs utilisateurs soient d'accord sur les ontologies partagées [15,16]. Une fois que les acteurs humains d'un projet sont d'accord sur une ontologie, la communication avec le système se fait naturellement, en utilisant cette ontologie. De plus l'adaptation des ontologies à la description de textes en langage naturel, semi-structurés [13] améliore la communication dans le sens homme-machine.

Les ontologies permettent le partage de la compréhension et la communication dans des contextes particuliers selon les besoins. Ainsi, on peut utiliser l'ontologie pour créer un réseau de relations qui définit les connexions entre les composants du système. Cette caractéristique de communication est offerte grâce à la non-ambiguïté des termes utilisés et définis par l'ontologie dans les systèmes.

— *L'ingénierie des systèmes*

L'ontologie peut servir divers aspects du développement des systèmes d'information. Dans l'ingénierie des systèmes, elle joue un rôle important sur trois aspects : la spécification, la fiabilité et la réutilisation.

— *L'interopérabilité*

La communication et l'échange d'informations et de ressources sous forme de programmes, de données ou de services, constituent un point important entre un ensemble des systèmes. Ces systèmes sont appelés systèmes distribués, par exemple, le réseau mondial Internet. Dans

ce domaine, un problème crucial concerne la capacité de deux applications (ou plus) à collaborer et à échanger de l'information afin d'atteindre un but global. On parle alors d'interopérabilité entre les systèmes pour la réalisation du but.

4.3.6. Types d'ontologies

La classification décrite dans [5,6,74], distingue trois (03) types d'ontologies suivant le type de conceptualisation :

— *Les ontologies terminologiques*

Spécifient les termes utilisés pour représenter la connaissance dans un domaine donné.

— *Les ontologies d'information*

Spécifient la structure des bases de données.

— *Les ontologies de représentation*

Des connaissances spécifient le modèle des connaissances du domaine considéré.

La seconde dimension classe les ontologies en quatre types par rapport à une tâche particulière :

— *Les ontologies d'application*

Contiennent toutes les définitions nécessaires pour modéliser la connaissance pour des applications particulières. Elles ne sont pas réutilisables en elles-mêmes puisqu'elles sont dédiées à ces applications.

— *Les ontologies de domaine*

Définissent des conceptualisations spécifiques à certains domaines. Les méthodologies d'ingénierie des connaissances font une distinction explicite entre ontologies de domaine et connaissances du domaine : alors que les connaissances du domaine décrivent des situations effectives dans un certain domaine, l'ontologie de domaine pose les contraintes sur la structure et le contenu (i.e. la grammaire et le vocabulaire) des connaissances du domaine.

— *Les ontologies génériques*

Sont similaires aux ontologies de domaine mais les concepts qu'elles définissent sont communs à plusieurs domaines. Elles définissent des concepts tels que les événements, les processus, les actions, les entités physiques.

— *Les ontologies de représentation*

Ces ontologies fournissent les primitives nécessaires à la description des ontologies génériques et de domaine. Elles fournissent des éléments de représentation pour les ontologies sans faire référence aux entités du monde réel.

Dans [9,17,18,19], les auteurs décrivent une classification des ontologies selon leurs niveaux de généralité :

— *Les ontologies de haut niveau (top-level ontologies)*

Elles décrivent les concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc., qui sont indépendants d'un problème ou d'un domaine d'application particulier.

— *Les ontologies de domaine* (*domain ontologies*)

Ces ontologies et les ontologies de tâche (task ontologies) décrivent, respectivement, le vocabulaire lié à un domaine générique (comme la médecine, ou les automobiles) ou une tâche ou une activité générique (comme le diagnostic ou la vente), en spécialisant les concepts présentés dans les ontologies de haut niveau. Elles donnent une représentation formelle des concepts du domaine étudié ainsi que des différentes relations qui lient ces derniers ; elle ne contient pas les concepts pédagogiques, narratifs et structurels.

— *Les ontologies d'application (application ontologies)*

Ces ontologies décrivent des concepts dépendant à la fois d'un domaine et d'une tâche particulière, qui sont souvent des spécialisations des deux ontologies relatives. Ces concepts correspondent souvent aux rôles joués par des entités de domaine tout en exécutant une certaine activité, comme l'unité remplaçable ou le composant disponible.

Les différents niveaux sont récapitulés dans la figure suivante :

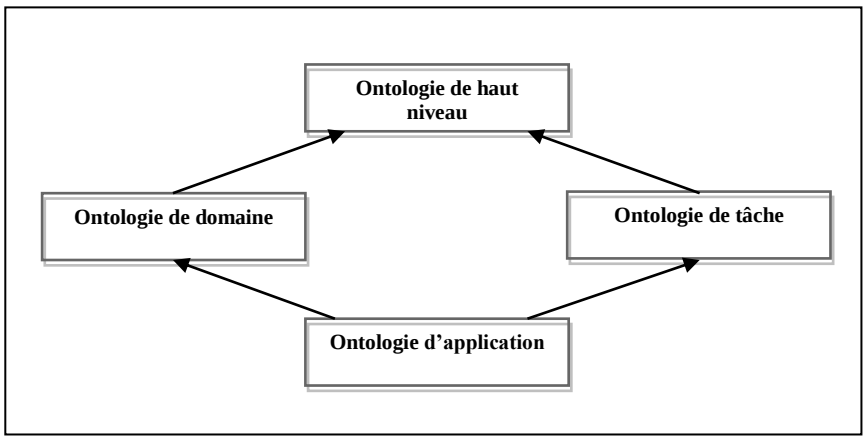


Figure 4.2 Différents types d'ontologie selon leurs degrés de dépendance vis-à-vis d'une tâche particulière ou d'un point de vue. *(Les flèches représentent des relations de spécialisation).*

Par conséquent, une ontologie peut être vue comme une théorie qui distingue les concepts particuliers, c'est-à-dire les objets concrets, physiques, les événements, les régions, etc., et les concepts universels, c'est-à-dire les propriétés, rôles, relations, états, etc.

4.3.7. Avantages d'utilisation des ontologies

Dans [5], les avantages d'une utilisation des ontologies sont divisés en trois catégories :

1. *L'assistance pour la communication*

L'ingénierie ontologique est une réponse aux besoins de communication entre personnes, entre personnes et systèmes, et entre systèmes. Les ontologies, par l'explicitation des concepts et des relations existant entre ces concepts, rendent la connaissance manipulable par tous les acteurs du système : elle apporte donc une interprétation consensuelle du système par les utilisateurs. Elles fournissent un vocabulaire commun sur lequel des descriptions peuvent être construites.

2. *L'interopérabilité entre les modules d'un système informatique*

Les ontologies sont utilisées comme support d'échange entre les différents modules des systèmes. Elles leur fournissent un accès commun à l'information et une compréhension

partagée des concepts. Elles amènent donc une utilisation et une réutilisation plus efficaces des sources de connaissances.

3. *Des améliorations pour l'ingénierie logicielle en terme de :*

- *Spécification.* Les ontologies peuvent aider dans la tâche d'identification des besoins et dans la spécification des systèmes à base de connaissances.

- *Fiabilité.* Une représentation formelle rend possible l'automatisation de tests de consistance, ce qui rend les logiciels plus fiables.

- *Réutilisabilité.* L'ingénierie ontologique définit la structure des connaissances d'un domaine. Une ontologie explicite les concepts d'un domaine, leurs propriétés et leurs relations. Par conséquent, en fonction de sa généricité, elle pourra être réutilisée dans un panel d'applications différentes.

- *Acquisition des connaissances.* Les ontologies permettent de réduire le goulot d'étranglement de l'acquisition des connaissances nécessaires pour enrichir la base de connaissances. En effet, une ontologie peut être utilisée comme base pour guider ce processus d'acquisition. Elle est un outil puissant pour la modélisation des connaissances.

4.3.8. Étapes de construction des ontologies

Dans [8,138], la construction d'une ontologie suppose certaines obligations qui découlent du choix d'utiliser certains concepts plutôt que d'autres pour représenter un phénomène. Ce sont les exigences ontologiques. C'est une tâche capitale dans la construction des ontologies puisque de la sélection de ces exigences découlent toutes les autres étapes de détermination du langage de connaissance et de construction de la base.

Dans [5], et d'après l'auteur, la construction d'ontologies s'inscrit dans un cycle de vie classique, composé de quatre phases nominales :

1. *la spécification*

Permet de fixer le but de la construction de l'ontologie et les utilisateurs de celle-ci. Elle fixe les limites du domaine à modéliser et l'utilisation qui sera faite des connaissances qu'elle permet de représenter.

2. *la conceptualisation*

Permet de structurer le domaine de connaissances à représenter. Il s'agit là de proposer un modèle identifiant et structurant les concepts du domaine d'étude.

3. *la formalisation*

Permet le passage du modèle conceptuel obtenu dans la phase de conceptualisation à un modèle formel. Cette phase amène à choisir un langage de représentation des connaissances. Le formalisme de ce langage est important et doit être choisi pour sa capacité à bien représenter les différents aspects de l'ontologie.

4. *l'implémentation*

Permet de transformer le modèle formel en une entité manipulable par un système informatique.

La phase de conceptualisation représente la plus grande partie du travail. Elle est primordiale puisqu'elle détermine la structure de l'ontologie et qu'il faut autant que possible qu'elle reste fixe. L'évolution de l'ontologie réside alors dans l'ajout d'extensions, c'est-à-dire l'ajout de nouveaux concepts par spécialisation des concepts déjà représentés. Il faut construire la hiérarchie des concepts puis spécifier les relations entre ceux-ci. Trois grandes approches existent pour cette phase :

1. *Les approches ascendantes* partent de textes, de documents techniques ou d'interviews afin d'en faire une analyse linguistique qui isole les termes importants du domaine. Les concepts les plus spécifiques sont alors identifiés et la structure est ensuite construite par généralisation de ces concepts.

2. *Les approches descendantes* partent de la tâche à réaliser pour expliciter les connaissances nécessaires à sa résolution. Les concepts les plus généraux sont

alors d'abord identifiés et la structure est ensuite construite par spécialisation. Cette approche est recommandée pour la réutilisabilité des ontologies qui représentent des considérations de haut niveau, ce qui est très intéressant pour la maintenance de cohérence.

3. *Les approches mixtes* identifient les concepts centraux puis les généralisent et les spécialisent pour développer la structure de l'ontologie.

Dans [14], il n'y a pas un consensus sur une méthodologie fixe pour la construction des ontologies, et selon l'auteur un processus de construction d'une ontologie opère selon quatre étapes fondamentales :

L'identification de l'objectif permet d'identifier, en termes généraux, l'objectif, la portée et les limitations de l'ontologie à construire.

La création de l'ontologie, l'étape la plus longue et la plus difficile, contient elle-même trois sous-étapes :

— *L'acquisition des connaissances* sert à définir les concepts dans un domaine donné et les relations entre eux, de manière à ne pas être ambiguës. Différentes techniques permettent de faire l'acquisition des connaissances. Elles peuvent se matérialiser par des entretiens informels avec des experts ou des entretiens structurés en vue de collecter des connaissances spécifiques et détaillées sur les concepts, leurs instances et leurs relations. Elles peuvent également être obtenues sous forme d'analyse informelle de texte pour définir les concepts fondamentaux ou bien sous forme d'une analyse formelle afin de définir les structures des connaissances.

— *Le codage*, une fois les concepts et leurs relations acquises, permet de représenter l'ontologie dans un langage formel. La formalisation de l'ontologie peut être de différents degrés.

• *Très informel* : l'ontologie s'exprime dans le langage naturel ;

• *Semi-informel* : l'ontologie s'exprime dans une forme structurée du langage naturel ;

• *Semi-formel* : l'ontologie est exprimée dans un langage artificiel défini formellement ;

• *Rigoureusement formel* : l'ontologie est exprimée dans un langage formel utilisant une sémantique formelle avec des théorèmes et preuves.

— *L'intégration des ontologies existantes* est l'étape qui permet de réutiliser les concepts déjà définis dans des ontologies existantes.

4.4. Méthodes de conception des ontologies

Le processus de développement d'une ontologie est un processus complexe où plusieurs acteurs interviennent dans les différentes étapes du processus. Il s'agit donc d'une équipe pluridisciplinaire. Pour cela, il faut utiliser des méthodes ou méthodologies pour seconder le processus de construction des ontologies [78].

Cependant, dans [79], et selon l'auteur, il n'existe pas une méthodologie parmi celles qui sont proposées dans la littérature qui est complètement maturée par rapport aux méthodologies du génie logiciel ou de l'ingénierie des connaissances. Les méthodes et les méthodologies recensées permettent la construction d'ontologies à partir de zéro (from scratch), c'est-à-dire, à partir des données brutes ou par réutilisation d'autres ontologies, la réingénierie, l'intégration ou la fusion avec d'autres ontologies, la construction collaborative ainsi que l'évolution des ontologies construites. La conception d'ontologies est une tâche difficile qui nécessite la mise en place de procédés élaborés afin d'extraire la connaissance d'un domaine, manipulable par les systèmes informatiques et interprétable par les êtres humains. Deux types de conception existent : la conception entièrement manuelle et la conception reposant sur des apprentissages [50].

4.4.1. Conception manuelle

Les méthodes présentées dans cette section sont celles utilisées pour la construction des ontologies à partir de zéro (from scratch) ou par réutilisation d'autres ontologies où la conception est réalisée manuellement.

— **La méthode d'Uschold et King** [80,81]

Ils ont proposé une première méthode de construction d'ontologie inspirée de leur expérience acquise lors du développement des ontologies dans le domaine de la gestion des entreprises (Enterprise Ontology). La figure suivante présente le processus de la méthode. Cette dernière repose sur les quatre étapes suivantes :

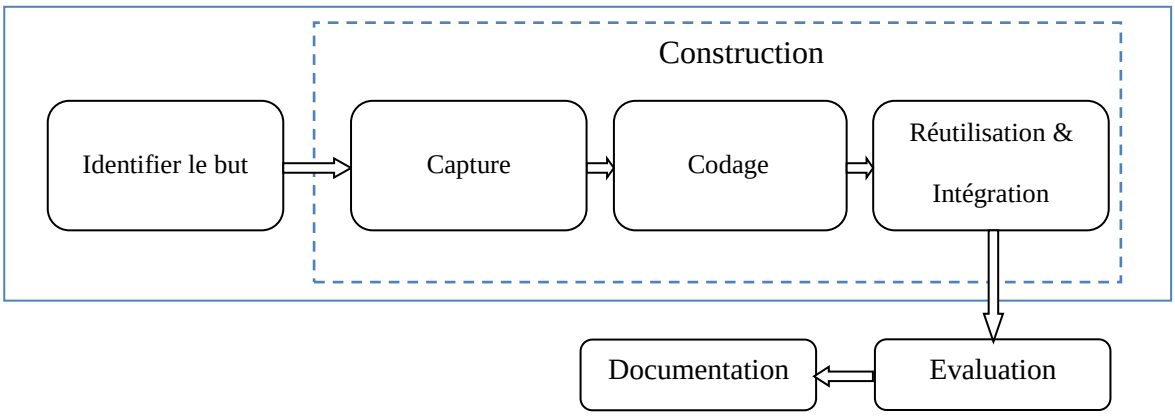


Figure 4.3 Processus de la méthode manuelle proposée par Uschold et King

Etape1 : Identifier le but et la portée de l'ontologie dont les raisons pour lesquelles l'ontologie en cours de construction sont clarifiées ainsi que les utilisateurs potentiels de l'ontologie ;

Etape2 : Construire l'ontologie. Cette étape est divisée en trois activités qui sont :

— Capture de l'ontologie : identifier les concepts et les relations fondamentaux, produire des définitions précises et non ambiguës à ces éléments en langage naturel, identifier les termes dénotant ces éléments et enfin essayer d'arriver à un agrément.
— Codage de l'ontologie : la représentation explicite de la conceptualisation dans un langage formel ;
— Réutiliser et intégrer éventuellement des ontologies existantes. Cette activité peut être, effectuée en parallèle avec l'activité de capture et/ou de codage ;

Etape3 : Evaluer l'ontologie ;

Etape4 : Documenter l'ontologie.

— **La méthode de Grüninger et Fox** [82]

Cette méthode est basée sur l'expérience du développement de TOVE (*Toronto Virtual Enterprise*) *Project Ontology*, qui est une ontologie dans le domaine de la modélisation des activités et des processus d'affaires :
Etape1 : capturer les scénarios motivants : le développement d'ontologies est motivé par des scénarios qui se présentent dans l'application. Les scénarios motivants sont des problèmes qui ne sont pas abordés de manière adéquate par les ontologies existantes. Un scénario motivant propose ainsi un ensemble de solutions possibles intuitivement pour les problèmes présentés dans ces scénarios. Ces solutions offrent une sémantique informelle destinée aux objets et aux relations qui seront ultérieurement inclus dans l'ontologie. Toute proposition d'une nouvelle ontologie ou d'une extension d'une ontologie doit décrire un ou plusieurs scénarios motivants ;

Etape2 : formuler les questions de compétence de façon informelle : ils sont basés sur les scénarios obtenus dans l'étape précédente. Une ontologie doit être en mesure de représenter ces questions en utilisant sa terminologie, et être en mesure de caractériser les réponses à ces questions en utilisant les axiomes et les définitions. Ces questions de compétences sont informelles, car elles ne sont pas encore exprimées dans un langage formel d'ontologie ;

Etape3 : spécification de la terminologie de l'ontologie en langage formel ;

Etape4 : formuler les questions de compétence de façon formelle, en utilisant la terminologie de l'ontologie : Une fois que les questions de compétence ont été posées d'une manière informelle et la terminologie de l'ontologie a été définie, les questions de compétence sont définies formellement ;

Etape5 : spécification des axiomes et des définitions pour les termes de l'ontologie en langage formel : Les axiomes de l'ontologie spécifient les définitions des termes et les contraintes sur leur interprétation. Les axiomes doivent être fournis pour définir la sémantique, ou le sens de ces termes ;

Etape 6 : établir des conditions pour caractériser la complétude de l'ontologie : Une fois que les questions de compétence ont été formellement décrites, il faut définir les conditions dans lesquelles les réponses à toutes les questions soient complètes.

4.4.2. Conception automatique

Ces méthodes focalisent sur l'étape d'acquisition automatique afin de minimiser le coût et l'effort du travail manuel. Ainsi, ces outils permettent d'extraire les termes candidats d'un domaine et leurs relations à partir d'un corpus textuel.

Néanmoins, cela nécessite l'intervention d'un expert du domaine pour la sélection et la validation des termes obtenus.

À ce sujet, on peut citer à titre d'exemple la méthode ARCHONTE et le travail d'Audrey Baneyx basé sur cette dernière ainsi que TERMINAE. Nous décrivons ces méthodes dans cette section.

— **La méthode ARCHONTE** (ARCHitecture for ONTological Elaborating)

Cette méthode a été mise au point par B. Bachimont [68], au sein du groupe *Terminologie et Intelligence Artificielle* [3] , pour construire des ontologies s'appuie sur la sémantique différentielle. La figure suivante présente le processus de la méthode. La construction d'une ontologie comporte trois étapes :

Etape1 : Normalisation

Choisir les termes pertinents du domaine et normaliser leur sens, à partir d'un corpus textuel qui est la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur est associé, puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père ;

Les principes différentiels sont :

[3] http://estime.spim.jussieu.fr/TIA/

- *le principe de communauté avec le père :* il faut expliciter en quoi le fils est identique au père qui le subsume ;
- *le principe de différence avec le père :* il faut expliciter en quoi le fils est différent du père qui le subsume. Puis que le fils existe, c'est donc qu'il est distinct du père ;
- *le principe de différence avec les frères :* il faut expliciter la différence de la notion considérée avec chacune des notions sœurs car toute notion doit se distinguer de ses sœurs sinon il n'y aurait pas lieu de la définir ;
- *Le principe de communauté avec les frères :* il faut expliciter la communauté existante entre la notion considérée et chacune des notions sœurs. Ce principe de communauté doit être différent du principe de communauté existant avec le parent. La communauté entre les notions filles doit permettre de définir des différences mutuellement exclusives entre les notions filles.

Ces principes différentiels permettent de fixer le cadre interprétatif des concepts. Cela revient à associer aux termes une signification qui fasse abstraction des variations de sens liées aux différents contextes textuels dans lesquels ils peuvent apparaître. Les concepts sont donc normés puisqu'ils sont décrits selon un certain point de vue, en l'occurrence celui de la tâche à réaliser. Par conséquent, cela permet de passer à « l'ontologie différentielle ».

Etape2 : Formalisation

Formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation… En d'autres termes, il s'agit de définir des concepts selon une sémantique formelle et extensionnelle (c.à.d. les concepts sont liés à un ensemble de référents dans le monde) et de formaliser les relations qui existent entre les concepts en définissant leur arité et les ensembles d'extensions de concepts qu'elles relient. Il faut passer de la dimension linguistique et interprétative de la taxinomie à « l'ontologie référentielle » ou « l'ontologie formelle » composée de concepts dont le sens est décontextualisé.

Etape3 : Opérationnalisation

L’opérationnalisation dans un langage de représentation des connaissances. Cette étape marque le passage à « l’ontologie computationnelle ».

4.5. Construction d’ontologies à partir des textes

La construction d’ontologies à partir de textes constitue un sous-domaine à part entière de l’ingénierie des ontologies. Le recours aux textes est légitimé par les travaux menés en linguistique dont l’hypothèse principale est que les textes sont porteurs de connaissances stabilisées et partagées par des communautés de pratiques. En outre, même s’ils ne les remplacent totalement, les textes sont plus facilement disponibles que les experts qui manquent de temps pour participer au processus de construction [75,76].

Un cadre méthodologique en quatre étapes : *constitution d’un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l’ontologie*, est commun à la plupart des méthodes de construction d’ontologies à partir de textes. Ces étapes, relativement indépendantes, réalisent un double mouvement permettant de passer du niveau textuel (la connaissance est décrite dans des corpus) au niveau conceptuel (la connaissance est décrite via des concepts dénotés par les entités linguistiques et les relations entre ces concepts) et de l’informel vers le formel.

— **Extraction d’information basée sur les ontologies**

L'articulation entre textes et ontologies peut être multiple. En effet, en vue de traitements automatiques plus efficaces, les textes peuvent être considérés comme sources de connaissances pour enrichir les ontologies et réciproquement [74].

— **Des textes vers les ontologies**

Nous pouvons résumer l'enrichissement de l’ontologie à partir des textes en deux points :

- Les ontologies peuvent être construites en se basant sur le texte comme source de connaissance. Dans ce cas, le terme mentionné dans la littérature est souvent « ontology learning » (apprentissage d'ontologie) où les approches proposées cherchent à automatiser le plus possible ce procédé en se basant souvent sur le traitement automatique de la langue naturelle et sur des connaissances linguistiques.

- Les textes peuvent aussi contenir, par exemple, des instances de concepts permettent d'enrichir une ontologie existante. Dans ce cas, le terme mentionné dans la littérature est souvent « ontology population » (peuplement d'ontologie). Dans ce cadre, il s'agit d'un typage d'instances de concepts. Les travaux utilisant le terme « peuplement d'ontologie » pour le « typage d'instances » considèrent que ces informations extraites font partie de l'ontologie elle-même et non d'une base de connaissances.

— **Des ontologies vers les textes**

L'apport des ontologies pour les textes correspond aux annotations sémantiques (semantic annotation ou knowledge mark-up). Il s'agit de caractériser le contenu informationnel à l'aide d'une ontologie ou d'une base de connaissances. D'une manière plus simple, cela correspond à faire un étiquetage sémantique du texte ou de portions du texte à l'aide d'instances de concepts ou avec des relations les reliant. Cet étiquetage suit un ou plusieurs schémas d'annotations définis par une (ou plusieurs) ontologie(s) correspondant aux tâches pour lesquelles cette annotation sémantique est construite.

4.6.Langages et plates-formes pour les ontologies

Il existe de nombreux langages informatiques, plus ou moins récents, spécialisés dans la création et la manipulation des ontologies. Nous en décrivons quelques-uns dans la suite.

— **XML RDF et OWL**

Les ontologies peuvent être décrites en XML (Extensible Markup Language). C'est un langage qui permet de décrire des méta-données en facilitant leurs traitements et leurs échanges.

D'autre part, RDF (Resource Description Framework) est un modèle de graphe destiné à décrire, de façon formelle, les ressources Web et leurs méta-données et permettre le traitement automatique de telles descriptions.

DFS (Resource Description Framework Schema) est un langage extensible qui permet la représentation des connaissances. Il appartient à la famille des langages du Web sémantique publiés par le W3C. Il fournit des éléments de base pour la définition d'ontologies ou de vocabulaires destinés à structurer des ressources RDF.

Cependant, RDF et RDFS souffrent de limites, comme l'impossibilité de raisonner et de mener des raisonnements automatisés sur les modèles de connaissances établis à l'aide de ces langages. En conséquence, un nouveau langage, OWL (Web Ontology Language), est apparu.

Plus tard OWL (Web Ontology Language) est apparu. C'est un dialecte XML fondé sur une syntaxe RDF. Il fournit les moyens pour définir des ontologies Web structurées. Il se différencie du couple RDF / RDFS par le fait que c'est un langage d'ontologies, contrairement à RDF.

De nombreux éditeurs d'ontologies sont apparus. Protégé[4] est l'un des éditeurs d'ontologie les plus utilisés. Il peut lire et sauvegarder des ontologies dans la plupart des formats d'ontologies : RDF, RDFS, OWL.

— **LOOM**

LOOM est une plate-forme pour la représentation des connaissances. Son objectif principal est de construire des applications intelligentes. Les connaissances déclaratives dans LOOM sont composées de définitions, de règles, de faits, etc. Pour compiler les connaissances déclaratives,

LOOM utilise un moteur déductif. Ce dernier est un classifieur qui utilise le chaînage-avant, l'unification sémantique et des technologies orientées objet.

— **ONTOLOGUA**

Ontolingua [5] est un mécanisme qui permet aux utilisateurs de créer et manipuler des ontologies. Il supporte les ontologies portables pour qu'elles soient traduites dans différents systèmes. Ontolingua est basé sur le langage KIF (Knowledge Interchange Format). Celui-ci est conçu pour l'échange de connaissances entre des systèmes informatiques répartis.

— **OIL**

OIL (Ontology Inference Layer)[6] est un langage dédié à la spécification et à l'échange des ontologies sur le Web. Il permet la représentation et l'inférence d'ontologies, en combinant des primitives de modélisation des langages de frame avec la sémantique formelle et les

[4] http://protege.stanford.edu/
[5] http://www.ksl.stanford.edu/software/ontolingua/
[6] http://www.ontoknowledge.org/oil/

modes de raisonnement des logiques descriptives. Il se base sur des formalismes tels que RDF/RDFS et XML, ce qui garantit sa totale compatibilité avec ces formalismes standards ou des formalismes en cours de standardisation. Les liens existant entre la structure d'un document et la modélisation du domaine couvert par ce document sont étudiés dans [54,55] à travers d'une comparaison entre OIL et les schémas XML.

— **SHOE**

SHOE (Simple HTML Ontology Extensions) est une extension du langage HTML qui permet aux auteurs de pages Web de générer une annotation de leurs documents, compréhensible par la machine. Ce langage peut être utilisé par des agents pour la gestion des pages Web [57].

4.7. Conclusion

Dans ce chapitre, nous avons présenté brièvement les notions les plus couramment utilisées dans le domaine d'ingénierie ontologique qui est vu comme sous-domaine de l'ingénierie des connaissances, en focalisant sur les méthodes de construction des ontologies dues à leur importance dans notre travail, et qui constituent l'un de nos axes de recherche sur lequel s'articule l'identification d'opinion.

Pour construire une ontologie, nous avons présenté quelques méthodes manuelles et quelques méthodes automatiques basées sur l'apprentissage, ces méthodes sont proposées par certains chercheurs et chacune d'elles présente des points faibles et des points forts, sachant qu'il n'existe pas une méthode consensuelle pour la construction des ontologies.

Chapitre 5

ONTOMART : Notre approche proposée

5.1. Introduction

Dans les dernières années, et notamment avec l'apparence du Web 2.0 qui a permet aux développeurs de mettre à la disposition des internautes des outils simples et efficaces comme les réseaux sociaux et les blogs, l'information est devenue accessible et à la portée de tout le monde. À travers ces outils, les utilisateurs peuvent poster librement des commentaires, des jugements, ou des critiques sur une personne, un service ou un produit.

Pour bénéficier des textes évaluatifs disponibles, en vue d'améliorer la qualité d'un produit ou d'un service offert aux clients ou de connaître l'avis du grand public sur une personne politique, etc., les entreprises, les politiciens, ainsi que les clients ont besoin d'outils puissants pour suivre les opinions, les sentiments, les jugements et les croyances qui se cachent derrière ces textes.

La nouvelle discipline, fouille d'opinion et analyse des sentiments (*ang. Opinion Mining and Sentiment Analysis*), est née pour répondre à la demande croissante en qualité d'analyse d'opinions, en mettant en place des systèmes capables de traiter automatiquement des textes évaluatifs, et rechercher les valeurs prédictives des jugements. Dans ce contexte, plusieurs approches ont été proposées par les chercheurs, et plusieurs ressources lexicales ont été construites.

La fouille de données d'opinion est un domaine qui s'intéresse essentiellement à la détection et la catégorisation d'opinions. La prise en compte du contexte des mots dans les évaluations, le domaine étudié, ainsi la complexité grammaticale et morphologique de certaines langues comme l'arabe dans notre étude, peuvent apporter une forte ambiguïté, pas seulement au niveau sémantique mais aussi au niveau subjectif, qui rend difficile la tâche d'identification des opinions et leurs cibles.

Les approches, qui se basent sur les lexiques des sentiments (Section 3.2.1 du chapitre 3) pour détecter les indices de subjectivité dans les textes, attribuent une polarité positive ou négative à chaque entrée sans la prise en compte du domaine étudié. Contrairement à ces approches, nous proposons dans notre étude une approche d'identification d'opinions basée sur l'exploration ontologique du domaine étudié, qui vise essentiellement à résoudre le problème d'ambiguïté subjective dans la phase d'identification de la subjectivité, en employant pour chaque domaine étudié, une ontologie couplée avec un lexique de sentiments, où l'attribution de polarité aux mots de lexique dépend fortement du domaine étudié. Nous pouvions résumer les grandes lignes de notre approche dans les points suivants :

(1) Extraction des caractéristiques explicites et implicites du domaine étudié, à partir des textes évaluatifs, soit par projection simple de l'ontologie du domaine, soit par exploration des relations sémantiques de l'ontologie ;

(2) Identification des expressions d'opinion en employant le lexique de sentiments avec un ensemble de règles linguistiques ;

(3) Association des expressions d'opinions identifiées dans la phase (2) avec les caractéristiques extraites dans la phase (1). Notons que les caractéristiques associées aux expressions d'opinion sont les cibles de passage d'opinions dans les textes évaluatifs ;

(4) Classification des opinions identifiées au niveau des caractéristiques (*ang. Feature-level classification*).

Le chapitre est organisé comme suit : nous présentons dans la section 5.2, les motivations de la démarche de notre travail en citant les principales approches proposées dans le domaine de

fouille d'opinion. Dans la section 5.3, nous présentons une description générale de notre système. Nous décrivons dans la section 5.4, les entrées de notre approche qui sont : l'ontologie du domaine, le lexique des sentiments, et l'ensemble des textes évaluatifs. Enfin, dans la section 5.5, nous expliquons l'architecture de notre approche et ses différents modules illustrés par des exemples.

5.2. Motivation

Le domaine de fouille d'opinions réunis différentes problématiques de recherche : construction de ressources lexicales [102, 134, 136], classification d'opinions [43, 90, 97, 111, 129, 132, 135], détection de passage d'opinions et leurs catégorisations sémantiques [44, 130], résumé automatique de textes d'opinion [129,127], etc.

Dans le domaine de fouille d'opinions, deux grandes familles d'approches se distinguent : celles qui font une simple extraction des caractéristiques d'objets [104, 122, 130, 135] et de celles qui les organisent en une hiérarchie à l'aide des taxonomies [123, 124, 125] ou ontologies [126, 127, 128, 129]. Le processus d'extraction des caractéristiques concerne principalement les caractéristiques explicites.

Cependant, les travaux utilisant des ontologies du domaine exploitent l'ontologie comme une taxonomie en utilisant seulement la relation « est un » entre les concepts. Ils n'ont pas vraiment utilisé toutes les données stockées dans l'ontologie, comme les composants lexicaux et d'autres types de relations. Nous croyons que nous pouvons obtenir plus d'avantages si nous explorons toutes les capacités offertes par les ontologies.

5.3. Description

Dans la section précédente (section 5.2) de ce chapitre, nous constatons que la plupart des chercheurs qui ont utilisé des modèles de représentation des connaissances (taxonomies ou ontologies) pour représenter les concepts des domaines étudiés, ont approuvé une amélioration remarquable de la performance de leurs systèmes d'extraction d'opinions.

Cependant, les travaux de recherche, qui ont utilisé les ontologies de domaine pour la représentation des connaissances, ont les exploité comme des taxonomies en utilisant

uniquement la relation « est un » (*ang. is-a*) entre les concepts. Ils n'ont pas vraiment utilisé ce que l'ontologie peut offrir comme relations sémantiques.

Nous croyons que l'utilisation adéquate des ontologies peut aider à améliorer le processus d'extraction des concepts de domaine, si nous exploitons parfaitement la couche sémantique qui peut offrir l'ontologie.

Dans le contexte de notre approche, nous proposons un système d'identification d'opinions à base d'ontologies dont l'objectif d'utilisation des ontologies est relatif à deux aspects :

- Représentation et réutilisation des concepts du domaine étudié ;
- Extraction des concepts explicites et implicites à partir des textes évaluatifs.

Pour chaque domaine étudié, notre approche utilise une ontologie du domaine pour la conceptualisation des connaissances, et un lexique de sentiments pour la fouille des expressions d'opinion dans les textes évaluatifs.

Nous avons opté pour une architecture modulaire, où chaque module est décrit séparément des autres, qui enfin vont coopérer pour décrire l'architecture générale du système.

5.4. Entrées de notre système

Nous découpons les entrées de notre approche autour des trois axes suivants : Définition de l'ontologie du domaine, la construction de lexique des sentiments et la collecte des textes évaluatifs.

5.4.1. Ontologie du domaine

Notre système s'appuie sur l'utilisation des ontologies de domaine. Par opposition aux autres types d'ontologies, les ontologies de domaine se limitent à représenter la connaissance d'un domaine particulier. Notre choix est motivé par le fait que les ontologies de domaine restreignent l'interprétation des concepts qu'elles définissent au contexte spécifié par le domaine. Cela a l'avantage de limiter l'ambiguïté des termes définis dans l'ontologie pour référencer les concepts, facilitant ainsi leur détection dans les textes.

Les principaux objectifs d'utilisation des ontologies de domaine dans notre système sont :

— *Structuration des concepts*

La structuration des concepts dans une ontologie permet de définir les termes les uns par rapport aux autres, chaque terme étant la représentation textuelle d'un concept.

Comme nous l'avons vu dans le chapitre 4 (section 4.3.1) de ce présent manuscrit, les ontologies sont des outils qui fournissent beaucoup d'informations sémantiques. Ils aident à définir les concepts, les relations et les entités qui décrivent un domaine avec un nombre illimité de concepts. Cet ensemble de concepts peut être une ressource lexicale importante pour extraire les caractéristiques explicites et implicites d'un produit ou d'un objet. Par exemple, dans la critique suivante :

طريق طويل **و** متعرج (une route longue et sinueuse)

Comme indiqué dans la figure 5.1, le mot d'opinion négative متعرج (sinueuse) est ambigu, car il n'est pas associé à une caractéristique lexicalisée. Toutefois, si le terme طويل (longue) est stocké dans l'ontologie comme instance de la propriété مسافة (distance) du concept طريق (route), le mot d'opinion متعرج (sinueuse) peut être facilement associé à la caractéristique طريق (route) (à noter que la conjonction joue un rôle important dans le processus de désambiguïsation).

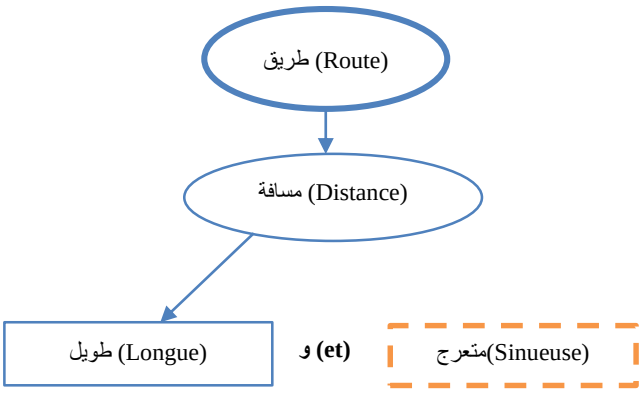


Figure 5.1 Exemple d'association d'une propriété non instanciée à un concept ontologique

Nous discutons ce point en détail dans la phase d'extraction des cibles de passage d'opinion (section 5.5.2) de ce chapitre.

— *Extraction des caractéristiques*

Les ontologies fournissent une structuration pour les caractéristiques grâce à la représentation hiérarchique et grâce à la capacité à définir les relations reliant nombreux concepts. C'est également une ressource précieuse pour structurer les connaissances de domaine lors de la tâche d'extraction des caractéristiques. En outre, les relations entre les concepts et l'information lexicale peuvent être utilisées pour extraire des caractéristiques implicites.

Par exemple, si le concept سيارة (voiture) est lié au concept طريق (route) par la relation عبر (franchir), une opinion négative envers الطريق (route) peut être extraite de la phrase :

عبرت السيارة بصعوبة (la voiture a franchi avec difficulté).

Deux objectifs principaux sont derrière l'utilisation des ontologies du domaine dans notre approche :

- Extraction des caractéristiques explicites par projection simple de l'ontologie ;
- Extraction des caractéristiques implicites par exploitation de la couche sémantique de l'ontologie ;

Dans la section 5.5.2, nous détaillons les étapes d'extraction des caractéristiques explicites et implicites à partir des textes évaluatifs en explorant l'ontologie du domaine.

Comme nous l'avons vu dans le chapitre 4, il existe plusieurs techniques pour construire une ontologie : automatique, semi-automatique ou manuelle. Dans notre approche, nous nous intéressons pas beaucoup de techniques de construction des ontologies, mais plutôt à ce que l'ontologie peut apporter comme avantages à notre approche dans la phase d'extraction des caractéristiques explicites et implicites, contrairement aux approches qui se limitent seulement à l'extraction des caractéristiques explicites, qui sont appuyées sur l'utilisation des vocabulaires du domaine.

5.4.2. Lexique de sentiments

Pour exprimer nos idées, nos préférences ou nos opinions sur un sujet, ou sur un objet ou l'un de ses caractéristiques, nous utilisons un langage évaluatif. Ce langage incorpore différents éléments linguistiques qui peuvent permettre la détection et l'analyse de la subjectivité. Ces éléments découlent de la théorie de l'évaluation cognitive (*ang. Appraisal Theory*) qui suggère que les opinions et les sentiments en générale proviennent des évaluations cognitives personnelles que l'individu exprime sur un objet ou un évènement.

L'ensemble des mots qui expriment l'opinion peuvent être subdivisés en catégories représentant la modélisation. La modélisation se définit comme l'inscription dans le discours de la prise d'attitude du locuteur à l'égard du contenu d'un énoncé. Les cinq catégories de modalité de l'évaluation sont :

— l'**opinion** : un fait présupposé est évalué par le locuteur qui révèle du même coup son point de vue ;

— le **jugement** (favorable ou défavorable) : le locuteur pose dans son énoncé, une action réalisée et juge si cet acte est bon ou mauvais ;

— l'**accord** ou le **désaccord** : on présuppose qu'une demande a été adressée au locuteur qui va dire s'il adhère ou non à la vérité d'un propos tenu par un autre ;

— l'**appréciation** (jugement intellectuel) : le locuteur évalue la valeur d'un fait en révélant ses propres sentiments selon son champ d'appréciation ;

— l'**acceptation** ou le **refus** : on présuppose qu'une demande d'accomplissement d'un acte a été adressée au locuteur qui peut répondre favorablement ou non à cette demande.

Dans chacune de ses modalités, la valeur peut être graduée selon sa force : soit du négatif au positif dans les cas de l'opinion, du jugement ou de l'appréciatif soit du plus ou moins pour l'accord/désaccord et acceptation/refus.

À ces modalités, viennent se combiner trois types de vocabulaire de l'opinion :

— le vocabulaire **affectif** (l'ensemble des mots impliquant une réaction émotionnelle : effrayant, pleurer, séduire, étonnant) ;

— le vocabulaire **appréciatif** (impliquant un jugement de valeur positif ou négatif : intéressant, de qualité, trop cher, banalité) ;

— le vocabulaire **connoté** (mots possédant une signification affective en plus de son sens premier : force, liberté, original).

Trancher entre le caractère objectif ou le caractère subjectif d'un texte revient à détecter dans le texte les indices de la présence du subjectif. Plusieurs indices peuvent en fait témoigner de la présence de la subjectivité dans un texte ou une phrase. La présence d'adjectifs constituerait un bon indicateur du caractère subjectif des phrases [131].

D'autres proposent d'identifier des indices co-occurrents de subjectivité pour en déduire le caractère subjectif des phrases [34]. On peut aussi tenter de reconnaître la subjectivité d'un texte en utilisant différents indices et caractéristiques (emplacement des mots, etc.)[132].

Enfin, certains ont tenté de séparer les opinions des faits, au niveau du document et au niveau de la phrase, en utilisant un classificateur naïf Bayes [98].

Bing Liu [93] admet aussi qu'en matière d'analyse de l'opinion, la distinction entre l'objectif et le subjectif n'est pas forcément pertinente. Ainsi, une phrase contenant de la subjectivité peut ne pas exprimer une opinion ni positive ni négative. Ex. : *Je pense qu'elle me l'a dit.* Aussi, une phrase objective peut être porteuse d'opinion, ce qui n'est pas vraiment traité par l'analyse des sentiments. L'expression peut comporter une opinion implicite ou transmettre une information qui peut produire chez le lecteur des effets négatifs d'appréciation de la cible. Ex. : Mon téléviseur X est tombé en panne hier.

Traditionnellement, l'extraction d'opinion dans les textes est basée sur la recherche des adjectifs, pour cela, les méthodes existantes sont souvent basées sur des dictionnaires

généraux. Malheureusement, ce type d'approche trouve ses limites : pour certains domaines, des adjectifs peuvent être inexistants, voire contradictoires [44].

Les approches lexicales utilisent des dictionnaires de mots subjectifs, considérés comme des références universelles à partir de l'anglais. Ces dictionnaires peuvent être généraux comme le *General Inquirer*[1], *Sentiwordnet*[2], *Opinion Finder*[3], *NTU Sentiment Dictionary (NTUSD)*[4], etc. Dans ces dictionnaires, une polarité est associée *a priori* à chacun des mots. Quel que soit le contexte dans lequel il sera inséré, le mot devrait ainsi avoir toujours la même polarité. On donne ensuite au document un score d'opinion en fonction de la présence de mots issus de ces dictionnaires dans le texte.

a) Initialisation du dictionnaire

Dans notre approche, nous proposons de construire pour chaque domaine étudié un lexique de sentiments. Nous croyons que la séparation des lexiques par domaine peut diminuer considérablement l'ambiguïté sémantique des mots d'opinion. Par exemple le mot صغير (petit) peut être positif dans un domaine et négatif dans un autre.

Exemple :

الهاتف صغير (le mobile est petit) صغير (petit) est positif dans le domaine d'électronique.

الغرفة صغيرة (la chambre est petite) صغير (petit) est négatif dans le domaine d'urbanisme.

Une idée inspirée des travaux de Harb et al. (2008) [44] consiste à construire un lexique avec les mots de sentiments les plus connus dans le domaine étudié. Pour cela, nous considérons deux ensembles P et N de mots germes classiquement utilisés dans la littérature dont les orientations sémantiques sont respectivement positives et négatives.

Exemple :

P = {جيد، جميل، ممتاز ،ايجابي ،صحيح، محبوب} {aimable, correct, positive, excellent, beau, bien} ;

[1] http://www.wjh.harvard.edu/~inquirer/
[2] http://sentiwordnet.isti.cnr.it/
[3] http://www.cs.pitt.edu/mpqa/opinionfinder.html
[4] http://onlinelibrary.wiley.com/doi/10.1002/asi.20630/full

N = {سيء، قبيح ،رديء ،خاطئ، مكروه} {désagréable, faux, mauvais, laid, mauvais} ;
Le dictionnaire aura alors pour chaque entrée : le mot de sentiment et sa polarité positive ou négative.

Mot de sentiment	Polarité
جيد	+
جميل	+
سيء	-
قبيح	-
…etc.	

Tableau 5.1 Structure simplifiée du dictionnaire des sentiments

b) Enrichissement du dictionnaire

Certaines approches de détection des sentiments utilisent des dictionnaires généraux existants [135, 136] ou construits manuellement [134], d'autres utilisent des dictionnaires construits automatiquement, comme dans [44], où les auteurs ont proposé une nouvelle approche de création automatique de dictionnaire d'adjectifs qui intègre les connaissance du domaine, puis l'enrichissement de ce dictionnaire à partir d'un corpus d'apprentissage.

Afin d'établir des associations entre les différentes expressions d'opinion pour enrichir le dictionnaire d'opinions, il est tout d'abord nécessaire de connaître la fonction grammaticale de chacun des mots du corpus d'apprentissage. Pour ce faire, un outil d'étiquetage automatique de textes qui attribue à chaque mot une catégorie grammaticale et fournit les mots sous une forme lemmatisée (forme canonique) est indispensable dans ce cas.

En fait, la création des ressources lexicales constitue un sous-domaine de recherche à part [44,134,35,136], et pour ne pas s'écarter de notre sujet de recherche qui focalise sur l'identification d'opinions par exploration des ontologies, nous ne donnerons pas une grande importance à la création du dictionnaire des sentiments, et nous nous limiterons à l'adaptation d'un lexique existant, qui regroupe une liste prédéfinis d'expression d'opinion.

5.4.3. Textes évaluatifs

Afin d'évaluer l'efficacité de notre approche, nous aurons besoin d'un ensemble de textes évaluatifs propres au domaine étudié. Le réseau Internet est considéré comme la source la plus riche pour collecter les textes à caractère subjectif : les réseaux sociaux comme Facebook et Twitter, les blogs, les sites de commerces électroniques,…etc. Les méthodes de collecte des textes peuvent être manuelles ou automatiques.

Nous croyons qu'une opinion est exprimée dans le même segment dans lequel les caractéristiques sont présentes, i.e. les caractéristiques et les expressions d'opinion sont près les uns des autres. Dans ce cas, la segmentation du texte joue le rôle d'éloigner les caractéristiques inutiles des expressions d'opinion.
Pour extraire les unités lexicales et déterminer si une unité est une caractéristique ou un mot d'opinion, plusieurs tâches linguistiques peuvent être appliquées :

a) Segmentation

Pour analyser un texte, nous devons mener à bien sa segmentation en paragraphes, phrases et propositions. Dans notre travail, nous avons utilisé *Stanford Segmenter* pour segmenter le texte en segments bruts. La figure suivante illustre un texte arabe (zone de texte en haut) segmenté en segments bruts (zone de texte en bas).

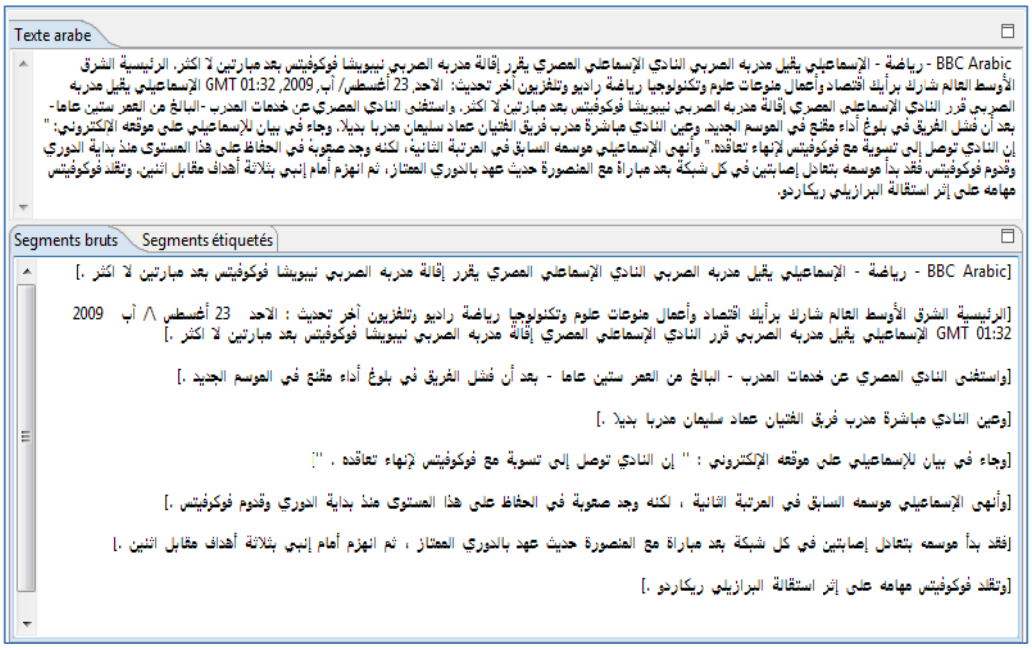


Figure 5.2 Exemple d'un texte arabe segmenté

b) Étiquetage

Après la segmentation du texte (étape précédente), nous devons associer à chaque terme son étiquette lexicale. L'étiquetage (*ang. POS Tagging : Part-Of-Speech Tagging*) est le processus de marquage des unités lexicales, i.e. l'attribution des classes grammaticales aux unités lexicales. Dans notre travail de recherche, cette étape est primordiale, car sans étiquetage des unités lexicales nous ne pouvons pas extraire ni les expressions d'opinion ni les caractéristiques. Nous avons utilisé *Stanford POS Tagger* pour localiser toutes les unités lexicales dans le texte. La liste des unités marquées sera utilisée dans l'étape de détermination des dépendances grammaticales. La figure suivante montre un simple texte arabe étiqueté.

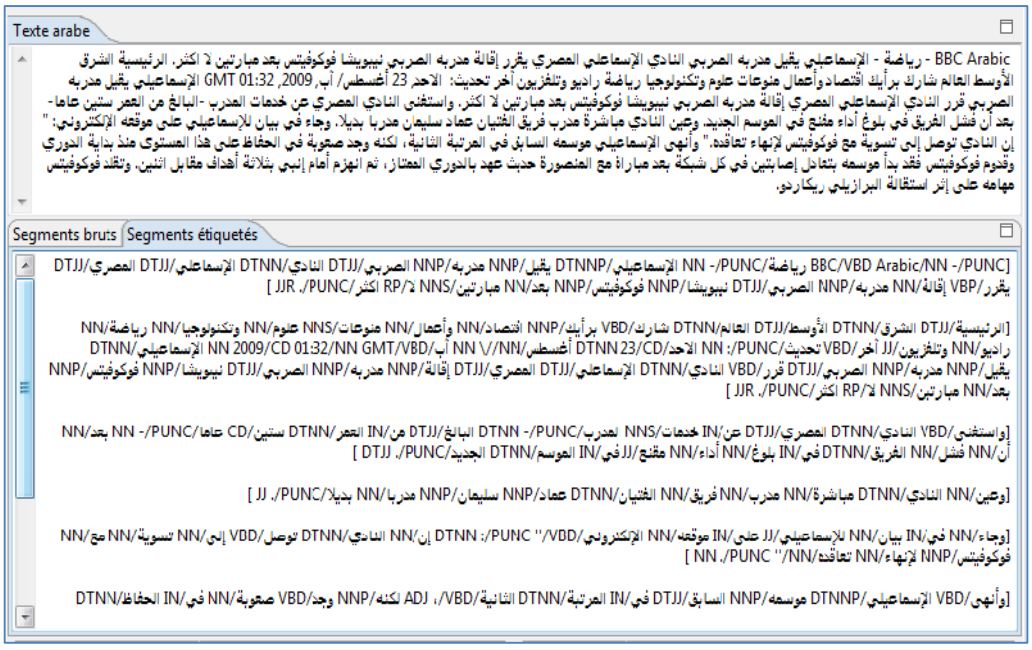


Figure 5.3 Exemple d'un texte arabe étiqueté

5.5. Architecture

Un texte évaluatif est composé de un ou plusieurs segments, chaque segment peut être subjectif ou objectif, chaque segment subjectif contient au moins une expression d'opinion (EO), notons qu'une EO est une expression explicite exprimée sur un objet ou sur l'un de ses caractéristique, elle est composée d'un nom, un adjectif, ou un verbe avec certains modificateurs comme les négations et les adverbes.

Exemple :

منـظر **جميل جدا** (**Très belle** vue)

Ce simple exemple contient une expression d'opinion composée d'un adjectif جميل (belle) et un modificateur جدا (très). Notons que les modificateurs jouent un rôle important dans la détermination de l'intensité des opinions, certains modificateurs comme جدا (très)،كثيرا (beaucoup)،بقوة (puissamment) indiquent une intensité forte, tandis que des modificateurs comme قليلا، (peu)، ببطء (doucement) peuvent indiquer une intensité faible.

Notre approche qui s'intéresse à l'identification des expressions d'opinion et l'extraction des cibles de passage d'opinions par une exploration ontologique du domaine étudié, a besoin de trois composants de base :

— Un lexique de sentiments : comme nous l'avons vu précédemment, il s'agit d'une ressource lexicale L des expressions d'opinion propres au domaine étudié ;

— Une ontologie de domaine O regroupant les concepts du domaine étudié, où chaque concept et chaque propriété est associée à un ensemble d'étiquettes linguistiques ;

— Un ensemble de textes évaluatifs T collectés d'Internet.

Nous détaillons dans ce qui suit les différents modules de notre approche.

5.5.1. Identification des expressions d'opinion

Nous rappelons que l'EO est la plus petite unité au sein d'un segment subjectif (SS). Elle est composée d'un seul et seul mot d'opinion (un nom, un adjectif ou un verbe), éventuellement associé à certains modificateurs comme les mots de négation et les adverbes. Par exemple : حقا ليس جيدا (vraiment pas bon) est une EO composée d'un modificateur حقا (vraiment), d'une négation ليس (pas), et d'un adjectif جيدا (bon). Une EO peut aussi être simplement un adverbe comme dans متحمسا (enthousiaste). Enfin, nous croyons également que l'extraction des expressions de recommandation, telles que : اشتر هذا الكتاب، سوف لن تندم (achète ce livre, tu ne

vas pas regretter), qui sont très fréquentes dans les commentaires, peut améliorer la performance de processus d'identification des expressions d'opinion.

Nous croyons qu'il est difficile, voire impossible de créer un lexique regroupant tous les mots de sentiments propres au domaine étudié, et le problème qui se pose dans ce cas est comment identifier les autres mots de sentiments dans les textes évaluatifs sans leur présence dans le dictionnaire ?

Une réponse à cette problématique a été retrouvée dans [133], les auteurs ont proposé une liste de règles linguistiques pour impliquer les opinions. Si nous combinons ses règles avec notre lexique de sentiments, nous pourrions identifier de nouveaux mots de sentiments avec leurs polarités. Les règles proposées sont les suivantes, que nous avons les illustré par des exemples en langue arabe :

1) Règle de conjonction intra-phrase

Par exemple, nous avons la phrase suivante :

عمر البطارية طويل جدا (La durée de vie de la batterie est très *longue*).

Il n'est pas clair que le mot طويل (longue) exprime une opinion positive ou négative. L'algorithme proposé dans ce contexte essaye de déterminer si طويل (*longue*) exprime une opinion positive ou négative à partir des autres textes évaluatifs. Par exemple dans d'autres critiques, on trouve :

ألة التصوير هذه تلتقط صور جميلة و عمر البطارية طويل (Cet appareil-photo prend de belles photos et la batterie à une durée de vie longue).

À partir de cette phrase, nous pouvons découvrir que طويل (longue) est positif pour عمر البطارية (la durée de vie de la batterie) parce qu'il conjointe avec le mot positif جميلة (belle).
Cette règle est appelée *règle de conjonction*, qui signifie qu'une phrase exprime seulement l'orientation d'une opinion sans la présence des mots comme لكن (mais) qui changent la direction.

La phrase suivante est peu probable :

ألة التصوير هذه تلتقط صور جميلة **و** عمر البطارية قصير (Cet appareil-photo prend de belles photos et la durée de vie de la batterie est *courte*).

2) Pseudo règle de conjonction d'intra-phrase

Parfois, on ne peut pas employer une conjonction explicite و (et). Employons la phrase suivante :

عمر البطارية طويل (La durée de vie de la batterie est longue).

Nous n'avons aucune idée si طويل (longue) est positif ou négatif pour عمر البطارية (la durée de vie de la batterie). Une stratégie semblable peut être appliquée. Par exemple, dans une autre critique on peut avoir :

عمر البطارية طويل، انه لشيء رائع (La durée de vie de la batterie est longue, c'est magnifique).

La phrase indique que l'orientation sémantique de طويل (longue) pour عمر البطارية (durée de vie de la batterie) est positive due à رائع (magnifique), sans explicitement utiliser «و » (et).

3) Règle de conjonction d'inter-phrases

La règle de conjonction peut également être généralisée sur des phrases voisines. L'idée est qu'il est possible d'exprimer une opinion dans un ensemble de phrases consécutives. Des changements d'opinions sont indiqués par des mots comme لكن(mais), الا أن(cependant)…etc.

Par exemple, les passages suivants sont normaux :

نوعية الصورة جيدة. عمر البطارية طويل (La qualité d'image est bonne. La durée de vie de la batterie est longue) et نوعية الصورة جيدة الا أن عمر البطارية قصير (la qualité d'image est bonne. Cependant, la durée de vie de la batterie est courte).

Cependant, le passage suivant n'est pas normal :
نوعية الصورة جيدة. عمر البطارية قصير (La qualité d'image est bonne. La durée de vie de la batterie est courte).

Bien que nous ne sachons pas si طويل (long) (ou قصير (court)) est positif ou négatif pour la عمر البطارية (durée de vie de la batterie), si nous savons que جيد (bonne) est positif, puis nous pouvons impliquer que طويل (longue) est positif et قصير (court) est négatif pour عمر البطارية (durée de vie de la batterie).

4) Règle des synonymes et antonymes

Si un mot s'avère positif (ou négatif) dans un contexte pour une caractéristique, ses synonymes sont également considérés positifs (ou négatifs), et ses antonymes sont considérés négatifs (ou positifs).

Par exemple, dans l'exemple de la règle précédente, nous savons que طويل (*longue*) est positif pour عمر البطارية (durée de vie de la batterie). Alors, nous savons également que قصير (courte) est négatif pour عمر البطارية (durée de vie de la batterie).

Dans notre approche, et pour extraire les expressions d'opinions explicites, nous avons proposé de faire une projection simple du dictionnaire de sentiments sur les textes évaluatifs, les expressions implicites avec leurs polarités sont extraites en employant les règles linguistiques précédentes. La figure suivante illustre le processus d'extraction des expressions d'opinion à partir des textes.

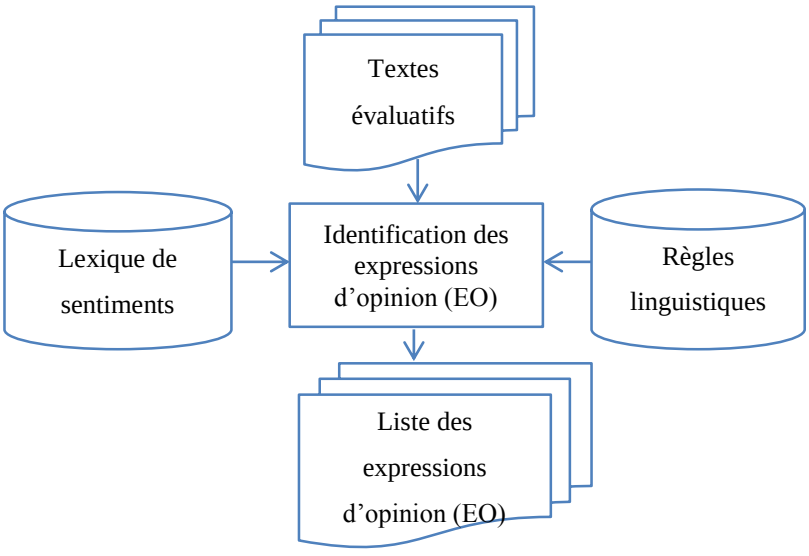


Figure 5.4 Processus d'identification des expressions d'opinion

Comme il y a deux types d'expressions d'opinion : explicites et implicites, nous avons proposé d'enrichir automatiquement le dictionnaire de sentiments à chaque nouvelle expression détectée en testant son existence dans le dictionnaire, si elle n'existe pas dans le dictionnaire, le système va créer une nouvelle entrée contenant le nom et la polarité de cette expression.

La Figure 5.5 illustre le nouveau processus d'identification des expressions d'opinion avec enrichissement de lexique :

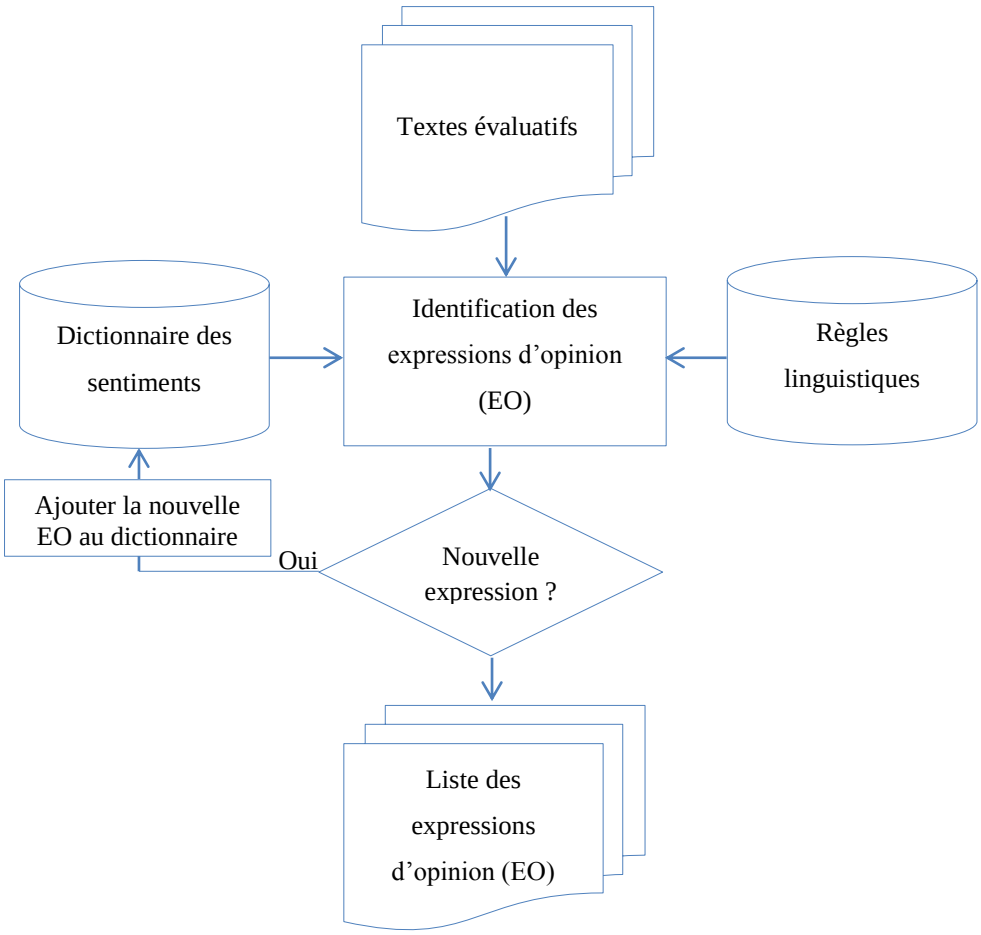


Figure 5.5 Processus d'identification des expressions d'opinion avec enrichissement de lexique des sentiments

5.5.2. Identification des cibles de passage d'opinion

L'identification de la cible d'une d'opinion fait l'objet d'une attention récente en fouille d'opinion, les méthodes existantes permettent principalement de traiter les cas où la cible se situe dans la même phrase que l'opinion [130].

La cible d'une opinion peut être explicite comme dans l'énoncé منظر جميل (belle vue) où la cible منظر (vue) est explicitement couplée avec l'expression d'opinion جميل (belle).

Comme, elle peut être implicite comme dans l'énoncé جميل حقا (vraiment beau) où il n'existe aucun signe indiquant la présence de la cible.

L'identification de la cible d'une opinion consiste à relier l'expression d'opinion avec l'objet évalué. Dans l'énoncé (1), il y a ainsi trois expressions d'opinion جميل (belle), رائع (merveilleuse), أعجبني كثيرا (aimer beaucoup) qui évaluent un objet cible unique منظر (vue).

«أعجبني كثيرا و رائع، جميل **منظر** », «Une **vue** belle et merveilleuse, j'ai beaucoup aimé» (1)

Dans [130], les auteurs ont cité les différents facteurs qui rendent complexe la tâche d'identification de la cible. Ci-après, nous présentons ces facteurs, en les illustrant par des énoncés en arabe :

— **Différentes formes textuelles pour un unique objet du monde**

Un même objet peut être représenté dans le texte par différentes expressions nominales ou pronominales. Dans (2) et (3), on parle de l'objet du monde مدينة لندن (ville de Londres) par une variante nominale métonymique[5] مدينة الضباب (la ville de brouillard) et par une anaphore pronominale هي (elle).

Dans un but applicatif, il est nécessaire de regrouper les opinions qui portent sur le même objet et de nommer l'objet évalué le plus précisément possible. On ne pourra ainsi pas considérer le pronom هي (elle) comme la cible de l'évaluation أبهر (impressionner). مدينة الضباب

[5] La **métonymie** consiste à remplacer un terme par un autre terme qui ne désigne pas la même chose mais qui lui est lié par un rapport logique. Le terme métonymie signifie en arabe soit «مجاز مرسل», soit «كناية»

(La ville de brouillard) est une réponse intermédiaire plus acceptable pour identifier l'objet cible réellement évalué لندن (Londres).

«استمتعت بإجازة رائعة في **مدينة الضباب**». (J'ai apprécié un merveilleux séjour dans la **ville de brouillard**) (2)

أبهرتنا بجمالها (elle nous a impressionné par sa beauté). (3)

— **Les relations méronymiques[6] entre objets**

Les objets sont potentiellement liés à d'autres objets par des relations méronymiques. Dès lors, même si évaluer un méronyme A d'un mot B peut être une façon d'évaluer indirectement B, il importe de considérer le méronyme A comme la cible exacte de l'évaluation. Ainsi, dans (4) et (5), on évalue tout d'abord علامة الرياضيات(la note de mathématique) liée à الامتحان (l'examen) et pas ce dernier dans sa globalité.

لم يوفق أحمد في **الامتحان** هذه المرة (Ahmed n'as pas réussi à l'**examen** cette fois-ci) (4)

فعلامة الرياضيات كانت سيئة للغاية (La **note de mathématique** était très mauvaise) (5)

— **Présence de plusieurs objets candidats autour de l'opinion**

Le troisième facteur est la présence de plusieurs objets distincts dans le contexte d'un même passage d'opinion. Dans (6), les deux passages d'opinion portent sur des cibles distinctes qu'il faut pouvoir déterminer parmi les quatre objets présents dans la phrase : تجارب (essais),الدواء (médicament), الخبراء (experts), مرضى (malades). L'objet le plus proche de l'opinion n'est pas nécessairement sa cible.

---

[6] La **méronymie** est une relation sémantique entre mots d'une même langue. Des termes liés par méronymie sont des méronymes.

La méronymie est une relation partitive hiérarchisée : une relation de partie à tout. Un méronyme A d'un mot B est un mot dont le signifié désigne une sous-partie du signifié de B. La relation inverse est l'holonymie (en arabe : اسم الكل).

للمرضى غير فعال تبن أنه **الخبراء** من طرف **الدواء** على **تجارب** عدة وبعد... (…et après plusieurs **essais** sur le **médicament** par les **experts**, il s'est avéré inefficace pour les **patients**) (6)

— **Proximité aléatoire entre l'opinion et sa cible**

Dernier facteur de complexité, la cible évaluée ne se situe pas toujours à proximité du passage d'opinion. Dans (7), le passage d'opinion de فشل (échec) porte sur l'objet النموذج (le modèle). La présence de nombreux autres objets dans le contexte النموذج (le modèle), مادة عازلة (matière isolante), تجارب (essais), المصممون (les designers) rend complexe l'identification de la cible réelle de l'opinion pour une approche automatique.

صُنع هذا **النموذج** من **مادة عازلة** لل**حرارة** ، بعد **تجارب** عديدة أثبت **المصممون** فشله

(Ce **modèle** a été fabriqué par une **matière isolante**, après de nombreux **essais**, il a prouvé son échec). (7)

Nous croyons que la couche sémantique peut résoudre les difficultés liées à l'identification de la cible. En fait, une ontologie peut offrir des relations de :

- Synonymie
- Antonymie
- Composition
- …etc.

Pour résoudre la problématique d'identification de la cible d'une opinion, il convient donc de considérer ces quatre facteurs. Le processus d'identification consiste alors à :

- Identifier les cibles explicites par projection simple de l'ontologie ;
- Identifier les cibles implicites par exploration de la couche sémantique de l'ontologie.

Dans cette section, nous nous intéressons spécifiquement à l'identification de la cible d'une opinion. Cette étape vise à extraire à partir des textes évaluatifs toutes les étiquettes de l'ontologie. Étant donné que chaque concept et ses étiquettes lexicales associées correspondent à des caractéristiques explicites, nous projetons simplement la composante

lexicale de l'ontologie sur le texte afin d'obtenir, pour chaque SS, l'ensemble des caractéristiques F.

Comme notre ontologie ne couvre pas tous les concepts et leurs propriétés dans le domaine étudié, de nombreux termes dans le texte peuvent être manqués.

Pour extraire les caractéristiques implicites, les propriétés de l'ontologie sont utilisées. Nous rappelons que ces propriétés définissent les relations entre les concepts de l'ontologie. Par exemple, la relation sémantique يشاهد (regarder) relie les concepts شخص (personne) et تلفاز (télévision).

1) Identification des cibles (caractéristiques) explicites

Une cible explicite dans le texte évaluatif correspond à un concept présent dans l'ontologie. Par une projection simple de l'ontologie sur le texte, nous pouvons facilement retrouver toutes les cibles recherchées. La figure ci-après présente l'architecture du sous-module d'extraction des cibles explicites :

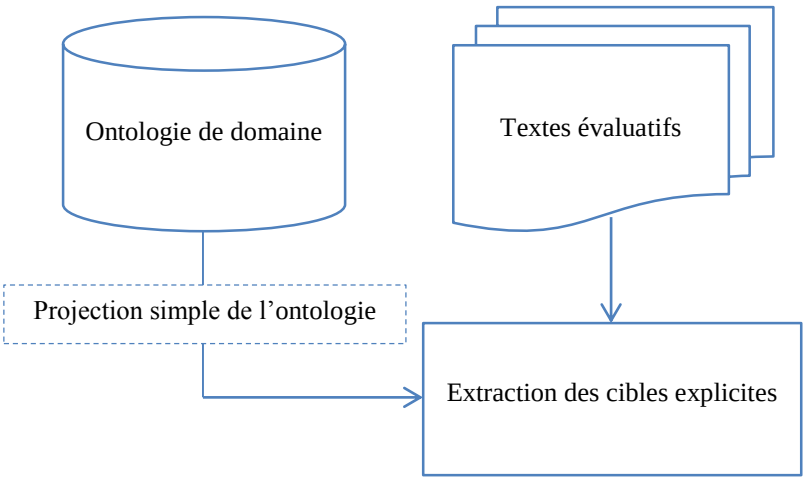


Figure 5.6 Extraction des cibles explicites

2) Identification des cibles implicites

Une cible implicite correspond à une caractéristique qui n'est pas directement visible dans le texte, mais détectable à partir des relations sémantiques de l'ontologie.

La présence des expressions d'opinion dans certains commentaires qui n'ont pas aucune dépendance grammaticale avec aucune caractéristique explicite, peut être utilisée pour repérer les caractéristiques implicites. Dans de tels cas, nous avons proposé d'utiliser les relations sémantiques entre les concepts ontologiques, les attributs et valeurs d'attributs pour localiser les caractéristiques sur lesquelles des opinions ont été exprimées.

Nous notons que le lien entre deux concepts peut être une relation de composition, d'inclusion, relation spatiale ... etc. Par exemple, dans le commentaire مقابلة دةيج (un bon match), le mot d'opinion دةيج (bon) est directement exprimé sur la caractéristique مقابلة (match) parce que دةيج (bon) et مقابلة (match) sont syntaxiquement dépendante, où une couple caractéristique-opinion explicite peut être extraite. Mais, si nous utilisons les propriétés de l'ontologie de domaine, il est possible d'extraire d'autres caractéristiques persistantes liées à la caractéristique مقابلة (match).

Dans la Figure 5.7, si 1_ فورميلا (Formula_1) est une instance du concept حدث_رياضي (Evènement_Sportif), et فيا (FIA) est une instance du concept منظمة_رياضية (Organisation_Sportive), une opinion positive exprimée sur فيا (FIA) peut être extraite. Dans ce cas, فيا (FIA) est extraite comme une caractéristique implicite, parce que les deux concepts منظمة_رياضية (Organisation_Sportive) et حدث_رياضي (Evènement_Sportif) sont sémantiquement liés par la relation ينظم (Organiser), où منظمة_رياضية (Organisation_Sportive) est le domaine et حدث _ رياضي (Evènement_Sportif) est la portée.

Figure 5.7 Exemple d'une relation sémantique entre deux concepts ontologiques

La figure ci-après présente l'architecture du processus d'extraction des cibles implicites.

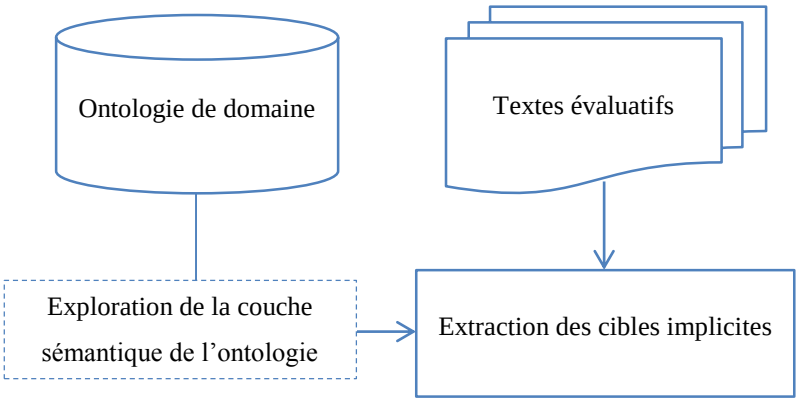


Figure 5.8 Extraction des cibles implicites

Nous regroupons les deux sous-modules précédents pour obtenir l'architecture du module d'extraction des cibles (explicites et implicites). La Figure 5.9 illustre ce module :

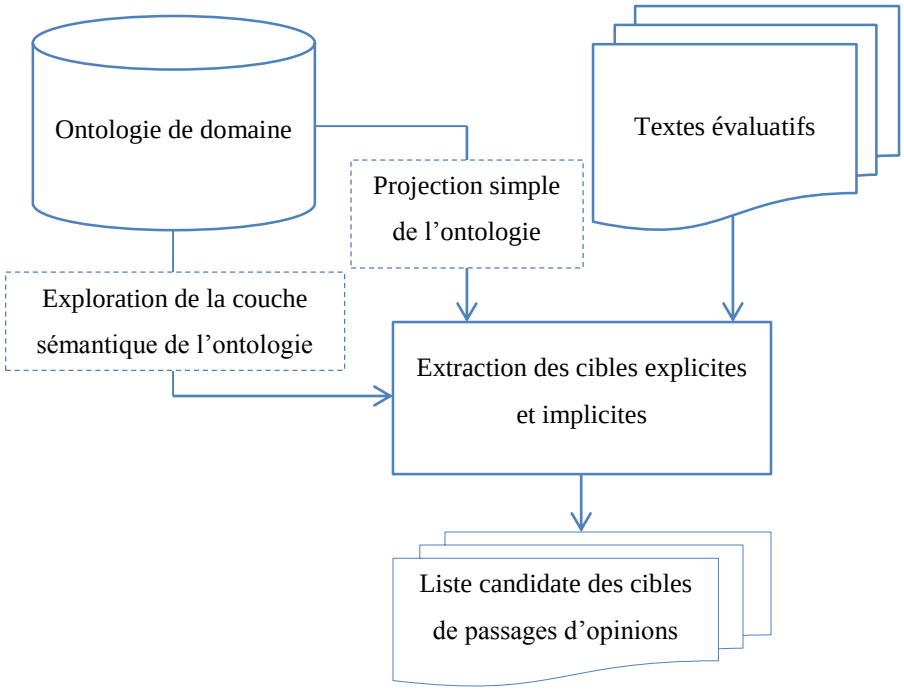


Figure 5.9 Module d'extraction des cibles explicites et implicites

5.5.3. Association des cibles avec les expressions d'opinion

Cette étape est la plus compliquée par rapport aux précédentes, elle cherche à associer toutes les expressions d'opinions extraites dans l'étape 5.5.1, avec les cibles extraites dans l'étape 5.5.2. Autrement dit, nous devons associer à chaque EO l'ensemble des cibles de passage de l'opinion, sous forme de couples (Ci, EOj).

En effet, une relation grammaticale entre deux unités lexicales comme les relations verbales, adjectivales et adverbiales peuvent être des associations entre les expressions d'opinion et les cibles de passage d'opinion (caractéristiques). Dans notre travail, nous avons utilisé *Stanford Parser* (Annexe 2), dont 7 dépendances typées ont été choisies comme modèles de 55 relations binaires proposées par Stanford Parser.

Le tableau suivant montre les relations utilisées.

Dépendance Typée	**Abréviation**
Adjectival Modifier	*amod*
Adjectival Complement	*acomp*
Adverbial Modifier	*advmod*
Open Clausal Complement	*Xcomp*
Nominal Subject	*nsubj*
Negation Modifier	*not*
Nominal subject	*nsubj*

Tableau 5.2 Modèles de dépendances grammaticales

Les sept modèles sélectionnés pour être utilisés dans notre approche sont illustrés ci-dessous par quelques exemples en arabe :

***amod**: adjectival modifier*

Un modificateur adjectival d'un NP (syntagme nominal) est une phrase adjectivale qui sert à modifier le sens du NP.

Exemple :

بمدينة ميلانو الإيطالية
ميلانو الإيطالية → amod(ميلانو, الايطالية)

***acomp**: adjectival complement*

Un complément adjectival du verbe est une phrase adjectivale qui fonctionne comme le complément (objet du verbe).

Exemple :

الذي بدى مرهقا → acomp(مرهقا,بدى)

***advmod**: adverbial modifier*
Un modificateur adverbial d'un mot est un adverbe (ADVP) qui sert à modifier le sens du mot.

Exemple :

لايزال الجمهور متحمسا → advmod(متحمسا, الجمهور)

***Xcomp**: open clausal complement*

Un complément phrastique ouvert (Xcomp) d'un VP ou un ADJP est un complément phrastique sans son propre sujet, dont la référence est déterminée par un objet externe. Ces compléments sont toujours non-finis.

Exemple :

قال أنك تحب السباحة → Xcomp(السباحة,تحب)

***nsubj**: nominal subject*

Un sujet nominal est un syntagme nominal qui est le sujet syntaxique d'une clause. Le gouverneur de cette relation ne soit pas toujours un verbe, où le verbe est un verbe copule[7], la racine de la clause est le complément du verbe copule, qui peut être un adjectif ou un nom.

Exemple :

أحمد هو المحترف → Xcomp(أحمد,المحترف)

Pour chaque caractéristique explicite, les dépendances typées sont utilisées pour extraire la liste des expressions qui lui sont associées i.e. une liste des couples explicites caractéristique-opinion. Par exemple, dans l'exemple suivant, فندق (hôtel) et إطلاله (vue) sont deux caractéristiques explicites.

فندق ممتاز مع إطلاله جميلة (Un excellent hôtel avec une belle vue)

Le graphe de dépendances de la Figure 5.10, montre les différentes relations grammaticales entre les unités lexicales :

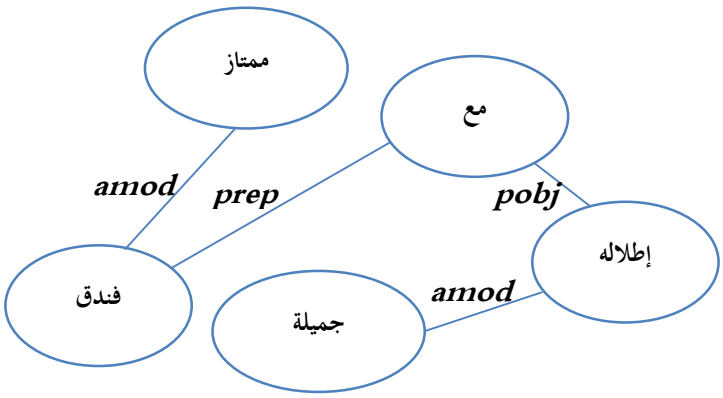


Figure 5.10 Graphe de dépendance grammaticale du segment « فندق ممتاز مع إطلاله جميلة » (Un excellent hôtel avec une belle vue)

La figure suivante montre, les étapes d'extraction des dépendances grammaticales en utilisant Stanford Parser :

[7] En arabe, un verbe copule signifie عماد(دعامة) qui est appelé aussi ضمير الفصل

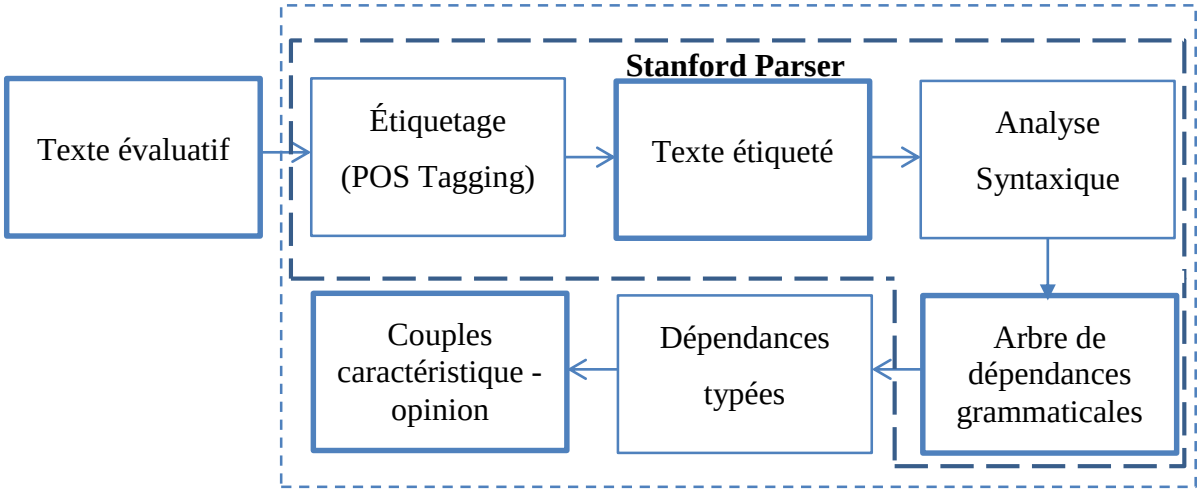


Figure 5.11 Processus d'extraction des dépendances grammaticales

Au cours de cette étape de construction des couples caractéristique-opinion, nous distinguons les cas suivants :

a) Caractéristiques connues et expressions d'opinion connues

Par exemple, si le lexique de sentiments propre au domaine étudié contient les mots مريح (détendu) et خلاب (enchanteur) et l'ontologie contient les termes جو (atmosphère) et طبيعة (nature), le système peut extraire à partir du texte évaluatif suivant :

جو مريح و طبيعة خلابة (une atmosphère détendue et une nature enchanteresse), les couples:

(جو, مريح) (Atmosphère, détendu)

(طبيعة, خلابة) (Nature, enchanteresse).

Cet exemple est assez simple, mais dans de nombreux cas, les cibles et les expressions d'opinion ne sont pas proches les uns des autres qui rend difficile de trouver un lien entre eux.

En fait, comme nous l'avons présenté dans la section 5.5.1 de ce chapitre, l'utilisation des règles de conjonction peuvent améliorer considérablement la performance du processus de recherche des expressions d'opinion inconnues.

Notre système traite les conjonctions (y compris les virgules) comme dans :

جو مريح، منشط **و** هادئ (Atmosphère confortable, activant et détendu).

Si lexique de sentiments contient le mot مريح (confortable) et l'ontologie de domaine contient le concept جو (atmosphère), notre système va extraire les couples suivantes :

(مريح، جو) (atmosphère, confortable)

(منشط، جو) (atmosphère, tonique)

(هادئ، جو) (atmosphère, calme)

Notre système va aussi attribuer automatiquement la polarité du mot مريح (confortable) à tous les mots qui sont conjointement liés. Si مريح (confortable) est un mot positif alors منشط (activant) et هادئ (calme) vont être considérés comme positifs.
L'utilisation des outils comme لكن (mais), إلا أن (cependant) peuvent changer la direction de l'opinion, du positif vers le négatif ou vice-versa. Par exemple, dans le texte évaluatif suivant :

طبيعة خلابة **لكن** الطريق وعر **و** طويل (une nature enchanteresse **mais** la route montueuse **et** longue)

Si le lexique de sentiments contient le mot خلاب (enchanteresse), mais ne contient pas le mot وعر (montueux), et l'ontologie de domaine contient les concepts طبيعة (nature) et طريق (route). Notre système va extraire les couples suivantes :

(خلابة، طبيعة) (Nature, enchanteresse) → Le mot خلاب (enchanteresse) est positif : à partir du dictionnaire des sentiments.

(وعر، الطريق) (Route, montueuse) → Le mot وعر (montueuse) est négatif, et ça dû au modificateur de la direction لكن (mais), qui a changé l'opinion du positif vers le négatif.

(طويل، الطريق) (Route, longue) → Le mot (longue) est négatif parce qu'il est conjointement lié avec le mot وعر (montueuse).

b) Caractéristiques connues et expressions d'opinion inconnues

Comme dans le texte évaluatif suivant :

رحلة مرهقة (voyage fatigant)

Si le mot d'opinion مرهق (fatigant) n'a pas été extrait à l'étape 5.5.1. Dans ce cas, le système va mettre à jour le lexique de sentiments en ajoutant automatiquement le nouveau mot récupéré dans le dictionnaire avec sa polarité.

L'expression d'opinion مرهقة (fatigant) sera associée avec le concept رحلة (voyage) pour produire la couple :

(مرهقة، رحلة) (Voyage, fatigant)

Mais, la question qui se pose comment peut-on connaitre la polarité du mot récupéré مرهِق (fatigant) ?

Nous avons toujours recours aux règles linguistiques mentionnées dans la section (5.5.1) de ce chapitre, par exemple, si nous trouvons dans certains textes évaluatifs, des expressions comme :

رحلة ممتعة **لكن** مرهقة (Voyage amusant **mais** fatigant),

Si le lexique des sentiments contient l'expression d'opinion ممتعة (amusant), avec une polarité positive, le système va automatiquement attribuer la polarité négative à l'expression d'opinion مرهقة (fatigant), et ça dû au modificateur de la direction لكن (mais) qui a changé la direction du positif vers le négatif.

c) Caractéristiques inconnues et expressions d'opinion connues

Comme dans le texte évaluatif suivant : مسلك صعب (route difficile), où le concept مسلك(chemin) n'a pas été extrait à l'étape 2 (section 5.5.2). Dans ce cas, l'ontologie de domaine peut être mise à jour par l'ajout d'une nouvelle étiquette à un concept ou une propriété existante ou par l'ajout d'un nouveau concept ou une nouvelle propriété dans le bon endroit de l'ontologie.
Dans notre exemple, le concept مسلك (chemin) peut être ajouté comme synonyme au concept طريق (route), معبر (passage), ممر (voie, sentier) ou مسار(piste)…etc.

Cependant, comme l'utilisateur peut exprimer une opinion sur différents objets dans un commentaire. Cette étape doit être faite avec soin. Pour éviter les erreurs, nous proposons de mettre à jour manuellement l'ontologie.

d) Expressions d'opinion seules

Comme dans le texte évaluatif suivant :

صعب، متعب و غير مريح (difficile, fatiguant et incommode)

Ce genre d'évaluations exprime une cible implicite. Dans ce cas, nous utilisons les propriétés de l'ontologie afin de récupérer le concept associé dans l'ontologie et nous faisons éventuellement recours aux règles linguistiques comme le cas de notre exemple, où le mot d'opinion صعب (difficile) est une étiquette de la propriété حالة (état) du concept طريق (route) comme indiqué dans la figure ci-après. Nous pouvions déduire en utilisant les propriétés de l'ontologie où l'expression d'opinion صعب (difficile) est associée au concept طريق (route), d'où la couple (صعب، طريق) (route, difficile) est extraite.

En appliquant les règles linguistiques de conjonction, notre système peut extraire deux autres couples :

(متعب، طريق) (Route, fatiguant)

(غير مريح، طريق) (Route, incommode)

Les expressions d'opinion : (difficile), (fatiguant) et (incommode) ont la même polarité parce qu'elles sont conjointement liées.

.

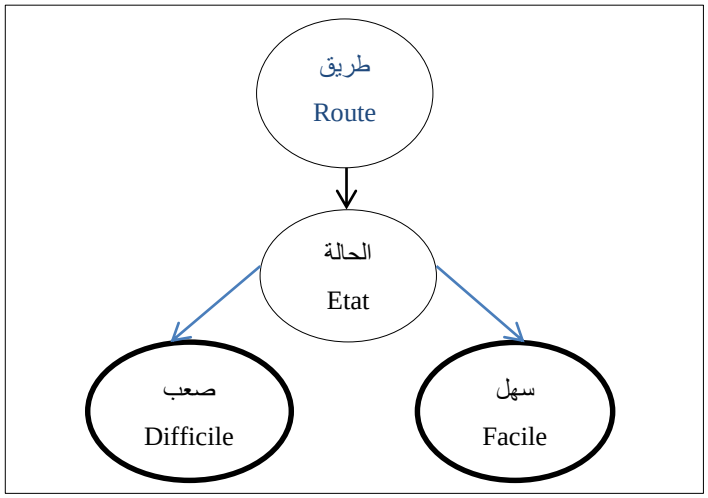


Figure 5.12 Exemple d'une propriété instanciée dans une ontologie de domaine

e) Caractéristiques seules

Comme dans l'évaluation suivante :

طريق شديد الانحدار (route à forte déclivité)

Même si le concept انحدار (déclin) n'est pas associé à aucun mot d'opinion, il est important d'extraire cette information car elle donne une opinion négative vers le concept طريق (route). Une évaluation avec des concepts seuls peut aussi être un indicateur de la présence d'une expression d'opinion implicite vers un concept.

5.5.4. Classification

Après l'extraction de tous les couples (c_i, e_j), nous arrivons à la dernière étape de notre approche qui est la catégorisation des opinions au niveau des caractéristiques extraites.

Dans la littérature, la catégorisation des textes consiste en l'attribution d'une valeur booléenne à chaque paire $(d_j, c_i) \epsilon\ D\ X\ C$, où D est un ensemble de documents et C un ensemble de catégories prédéfinies. Une valeur T attribuée à la paire $(dj\ ,\ ci\ )$ indique une décision de déposer d_j sous c_i, et une valeur F indique une décision de ne pas déposerd_j sous c_i.

Plus formellement, la tâche consiste à approximer une fonction inconnue d'une cible ϕ: D X C $\rightarrow$ $\{T, F\}$ appelée classificateur.

1) Agrégation des opinions au niveau de caractéristiques

Contrairement à la classification d'opinions au niveau des documents qui consiste à attribuer une valeur positive ou négative à un document en attribuant la polarité dominante, notre approche se base sur l'identification des opinions au niveau des caractéristiques, où chaque caractéristique est associée à une ou plusieurs expressions d'opinion sous forme de couples $(c_i, e_j) \in C \times E$, où C, est l'ensemble de cibles de passage d'opinions, et E est l'ensemble des expressions d'opinion.

Pour calculer l'orientation sémantique globale au niveau de chaque caractéristique, nous proposons d'utiliser une fonction d'agrégation d'opinions :
Une caractéristiques *f* peut être associé à une ou plusieurs mots d'opinions ($m_1, \ldots m_n$).

$$OS(f) = \frac{1}{n} \sum_{m_i \in E} OS(m_i)$$

Où : m_i est un mot d'opinion dans la liste *E*, l'ensemble des mots d'opinions associés à *f*. $OS(m_i)$ est l'orientation sémantique du mot m_i.

En appliquant la fonction d'agrégation précédente sur toutes les caractéristiques identifiées (cibles de passage d'opinions), nous obtenons une nouvelle liste de couples (f_i, OS_i), où f_i est une caractéristique de E, et OS_i est l'orientation sémantique agrégée et associée à f_i.

Notons qu'à un mot positif est assigné une orientation sémantique selon son degré de positivité ou de négativité qui varie de -1 à +1.

Si le score final est positif, alors l'opinion sur *f* est positive. Si le score final est négatif, alors l'opinion sur f est négative. Elle est neutre autrement.

2) Évaluation des méthodes de classification

Comme nous l'avons mentionné dans le chapitre 3 de ce présent manuscrit, pour mesurer l'efficacité d'un classificateur dans un problème à *n* classes, trois mesures principales sont utilisées : la précision, le rappel et le F-score.

$$Précision = \frac{\sum_{i=1}^{n} précision_i}{n}$$

$$Rappel = \frac{\sum_{i=1}^{n} rappel_i}{n}$$

Etant donné pour chaque classe i :

$$précision_i = \frac{nombre\ d'objets\ correctement\ attribués\ \text{à}\ la\ classe\ i}{nombre\ d'objets\ attribués\ \text{à}\ la\ classe\ i}$$

$$rappel_i = \frac{nombre\ d'objets\ correctement\ attribués\ \text{à}\ la\ classe\ i}{nombre\ d'objets\ appartenant\ \text{à}\ la\ classe\ i}$$

La précision indique le degré de vérité des résultats obtenus, tandis que le rappel indique leur pertinence. Il est intéressant de faire un compromis entre la précision et le rappel, c'est le rôle du $F_\beta score$ dont le coefficient β est inversement proportionnel à l'importance qui est donnée à la précision par rapport au rappel.

$$F_\beta score = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

Généralement, le $F_1 score$ est utilisé, qui est le compromis équilibré entre la précision et le rappel.

$$F_\beta score = 2\ \frac{Précision \times Rappel}{\text{Précision} + \text{Rappel}}$$

Habituellement, l'utilisation de ces mesures entre dans un processus de validation. Il y a deux types de validation : la validation par test et la validation croisée. Dans une validation par test, on apprend un modèle sur un corpus d'apprentissage et on test la classification sur un autre corpus. Dans une validation croisée, on ne considère pas précisément d'ensemble d'apprentissage et de test mais des sous-ensembles d'un corpus qui sont utilisés pour se valider les uns les autres. La validation croisée est considérée plus fiable qu'une simple validation par test.

L'évaluation des méthodes ne repose pas seulement sur les mesures d'évaluation, elle repose aussi sur la définition des ensembles d'apprentissage et de test. Dans notre problème, si le corpus d'apprentissage est trop proche du corpus de test, l'obtention de bons résultats n'assure pas que le modèle appris soit un bon modèle pour la fouille d'opinions.

3) Classificateur

Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinions. On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, les règles de décision ainsi que des méthodes combinant différents classificateurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les machines à vecteurs de supports (SVM) [40, 44, 108, 109, 110, 111, 112] et les classificateurs bayésiens naïfs (NB) [98, 109, 110]. Les machines à vecteurs de supports, appelées encore séparateurs à vaste marge (SVM), sont des classificateurs multi-classes.

Les Séparateurs à Vaste Marge, souvent écourtés à l'acronyme SVM, sur lesquels repose notre tâche de classification, sont caractérisés théoriquement par l'avantage de minimisation de l'erreur empirique et pratiquement par des algorithmes optimisés, avec possibilité de construire un noyau adapté aux données à traiter.

SVM est une méthode de classification par apprentissage supervisé, elle repose sur l'existence d'un classificateur linéaire dans un espace approprié. Cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur

l'utilisation des fonctions dites « noyaux » (*ang. kernel functions*) qui permettent une séparation optimale des données.

La notion d'apprentissage étant importante, il se divise en apprentissage supervisé et non supervisé. Le cas qui concerne les SVM est l'apprentissage supervisé. Les exemples particuliers sont représentés par un ensemble de couples d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vus et qui prédit les sorties pour les entrées qui n'ont pas encore été vues. Les entrées peuvent être des descriptions d'objets et les sorties sont la classe des objets donnés en entrée [47].

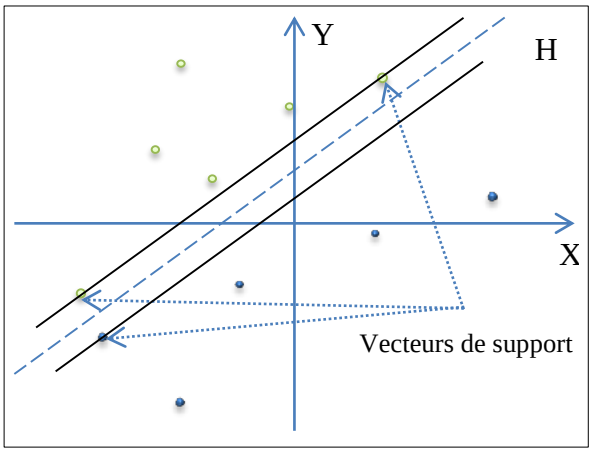


Figure 5.13 Représentation graphique des SVM

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan.

Dans la figure 5.13, on détermine un hyperplan qui sépare les deux ensembles de points. Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

L'une des propriétés remarquables des SVM est que l'hyperplan doit être optimal, c'est-à-dire parmi les hyperplans valides, il faut chercher celui qui passe au « milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr ». Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale.

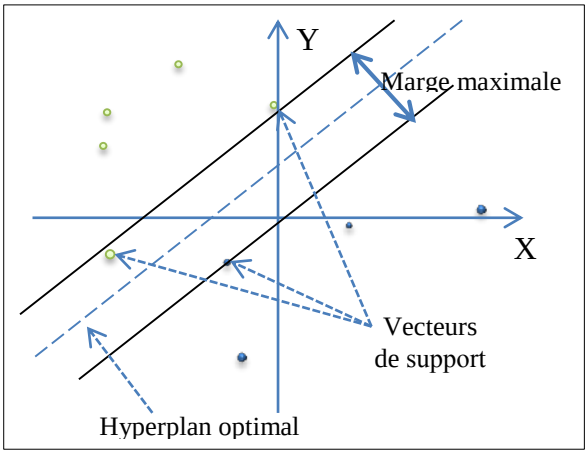


Figure 5.14 Marge maximale

On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de *séparateurs à vaste marge.* Graphiquement, nous pouvions schématiser le processus de classification dans la figure suivante :

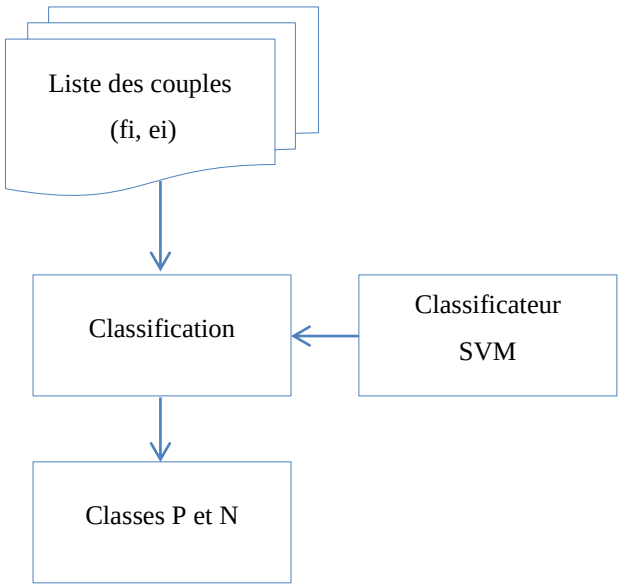


Figure 5.15 Processus de classification

En regroupant les modules présentés précédemment, nous obtenons l'architecture générale de notre approche :

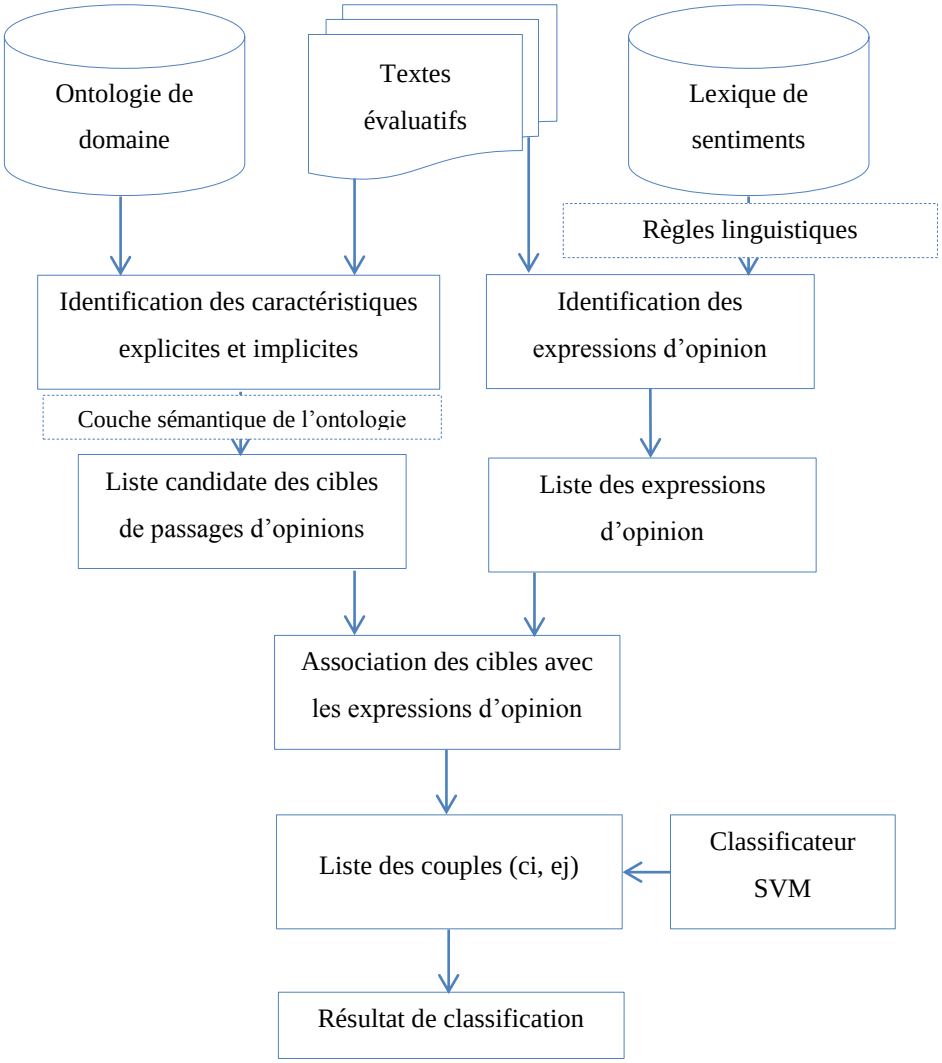


Figure 5.16 Architecture générale de notre approche

— Module d'identification des expressions d'opinion : utilise comme entrées le dictionnaire de sentiments couplé avec les règles linguistiques (section 5.2.2) ;

— Module d'identification des cibles de passages d'opinions : utilise l'ontologie de domaine comme entrée principale, avec des algorithmes de parcours de l'ontologie ;

— Modules d'association des expressions d'opinions avec les cibles candidates ; utilise les résultats des deux modules précédents, et fournit un résultat sous forme d'un ensemble de couples (c_i, EO_j);

— Classification d'opinions au niveau des caractéristiques (cibles) : utilise la liste des couples extraite dans le module précédent.

5.6. Conclusion

Ce chapitre présente une description générale de notre approche ONTOMART, qui se base sur l'exploration ontologique du domaine étudié. Elle se compose de trois modules principaux : extraction les cibles de passage d'opinions explicites et implicites à partir des textes évaluatifs, extraction des expressions d'opinion en se basant sur le lexique de sentiments, puis l'association des caractéristiques identifiées avec les expressions d'opinions correspondantes formant une liste de couples caractéristique-opinion, qui seront ensuite classifiée en faisant appel aux techniques les plus répondues dans le domaine de fouille d'opinions.

L'ontologie de domaine comme un modèle de représentation des connaissances du domaine, et le lexique de sentiments sont les entrées de notre système. Ce choix est motivé par le rôle important que peut jouer la séparation des lexiques et ontologies par domaine, pour diminuer l'ambiguïté subjective.

Dans le chapitre suivant, nous présentons l'implémentation et l'évaluation de notre approche, en testant sa performance, soutenue par une étude de cas.

Chapitre 6

Implémentation et Evaluation

6.1. Introduction

Dans ce chapitre, nous présentons l'implémentation de notre approche ONTOMART présentée dans le chapitre 5, puis l'évaluation et la discussion des résultats obtenus.

Comme nous l'avons mentionné dans le chapitre 5 de ce manuscrit, la mise en œuvre de notre système nécessite comme entrées : une ontologie conceptualisant le domaine étudié, un lexique reflétant les sentiments des personnes, qui mène à extraire leurs opinions envers les objets commentés et critiqués, et un corpus de textes pour évaluer notre approche.

La création de l'ontologie du domaine à étudier constitue elle-même une tâche difficile et cela dû à la complexité liée à la variabilité sémantique des concepts qui dépendent fortement du contexte, et elle nécessite elle-même comme entrées, un corpus de textes et un thésaurus regroupant la terminologie du domaine.

Le lexique des sentiments relatif au domaine étudié doit contenir les mots des sentiments les plus couramment utilisés, avec une possibilité d'enrichissement en utilisant des règles linguistiques comme la conjonction, la disjonction et les modificateurs de polarité comme la négation. En ce qui concerne le corpus de textes arabes, nous avons utilisé un corpus existant collecté auprès d'un ensemble de sites de la presse arabe

6.2. Choix du domaine étudié

Le choix du sport comme domaine de notre étude parmi plusieurs a été motivé par :

- La richesse des textes arabes disponibles sur le net en matière de terminologie qui regroupe les concepts du domaine, notamment les blogs sportifs où les blogueurs peuvent librement mettre en ligne des articles relatifs à toutes les disciplines sportives.

- Les intérêts qui peuvent être apportés par la conceptualisation de ce domaine pour les sportifs, et les établissements sportifs.

- D'après notre connaissance, le sport n'a pas subi une importance remarquable par les chercheurs notamment ceux qui travaillent dans le domaine de la fouille d'opinions.

6.3. Construction de l'ontologie

Toutes les méthodes de construction des ontologies à partir des textes utilisent des outils de Traitement Automatique des Langages Naturelles (TALN), et s'intéressent à la possibilité de combiner des corpus de textes d'un domaine avec un lexique de concepts pour extraire des termes, ainsi que les relations entre eux et construire, in fine, une ontologie du domaine à partir des termes identifiés [3,50,55,60,76].

En fait, il existe deux approches pour déceler les relations entre concepts :

- La première est basée sur la définition de patrons lexico-syntaxiques qui établissent une relation entre concepts du domaine. Ces patrons peuvent être relatifs aux relations hiérarchiques (hyperonymie, définition, méronymie), synonymie, etc.

- La deuxième approche, dite statistique, décèle des relations entre concepts (co-occurrences de termes, etc.) sans toutefois interpréter ces relations.

Dans cette section, nous nous intéressons à la création de notre ontologie propre au domaine étudié qui est le sport à partir des textes arabes, dont l'intérêt de cette ontologie consiste à conceptualiser le domaine étudié, ainsi son utilisation dans la phase d'extraction des mots d'opinion, expliquée dans la section 2.2 du chapitre 5.

Les concepts ontologiques extraits correspondent aux termes figurant dans les textes. Une fois les concepts ontologiques identifiés, l'objectif est de repérer les relations sémantiques liant les

termes désignant les concepts ontologiques dans les textes. En effet, ces relations traduisent des actions permettant de définir des relations sémantiques entre les concepts ontologiques.

Dans le domaine du sport, les relations sémantiques expriment des liens entre des concepts comme : منافسة (compétition), حدث (événement), مباراة (tournoi), لاعب (joueur),…etc. Parmi les liens sémantiques qui relient ces concepts, nous trouvons les relations verbales qui apparaissent fréquemment dans les textes, par exemple dans l'expression suivante :

فاز العداء بالمدالية الذهبية في سباق ال 1500 م (l'athlète a remporté la médaille d'or aux 1500 m)

La relation verbale فاز (remporter) relie les deux concepts العداء (athlète) et السباق (course).

Le processus de construction de notre ontologie doit assurer un coût réduit (temps, effort, niveau d'expertise…etc.), la qualité, la réutilisabilité et l'extensibilité.

6.3.1. Analyse terminolo-ontologique

Notre approche de construction de l'ontologie est semi-automatique car elle nécessite à chaque étape des interventions humaines, elle est basée sur une analyse terminolo-ontologique des textes et sur des patrons lexico-syntaxiques, qui consiste à analyser les termes et les relations extraits par les outils de traitement automatique des langages naturels, et construire à la fois les concepts d'une ontologie formelle en distinguant les classes, les propriétés et les valeurs des propriétés.

Nous utilisons un corpus de textes comme source de connaissances pour l'extraction des concepts et leurs propriétés. Dans le domaine du sport, les textes sont disponibles en format électronique. Nous pouvons donc utiliser des outils[1] d'analyse terminologique afin de les exploiter et de repérer les concepts et leurs propriétés automatiquement.

6.3.2. Méthodologie de construction

Notre méthodologie construit l'ontologie du domaine à partir des termes extraits des textes. Elle s'appuie sur l'utilisation d'un thésaurus regroupant des dizaines de mots et d'expressions,

[1] GATE : https://gate.ac.uk

un ensembles de patrons lexico-syntaxiques, d'un corpus existant de textes arabes, et l'utilisation des outils TALN qui permettent de passer du corpus de textes à l'ensemble des concepts du domaine et déceler les relations entre eux.

Notre méthode se compose des modules suivants :

– *Extraction de termes* : consiste à extraire les termes et leurs propriétés en prenant en entrée le corpus de textes pour détecter les objets. Puis, nous analysons le corpus avec un analyseur syntaxique pour extraire les paires et les triplets présents dans les mêmes syntagmes syntaxiques.

– *Construction du noyau de l'ontologie* : consiste à utiliser les paires (objet, propriété) pour la construction d'une hiérarchie de concepts.

– *Extraction des relations entre concepts* : prend en entrée les triplets extraits du texte, puis extrairait les relations transversales[2].

– Regroupement des deux deniers modules pour obtenir l'ontologie complète.

1) Construction du noyau

Le noyau d'une ontologie est l'ensemble de concepts généraux et leurs propriétés. Pour construire le noyau, nous projetons le thésaurus de domaine sur l'ensemble des textes.

Les concepts sont dénotés dans les textes par des termes simples ou composés, l'extraction des termes doit précéder l'étape d'extraction des concepts. A ce stade, nous effectuons une correspondance entre les termes extraits et les concepts associés à ces termes. Pour ce faire, nous nous basons sur une ressource externe qui est le thésaurus de domaine.

Dans ce qui suit, nous détaillons notre démarche pour l'extraction des concepts :

- Extraction des concepts à partir des textes par projection du thésaurus ;

[2] Des relations non taxonomiques entre concepts : relations verbales, adjectivales, adverbiales, etc.

- Identification des relations sémantiques entre concepts par utilisation des patrons lexico-syntaxiques ;

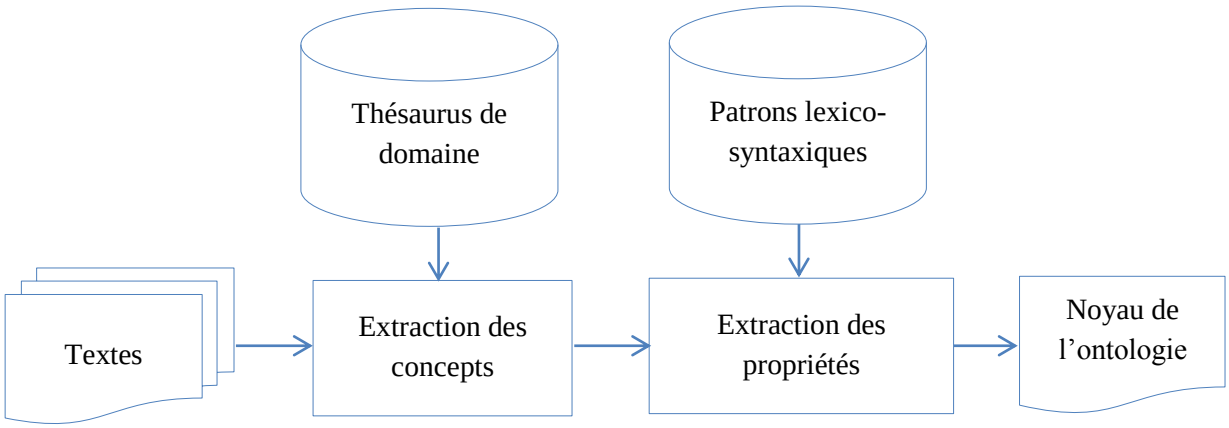


Figure 6.1 Module de construction du noyau

Le seul patron que nous utilisions pour identifier les propriétés des concepts est de type (objet, préposition, objet), par exemple dans l'expression suivante :

شهدت مسابقة تحدي السيارات **ذات** الدفع الرباعي مشاركة 20 متسابقاً

Si nous utilisons le mot **ذات** (à) comme préposition dans le patron (objet, préposition, objet), nous pouvons simplement extraire le triple suivant :

(سيارة، ذات، دفع رباعي), dont, دفع رباعي est une étiquette de la propriété نوع الدفع (motricité).

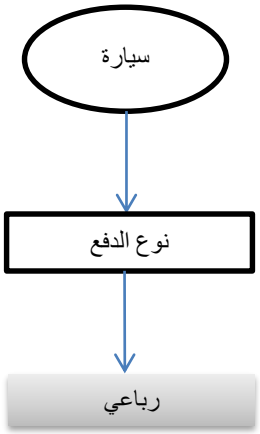


Figure 6.2 Une étiquette de la propriété نوع الدفع (motricité) relative au concept سيارة (voiture)

Il existe plusieurs prépositions qui jouent le rôle de déterminants des propriétés relatives aux concepts du domaine (exemple : ل، ب، ذات).

En effet, en langue arabe, la plupart des propriétés ne sont pas liées à des prépositions comme en français ou en anglais, mais elles sont définies comme génitif (مضاف اليه) comme dans les expressions suivantes :

(1) سباق ال 1500 م (Course 1500 mètres).

(2) سباق ال 100 م حواجز (Course les 100 mètres haies).

(3) بطل العالم في الملاكمة وزن الريشة (Le champion du monde en boxe poids plume).

(4) بطل العالم في الملاكمة وزن الذبابة (Le champion du monde en boxe poids mouche).

Les deux propriétés م 1500 (1500 m) et م 100 (100 m) respectivement dans l'expression (1) et l'expression (2), peuvent être des étiquettes de la propriété مسافة (distance) liée au concept سباق (course).

Les deux propriétés وزن الريشة (poids plume) et وزن الذبابة (poids mouche) respectivement dans l'expression (3) et l'expression (4) peuvent être des étiquettes de la propriété اختصاص (spécialité) liée au concept ملاكمة (boxe).

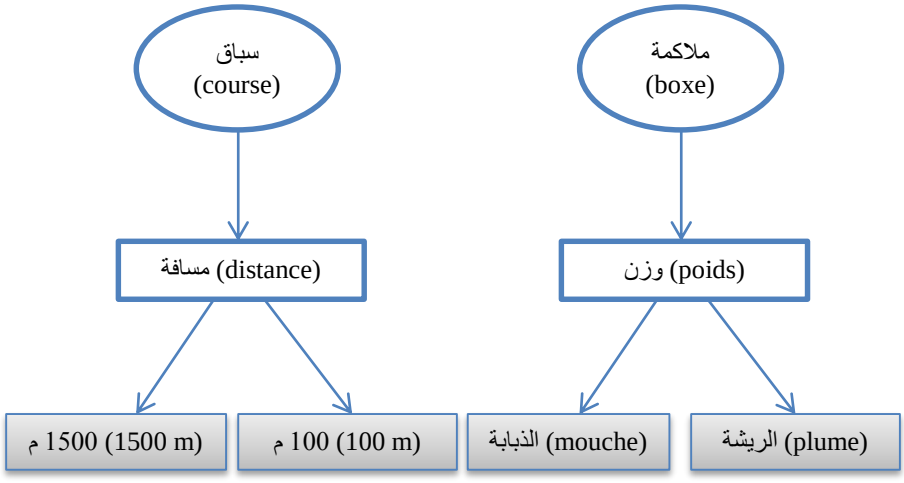


Figure 6.3 Exemple de deux propriétés définies comme génitifs

Dans notre étude, nous avons remarqué que l'intervention humaine est fréquente, notamment dans le cas de nomination de la propriété lié au concept. La propriété peut regrouper une ou plusieurs étiquettes, mais elle reste implicite pour le système, c'est comme dans notre exemple, les deux propriétés مسافة (distance), وزن (poids), qui sont notées manuellement.

2) Extraction des relations entre concepts

Afin d'extraire les relations sémantiques entre concepts, nous utilisons les mêmes ressources utilisées dans la phase d'extraction des concepts, et un ensemble de patrons lexico-syntaxiques pour repérer les relations sémantiques. L'application de tels patrons nécessite de traiter préalablement le texte en appliquant différents outils du TAL (tokenizer, lemmatiseur, analyseur syntaxique, etc.).

Les patrons les plus répondus sont résumés dans le tableau suivant :

Patron	**Exemples**
(Nom, Nom)	جولة الإعادة، مدرب الفريق، صفارة الحكم
(Nom, Adjectif)	رمية جانبية، تسديدة قوية، لاعب محترف
(Nom, Préposition, Nom)	القفز بالزانة، الانسحاب من السباق
(Nom, Préposition, Nom, Nom)	الفوز بكأس العالم، الحاصل على المدالية الذهبية
(Nom, Nom, Nom)	دائرة رمي المطرقة، دائرة رمي القرص
(Nom, Nom, Adjectif)	حاجز سباق حر، متسابق مسافات قصيرة
(Nom, Préposition, Nom, Nom)	الإبقاء على وضعية الانحناء

Tableau 6.1 Quelques patrons utilisés pour l'extraction des relations entre concepts

Dans certains cas, comme dans l'expression suivante :

بطل العالم للمرة الثالثة على التوالي في سباق ال 100 م حواجز ذكور (Le champion du monde pour la troisième fois consécutive dans les 100 mètres haies hommes)

3) Regroupement des concepts et les relations entre eux

Cette étape consiste à fusionner les deux étapes précédentes pour avoir l'ontologie complète.

Nous pouvions résumer les étapes de construction de notre ontologie comme suit :

Entrées

Thésaurus du domaine : Un vocabulaire propre au domaine étudié, regroupant plus de 2500 termes et d'expressions, construit manuellement, dont, plusieurs ressources sur le net ont été utilisées.

Les textes : les textes arabes relatifs au domaine du sport sont extraits d'un corpus existant.

Les patrons lexico-syntaxiques (Tableau 6.1) : Ils sont employés pour repérer les relations sémantiques entre concepts.

Traitement

L'ensemble des traitements automatiques des textes employés pour extraire les concepts et les relations sémantiques entre concepts. L'architecture de notre méthode peut être représentée schématiquement comme suit :

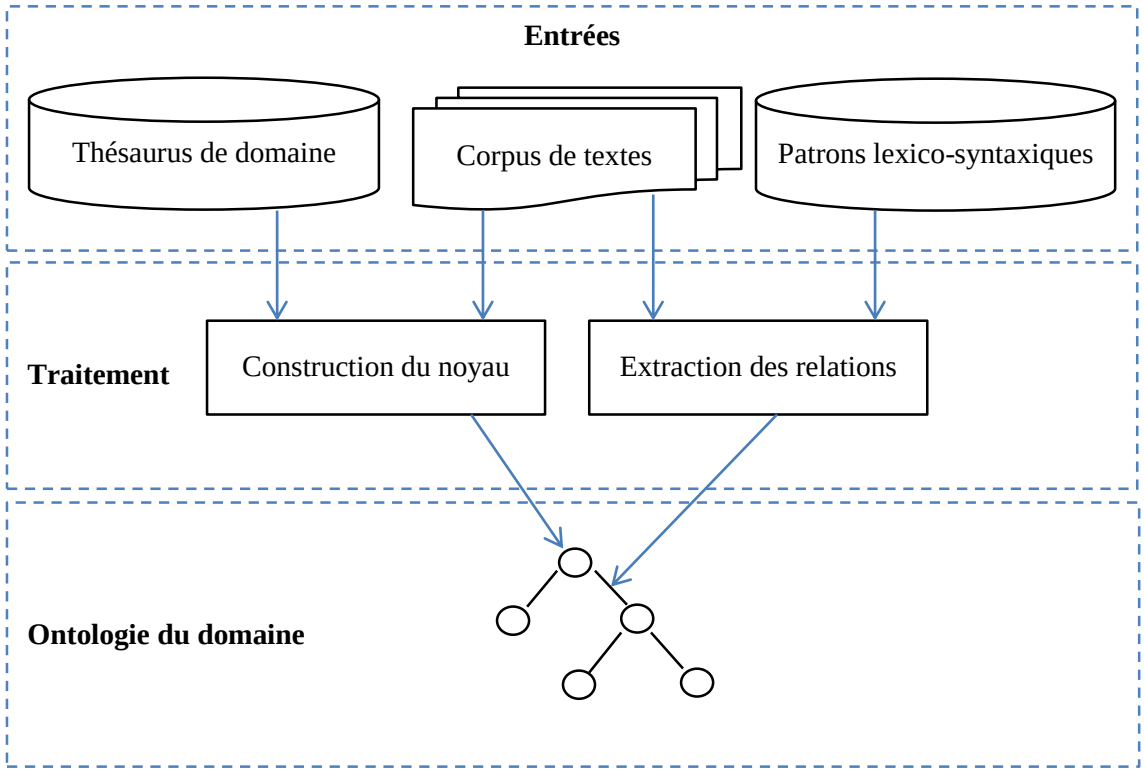


Figure 6.4 Etapes de construction de l'ontologie du domaine

Pour éditer notre ontologie, nous avons utilisé Protégé, elle contient 371 concepts, liés à 6 superclasses : أداة (Outil),رياضة (Sport),شخص (Personne),مجموعة (Groupe),منشأة (Structure), هيئه (Organisation). 42 propriétés d'objets, et 750 étiquettes pour les concepts et les propriétés d'objets. La figure suivante donne un aperçu d'un extrait de notre ontologie du domaine.

Notre ontologie est codée en OWL (Web Ontology Language), pour l'explorer, nous avons utilisé le Framework JENA (Annexe 1) comme un plugin dans un projet Java.

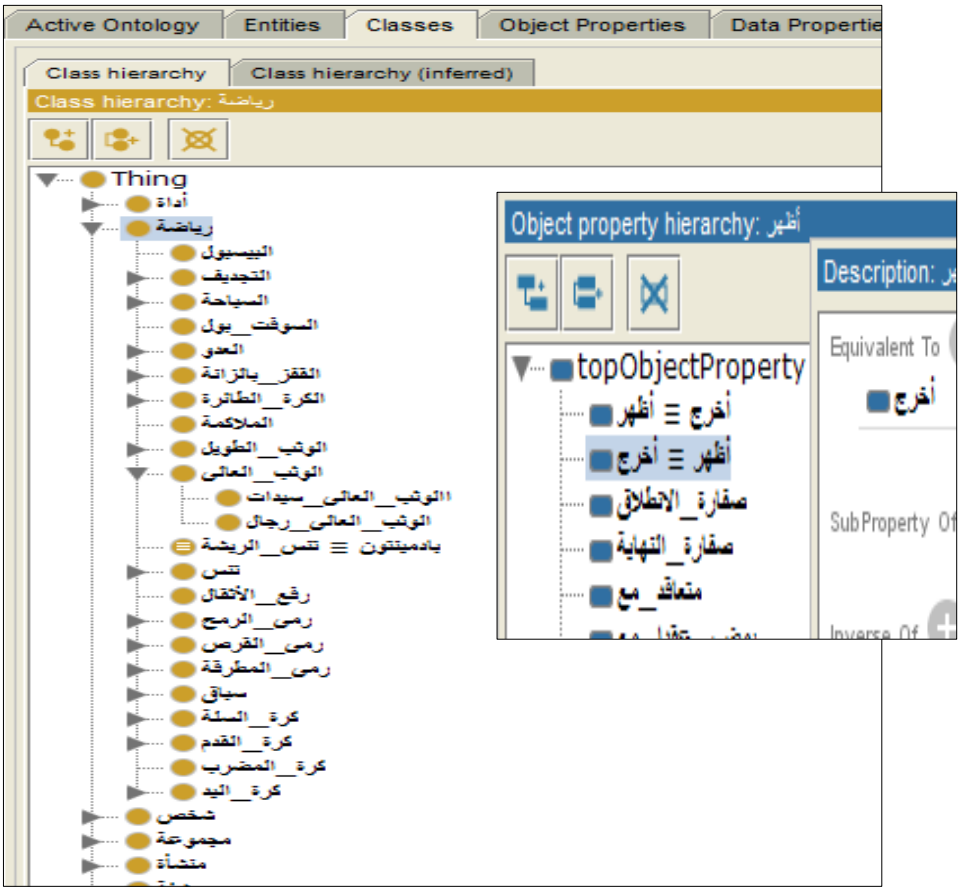


Figure 6.5 Un extrait de notre ontologie du domaine

6.4. Lexique de sentiments

Un lexique d'opinions (ou lexique de sentiments) est une ressource linguistique qui associé à chaque mot sa polarité sémantique et éventuellement une intensité mesurant le degré de positivité et négativité.

Dans le domaine de fouille d'opinions, chaque mot d'opinion est considéré comme unité d'information de base fournissant un indice pour la détection de subjectivité dans le document.
Beaucoup de ressources lexicales sont libres à utiliser : *Bing Liu's Opinion Lexicon*[3], *MPQA* Subjectivity Lexicon[4], *SentiWordNet*[5], *Harvard General Inquirer*[6], *LIWC*[7]…etc. Malheureusement, les lexiques de sentiment pour la langue arabe disponibles sur Internet ne couvrent pas tous les domaines et un déficit significatif a été observé comme dans notre étude de cas, où de nombreuses expressions d'opinion ne sont pas présentes.

D'après notre recherche sur le net, le seul lexique de sentiments pour la langue arabe publiquement disponible est la version arabe de *MPQA Subjectivity Lexicon* (retrouvé sur altec-center.org), qui est une traduction de la version anglaise.

Le lexique est divisé en 4 catégories selon l'intensité des mots : positive fort, positive faible, négative faible, et négative fort, distribué comme suit :

Catégorie	**Nombre d'expressions d'opinion**
Positive Fort	1327
Positive Faible	683
Négative Faible	876
Négative Fort	2663
Total	**5549**

Tableau 6.2 Distribution des mots de sentiments selon leurs intensités

6.5. Corpus

De nombreuses ressources telles que les sites web de commerce électronique, forums, blogs ... etc., relatifs à de nombreux domaines : électronique, cinématographie, politique, sport, etc.,

[3] http://www.cs.uic.edu/~liub/
[4] http://mpqa.cs.pitt.edu/
[5] http://sentiwordnet.isti.cnr.it/
[6] http://www.wjh.harvard.edu/~inquirer/
[7] http://www.liwc.net/

peuvent être utilisées pour créer des corpus textuels. La langue arabe souffre d'un manque remarquable de ressources textuelles subjectives dans de nombreux domaines, par exemple, d'après nos recherches sur Internet, nous n'avons pas trouvé des sites de e-commerce qui permettent aux clients de publier leurs opinions sur les produits et les services en langue arabe. Dans notre travail, nous avons utilisé un corpus existant appelé OSAC (Open Source Arabic Corpora) [144]. OSAC est recueilli auprès de nombreuses sources telles que *bbcarabic.com, cnnarabic.com, aljazeera.net, khaleej.com*...etc. Le corpus couvre de nombreuses catégories : nouvelles du Moyen-Orient, nouvelles du monde, commerce & économie, sport, presse internationale, sciences & technologies, art & culture, etc.

Pour le domaine sportif, le corpus OSAC contient 2419 textes couvrant plusieurs disciplines sportives : football, basketball, athlétisme…etc.

6.6. Expérimentation et évaluation

Selon la structure modulaire de notre approche, nous avons évalué chaque étape indépendamment : extraction des expressions d'opinion, extraction de caractéristiques explicites et implicites, la construction de paires caractéristique-opinion, et enfin la classification des caractéristiques selon la polarité associée en utilisant en utilisant un classificateur SVM.

6.6.1. Expérimentation

- **Extraction des expressions d'opinion**

Cette étape dépend fortement de la richesse de lexique de sentiments. Malheureusement, notre lexique ne couvre pas tous les mots d'opinion tels que dans le segment de texte ضربة مقصيه (coup de ciseaux), où l'adjectif مقصيه n'est pas extrait, cet adjectif peut exprimer une opinion positive envers ضربة (coup) ou une opinion implicite vers un joueur.

Rappel et *Précision* présentés au chapitre 3, sont les deux mesures de performance utilisées pour évaluer cette étape :

$$Rappel = \frac{Nombre\ d'expression\ correctement\ extraites}{Nombre\ d'expressions\ correctes}$$

$$Précision = \frac{Nombre\ d'expression\ correctement\ extraites}{Nombre\ d'expressions\ extraites}$$

Après un test à l'aide de 120 textes de notre corpus, nous avons extrait manuellement 352 expressions d'opinion que nous avons estimées comme pertinente contre 237 expressions d'opinion extraites automatiquement par notre système. De 237 expressions, nous avons constaté que 25 expressions ne portent pas aucune subjectivité comme dans le segment de texte مباراة ودية (match amical) où l'adjectif ودية (amical) n'a pas été utilisé pour exprimer une opinion.

La figure suivante montre un texte extrait de notre corpus étiqueté manuellement pour extraire les expressions d'opinion.

المنتخب العراقي يفوز على نظيره الفلسطيني: احتفل المنتخب العراقي بالفوز على المنتخب الفلسطيني في اول مباراة ودية له على ارضه بمدينة اربيل عاصمة اقليم كردستان العراق. وفاز الفريق العراقي على ضيفه الفلسطيني بثلاثة اهداف مقابل لا شيء، بفوز قال عنه مراسل بي بي سي في بغداد غابريل جيتهاوس انه خبر مفرح ونادر في العراق. وكانت تلك اول مباراة دولية للمنتخب العراقي في ارضه بسبب الحظر الذي فرضه الاتحاد الدولي لكرة القدم نتيجة تدهور الاوضاع الامنية التي اعقبت الغزو والاحتلال الامريكي عام 2003. وكانت آخر مباراة دولية للعراق ضد سورية في يوليو/تموز عام 2002 وفاز حينها العراق بهدفين لهدف. وكان المنتخب العراقي بطل آسيا عام 2007 يخوض مبارياته الدولية في السنوات الماضية في دول عربية مجاورة منها الأردن والإمارات. وقد خرج المنتخب العراقي من الدور الأول لكأس القارات التي اقيمت في جنوب افريقيا الشهر الماضي بعد تعادلين مع البلد المضيف ونيوزيلندا وهزيمة من اسبانيا بهدف واحد. واثر البطولة انتهى علاقة المدرب الصربي بورا ميلوتينوفيتش مع المنتخب، وخلفه في الموقع المدرب العراقي ناظم شاكر. كما ان العراق خرج من المرحلة الأولى لتصفيات آسيا المؤهلة لكأس العالم 2010 في جنوب افريقيا. ويبدو أن العراقيين يأملون في استكمال التجهيزات لاستضافة المباريات الدولية اعتبارا من عام 2011 حيث تبدأ الاستعدادات لتصفيات كأس العالم 2014.

Figure 6.6 Etiquetage manuelle pour l'extraction des expressions d'opinion

Nous avons observé aussi que certaines expressions objectives peuvent porter des opinions, comme dans le segment de texte بطل آسيا (champion d'Asie), où le mot بطل (champion) peut être utilisé pour exprimer une opinion positive sur la caractéristique فريق (équipe).

En utilisant ces valeurs pour évaluer la tâche d'extraction des expressions d'opinion, nous pouvons déduire que le rendement de cette étape avec un **Rappel = 0,6028 (60,28%)** et une **Précision = 0,8945 (89,45%)**, est bon par rapport à la complexité de la tâche d'extraction des expressions d'opinion, ainsi que la complexité morpho-syntaxique de la langue arabe qui rend difficile à repérer des expressions d'opinion dans le texte.

- **Extraction des caractéristiques**

Pour les caractéristiques explicites et implicites, deux mesures de performance ont été utilisés *Rappel* et *Précision* données par les équations suivantes :

$$Rappel = \frac{Nombre\ de\ caractéristiques\ correctement\ extraites}{Nombre\ de\ caractéristiques\ correctes}$$

$$Précision = \frac{Nombre\ de\ caractéristiques\ correctement\ extraites}{Nombre\ de\ caractéristuqes\ extraites}$$

Nous entendons par *caractéristiques correctement extraites* qui coïncident avec les caractéristiques étiquetées manuellement. *Caractéristiques correctes* indiquent les caractéristiques qui sont considérées comme pertinentes et devraient être extraites par notre système. *Caractéristiques extraites* indiquent le nombre total de caractéristiques extraites qui peuvent être pertinentes ou impertinentes.

a) Caractéristiques explicites

Comme nous l'avons mentionné précédemment dans le chapitre 5, une caractéristique explicite peut être l'image d'un concept ontologique sur laquelle une opinion peut être directement exprimée. Si nous réutilisons le même texte de la figure précédente, les noms et les syntagmes nominaux soulignés sont quelques caractéristiques étiquetées manuellement.

المنتخب العراقي يفوز على نظيره الفلسطيني: احتفل المنتخب العراقي بالفوز على المنتخب الفلسطيني في اول مباراة ودية له على ارضه بمدينة اربيل عاصمة اقليم كردستان العراق. وفاز الفريق العراقي على ضيفه الفلسطيني بثلاثة اهداف مقابل لا شيء، بفوز قال عنه مراسل بي بي سي في بغداد غابريل جيتهاوس انه خبر مفرح ونادر في العراق. وكانت تلك اول مباراة دولية للمنتخب العراقي في ارضه بسبب الحظر الذي فرضه الاتحاد الدولي لكرة القدم نتيجة تدهور الاوضاع الامنية التي اعقبت الغزو والاحتلال الامريكي عام 2003. وكانت آخر مباراة دولية للعراق ضد سورية في يوليو/تموز عام 2002 وفاز حينها العراق بهدفين لهدف. وكان المنتخب العراقي بطل آسيا عام 2007 يخوض مبارياته الدولية في السنوات الماضية في دول عربية مجاورة منها الأردن والإمارات. وقد خرج المنتخب العراقي من الدور الأول لكأس القارات التي اقيمت في جنوب افريقيا الشهر الماضي بعد تعادلين مع البلد المضيف ونيوزيلندا وهزيمة من اسبانيا بهدف واحد. واثر البطولة انتهى علاقة المدرب الصربي بورا ميلوتينوفيتش مع المنتخب، وخلفه في الموقع المدرب العراقي ناظم شاكر. كما ان العراق خرج من المرحلة الأولى لتصفيات آسيا المؤهلة لكأس العالم 2010 في جنوب افريقيا. ويبدو أن العراقيين يأملون في استكمال التجهيزات لاستضافة المباريات الدولية اعتبارا من عام 2011 حيث تبدأ الاستعدادات لتصفيات كأس العالم 2014.

Figure 6.7. Etiquetage manuelle pour l'extraction des caractéristiques explicites

Après un étiquetage manuel des textes pour localiser les caractéristiques pertinentes, nous avons trouvé 537 caractéristiques qui sont des instances de concepts ontologiques, considérées comme pertinentes.

Une caractéristique explicite représenté par un NP peut être constituées de plusieurs sous NPs comme dans l'exemple suivant :

(NP (NP (NNP ارضـه) (NNP بمدينـة) (NNP اربيل)) (NP (NN عاصـمة) (NP (NN اقليم) (NP (NNP كردستان) (DTNNP العراق))))))))).

Dans de tels cas, le seul NP principal est considéré comme une caractéristique candidate.

De 120 textes utilisés dans le test, et par projection de notre ontologie du domaine sur les textes étiquetés, notre système a extrait seulement 212 caractéristiques sans aucune caractéristique inutile (nous avons considéré que tous les concepts ontologiques sont pertinents car ils sont insérés manuellement), ce qui donne : **Rappel = Précision = 39,49%.**

Le facteur important qui peut être à l'origine de ce résultat modeste est que l'ontologie du domaine ne couvre pas tous les concepts et les instances pertinents.

b) Caractéristiques implicites

Cette étape chalengeuse a besoin d'une exploration profonde de la couche sémantique de l'ontologie en utilisant les concepts, les propriétés, et les hiérarchies entre concepts.

Le système utilise des expressions d'opinion et d'autres caractéristiques explicites comme indicateurs (indices) pour découvrir de nouvelles caractéristiques sur lesquelles des opinions ont été exprimées implicitement.

Du côté pratique, nous avons utilisé le Framework JENA en vue d'extraire pour chaque classe (concept) ses instances, relations, sous-classes, super-classes, domaines et portées. Comme illustré sur la figure suivante, en utilisant JENA, nous pouvons vérifier que la classe معاني _ رياضية (Concepts_Sportifs) comporte deux sous-classes, la sous-classe منظمة_رياضية (Organisation_Sportive), qui est liée à la sous-classe حدث_رياضي (Evènement_Sportif) par

l'intermédiaire de deux relations sémantiques ينظم (Organiser) et منظم_من_طرف (Organisé_par). منظمة_رياضية (Organisation_sportive) a trois instances: فيفا (FIFA), فيا (FIA), et يو أف سي (UFC). حدث_رياضي (Evènement_Sportif) a également trois instances كأس العالم (Coupe_du_Monde), سباق الدراجات النارية (Speedway), et مقابلة (match).

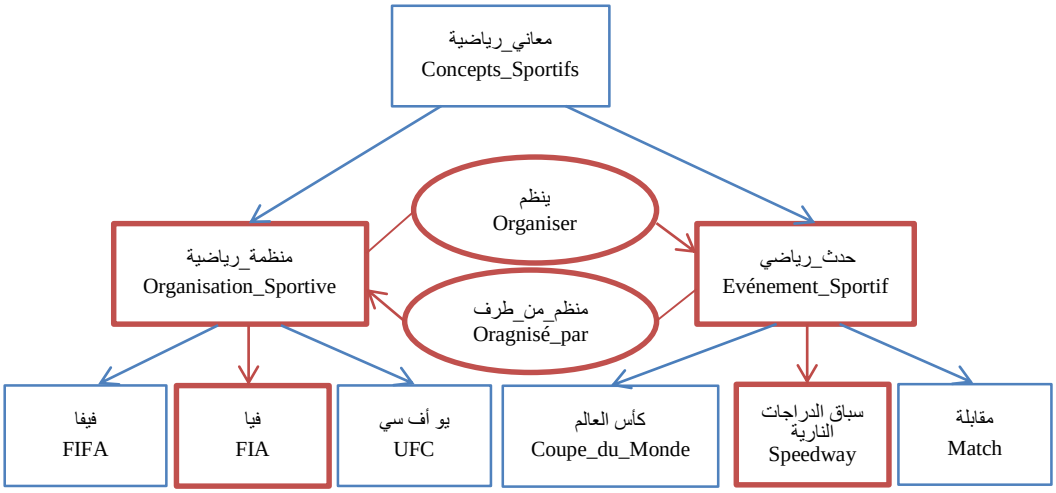


Figure 6.8 Quelques concepts ontologiques avec instances et relations

Dans le segment de texte أحسن سباق للدراجات النارية شاهدته (le meilleur speedway que j'ai vu), une opinion explicite est exprimée sur la caractéristique سباق للدراجات النارية (speedway) qui est une instance du concept حدث_رياضي (Evénement_Sportif), ce dernier est lié au concept منظمة_رياضية (Organisation_Sportive) en utilisant la relation منظم_من_طرف (Organisé_par). On peut déduire qu'une opinion positive est exprimée sur فيا (FIA), une instance de منظمة_رياضية (Organisation_Sportive).

Notons que notre ontologie n'est pas bien peuplée (absence de nombreux instances pour les concepts), ce qui a considérablement influencé le résultat de l'extraction des caractéristiques implicites. De 84 caractéristiques implicites pertinentes, seulement 49 ont été extraites par notre système. De 49 caractéristiques, 19 caractéristiques sont estimées comme caractéristiques correctement extraites par notre système, ce qui donne un **Rappel = 0,2261 (22,61%)** et de une **Précision = 0,3877 (38,77%)**. En raison de la tâche d'exploration complexe de l'ontologie, nous pouvons estimer que le résultat obtenu est relativement bon.

- **Construction des couples caractéristique-opinion**

Le but de cette étape est d'associer à chaque caractéristique explicite ou implicite extraite ses expressions d'opinion correspondantes. Le rôle d'analyse des textes est d'extraire les relations possibles entre les caractéristiques candidates et les expressions d'opinion. Dans notre travail, nous avons utilisé l'analyseur *Stanford Parser* pour extraire les dépendances grammaticales entre les caractéristiques et les expressions d'opinion.

Comme nous l'avons indiqué dans le chapitre 5, parmi les 55 modèles de dépendance proposés par Stanford Parser, nous avons choisi 7 comme modèles de dépendances binaires. Les 7 modèles sélectionnés pour être utilisés dans notre approche sont décrits dans le manuel d'utilisation des dépendances grammaticales de Stanford Parser [145].

Parmi les 212 caractéristiques explicites, 49 caractéristiques implicites et 237 expressions d'opinion extraites par notre système, et de 113 couples caractéristique-opinion prévues comme pertinents, 79 paires sont correctement construits par notre système de 12 couples évalués comme impertinents, ce qui donne un **Rappel = 59,29%**, et une **Précision = 84,81%**, où:

$$Rappel = \frac{Nombre\ de\ couples\ construits\ correctement}{Nombre\ de\ couples\ pertinents}$$

$$Précision = \frac{Nombre\ de\ couples\ costruit\ correctement}{Nombre\ de\ couples\ construits}$$

La figure suivante illustre un exemple d'association manuelle de certaines caractéristiques étiquetées (soulignées) à leurs expressions d'opinions correspondantes (en surbrillance) :

احتفل المنتخب العراقي بالفوز| مباراة ودية| فاز الفريق العراقي| خبر مفرح ونادر| تدهور الاوضاع الامنية| الغزو والاحتلال
الامريكي| فاز حينها العراق| المنتخب العراقي بطل آسيا| خرج المنتخب العراقي من الدور الأول لكأس القارات |وهزيمة من اسبانيا

Figure 6.9 Un exemple d'association manuelle des caractéristiques avec leurs expressions d'opinion

- **Classification**

Dans le domaine de fouille d'opinions, 4 mesures de performance sont généralement utilisées pour évaluer la tâche de classification ; ces mesures sont calculées à base d'une matrice de confusion. Le tableau 2 illustre notre matrice de confusion utilisée pour calculer la précision, le rappel, la pertinence et F1 (aussi appelé F-score ou F-mesure) :

	Caractéristiques prédites comme Positive	**Caractéristiques prédites comme Négative**
Caractéristiques classées comme Positive	**TP (True Positive)** : Nombre de caractéristiques classées correctement comme Positive	**FN (False Negative)** : Nombre de caractéristiques classées incorrectement comme Négative
Caractéristiques classées comme Négative	**FP (False Positive)** : Nombre de caractéristiques classées incorrectement comme Positive	**TN (True Positive)** : Nombre de caractéristiques classées correctement comme Négative

Tableau 6.3 Matrice de confusion

$$Précision = \frac{TP}{TP + FP}$$

$$Rappel = \frac{TP}{TP + FN}$$

$$Pertinence = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 \times Précision \times Rappel}{Précision + Rappel}$$

En comparant le résultat du classificateur avec une prédiction humaine, la Précision est la portion des occurrences classées correctement comme positive contre toutes les occurrences

prédites comme positives (TP+FP). Le Rappel est la portion des occurrences classées correctement comme positive contre toutes les occurrences attribuées à la classe Positive. La La Pertinence est un taux mesurant la performance du classificateur. F1 combine la Précision et le Rappel, qui est une moyenne harmonique de ces deux derniers.

Pour évaluer notre classificateur, nous devrions étendre la notion de Précision et de Rappel en définissant des valeurs distinctes pour les 5 classes (*positif fort, positif faible, neutre, négatif faible et négatif fort*). Pour cette tâche, nous avons utilisé un corpus d'apprentissage de 120 textes et 50 textes pour tester le classificateur. Football et Athlétisme sont les deux catégories de sport essentiellement couvertes par notre corpus sur lequel nous avons évalué le classificateur à l'aide de l'outil LIBSVM[8] (Annexe 3).

En vue de démontrer l'efficacité d'utilisation de la couche sémantique dans la phase d'extraction des caractéristiques implicites, nous avons effectué deux test séparés, le premier concerne seulement l'extraction des caractéristiques explicites, le deuxième concerne les caractéristiques explicites et implicites.

Les résultats du classificateur SVM sont montrés dans les deux tableaux suivants, où la précision et le rappel sont les valeurs moyennes de précision et de rappel des classes séparées :

Utilisation de l'ontologie sans la couche sémantique				
	Précision	**Rappel**	**Pertinence**	**F1**
Football	0.7727	0.7083	0.7500	0.7286
Athlétisme	0.7923	0.8900	0.8283	0.8581
Moyenne	***0.7825***	***0.7991***	***0.7891***	***0.7933***

Tableau 6.4 Résultats de classification sans l'utilisation de la couche sémantique de l'ontologie

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Utilisation de l'ontologie avec la couche sémantique				
	Précision	**Rappel**	**Pertinence**	**F1**
Football	0.9074	0.8167	0.8667	0.8409
Athlétisme	0.9061	0.9333	0.9181	0.9256
Moyenne	***0.9068***	***0.8750***	***0.8924***	***0.8833***

Tableau 6.5 Résultats de classification après utilisation de la couche sémantique de l'ontologie

6.6.2. Evaluation

Afin d'évaluer l'influence d'utilisation complète de l'ontologie, nous faisons une comparaison entre les deux expérimentations effectuées précédemment (section 6.6.1). La figure 6.10 donne un aperçu général sur les résultats obtenus :

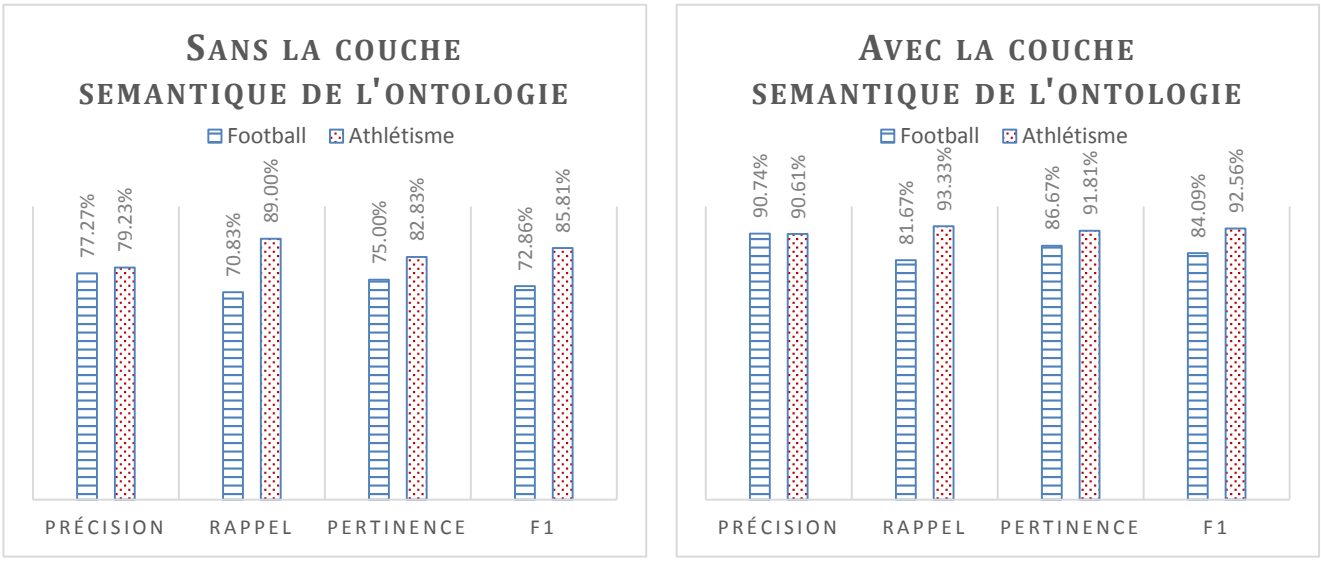


Figure 6.10 Comparaison des résultats obtenus

Les taux moyens obtenus dans la première et la deuxième expérimentation sont donnés dans la Figure 6.11, où Série 1 et Série 2 concernent respectivement les taux moyens d'expérimentation obtenus sans l'utilisation de la couche sémantique de l'ontologie et les taux moyens obtenus après utilisation de la couche sémantique :

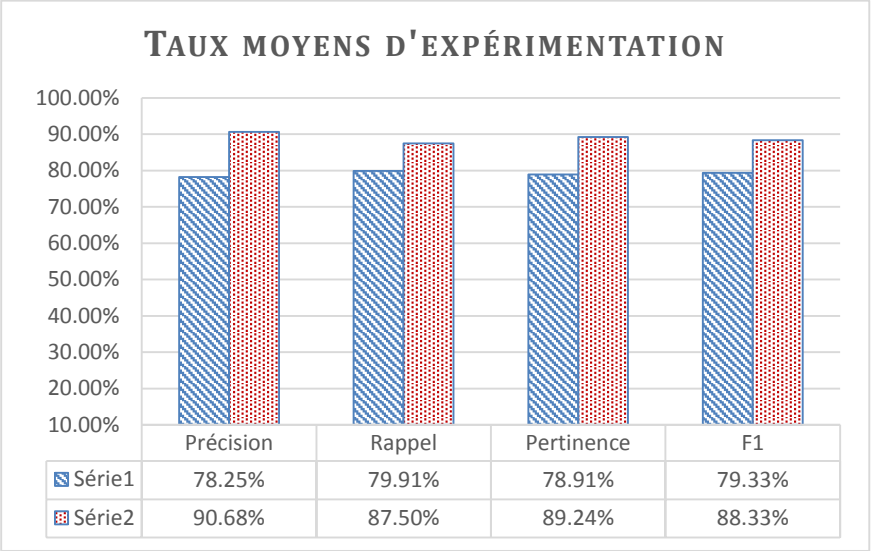


Figure 6.11 Taux moyens d'expérimentation

En général, en comparant les résultats prédits avec ceux obtenus dans tous les modules de notre système, les facteurs suivants ont influencé de manière significative les résultats obtenus :

- L'ontologie du domaine et le lexique de sentiments ne sont pas enrichis ce qui a respectivement diminué la performance de processus d'extraction des caractéristiques et le processus d'extraction des expressions d'opinion ;

- Certains concepts ne sont pas groupés, une opinion peut être exprimée sur une caractéristique ou sur l'un de ses synonymes qui associe des polarités différentes aux caractéristiques sémantiquement similaires ;

- Certaines expressions arabes morphologiques complexes n'ont pas été prises en compte : ce type d'expression a besoin d'une analyse morphologique profonde pour séparer les caractéristiques avant l'étiquetage du texte, sachant que les mots extraits peuvent inclure à la fois des expressions d'opinion et des caractéristiques.

6.7. Conclusion

Notre système d'identification d'opinions qui combine plusieurs outils et ressources linguistiques et qui s'articule sur l'utilisation des ontologies comme modèles de représentation des connaissances a démontré sa performance lors de l'exploration parfaite de l'ontologie du domaine.

Pendant l'évaluation de notre système, il était possible de prévoir un système plus performant lorsque les ressources linguistiques pour la langue arabe sont offertes d'une façon plus élaborée. C'est le cas de notre lexique de sentiments qui ne couvre pas toutes les expressions d'opinion contrairement aux autres lexiques disponibles pour l'anglais comme SentiWordNet qui est plus riche et plus structuré, ainsi que les expressions d'opinion synonymes qui ont la même polarité et intensité sont regroupés. Le corpus ne couve pas toutes les disciplines sportives d'où plusieurs concepts ontologiques pertinents ne sont pas extraits à partir des textes de notre corpus.

L'évaluation a également montré que le système peut être plus performant lorsque l'ontologie du domaine et le lexique des sentiments sont plus riches. Un tel système peut donc combiner plusieurs approches pour améliorer sa performance.

La complexité morphologique et syntaxique de certaines expressions arabes a diminué considérablement la performance d'identification des processus d'identification des expressions d'opinion et des caractéristiques.

Conclusion et Perspectives

Dans ce manuscrit, nous avons présenté une approche d'identification d'opinions dans les textes arabes, dans un cadre ontologique permettant de conceptualiser les connaissances du domaine étudié, dont le but est d'étudier la contribution des ontologies et leurs influences sur la performance du processus d'identification d'opinions.

Plusieurs ressources et techniques ont été employées pour extraire les caractéristiques (cibles de passage d'opinion), et les expressions d'opinion exprimées sur ces dernières : un corpus textuels, un lexique de sentiments, et une ontologie spécifique au domaine étudié.

Pour extraire les liens entre les caractéristiques et les expressions d'opinion en vue de construire la liste des couples caractéristique-opinion, nous avons utilisé l'outil Stanford Parser, dont 7 modèles ont été choisis pour déceler les relations grammaticales entre les concepts et les expressions d'opinions.

Après avoir testée et évaluée notre approche ONTOMART sur une partie de notre corpus textuel, nous avons constaté que l'exploration complète de l'ontologie a amélioré considérablement la performance de notre système. L'exploration de la couche sémantique de l'ontologie mène à extraire des caractéristiques implicites sur lesquelles des opinions a été exprimées.

Lors de l'évaluation, nous avons constaté que trois facteurs principaux ont été derrière la diminution des performances de notre système :

1) L'ontologie construite manuellement ne couvre pas tous les concepts du domaine étudié, ainsi le processus d'extraction de concepts ontologiques n'est pas doté d'un processus d'enrichissement automatique de l'ontologie ;

2) Le lexique de sentiments ne couvre pas toutes les expressions parce qu'il est construit comme un lexique général et n'est pas spécifique à notre domaine étudié.

3) Certaines expressions arabes morphologiquement et syntaxiquement complexes ont diminué la performance d'extraction des expressions d'opinion et leurs cibles.

Afin d'améliorer la performance de notre approche, nous proposons dans nos futurs travaux de recherches ce qui suit :

- En vue de diminuer l'ambiguïté subjectives entre expressions d'opinion, nous proposons de séparer les lexiques de sentiment par domaine, parce qu'un mot d'opinion peut être positif dans un domaine et négatif dans autre et vice-versa ;

- L'enrichissement automatique des lexiques de sentiments peut aider à construire un lexique pour couvrir un maximum d'expressions, ce qui mène à améliorer considérablement la performance du système d'identification ;

- Une ontologie du domaine riche, dotée d'un processus d'enrichissement automatique est indispensable pour améliorer la performance du processus d'extraction des cibles de passage d'opinions ;

- Une analyse plus profonde des expressions arabes complexes peut faciliter à la fois la localisation des expressions d'opinion et leurs cibles dans le texte.

Références bibliographiques

[1] Gruber T.R., (1991), "The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases", Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, Knowledge System Laboratory, Stanford University.

[2] Studer R., Benjamins V.R., Fensel D., (1998), "Knowledge engineering: principles and methods", in IEEE Transactions on Data and Knowledge Engineering, Volume 25, Issues 1–2, March 1998, pp.161–197

[3] Bourigault D., Aussenac-Gilles N., Charlet J., (2004), « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », In Techniques Informatiques et Structuration de Terminologies, Numéro Spécial de la Revue d'Intelligence Artificielle (RIA), 18(1), Hermès, Paris, 2004, pp. 87-110.

[4] Sowa J.F, (2006), "Knowledge Representation: Logical, Philosophical and computational foundations", Brooks Cole Publishing Co., Pacific Grove, 2000, 594 p.

[5] Renouf A., (2007), « Modélisation de la formulation d'applications de traitement d'images », Thèse PhD, Université de Caen, France.

[6] Handschuh S., (2005), "Creating Ontology-based Metadata by Annotation for the Semantic Web", Thèse de doctorat, Université de Karlsruhe, Allemagne.

[7] Charlet J., Bachimont B., Troncy R., (2004), « Ontologies pour le Web Sémantique », In Le Web sémantique, Charlet J., Laublet P., & Reynaud C., (Ed.), Hors-série de la Revue Information - Interaction - Intelligence (I3), 4(1), Cépaduès, Toulouse, 2004, pp. 69-100.

[8] Kaveh B., (2004), « Le rôle des ontologies de domaine dans la conception des interfaces de navigation pour des collections en ligne des musées : évaluation et proposition », Mémoire de DEA en Mangement et Technologies des Systèmes d'Information (MATIS), Université de Genève, Suisse.

[9] Nicola G., (1998), "Formal Ontology and Information Systems", Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

[10] Guarino N., (1997), "Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, SCIE 1997, M. T. Pazienza (Eds.), Springer Verlag, pp. 139-170.

[11] Gruber T.R., (1993), "Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, Vol.5, No. 2, pp.199-220.

[12] Martin P., (1995), "Using the WordNet Concept Catalogue and a Relation Hierarchy for Knowledge Acquisition". Proc. of Peirce'95, 4th, International Workshop on Peirce, University of California, Santa Cruz, USA, pp. 36-47.

[13] Klein, M., Fensel,D., Van Harmelen F., Horrocks, I., (2000), "The Relation between Ontologies and Schema-Languages: Translating OIL-Specifications to XML-Schema". In Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI-00, Berlin, Germany.

[14] Mellal, N., (2007), « Réalisation de l'interopérabilité sémantique des systèmes, basée sur les ontologies et les flux d'information », thèse de doctorat, Polytechnique de Savoie, France.

[15] Gómez-Pérez, A., Juristo N., Pazos J., (1995), "Evaluation and assessment of the knowledge sharing technology". In Towards very large knowledge bases, pp. 289–296.

[16] Motta, E., Buckingham S., (2000), "Ontology-Driven Document Enrichment: Principles, Tools and Applications". International Journal of Human-Computer Studies. Volume 52, Issue 6, pp. 1071–1109

[17] Sowa J.F, (1995), "Top-Level Ontological Categories. International Journal on Human-Computer Studies", Vol. 43, N°5/6, pp. 669-685.

[18] Uschold M., (1996), "Converting an Informal Ontology into Ontolingua: Some Experiences". A slightly abridged version of this paper appears in the Proceedings of the Workshop on Ontological Engineering held in conjunction with ECAI 96, Budapest.

[19] Van Der Vet P.E., Mars N.J.I., (1998), « Bottom-up Construction of Ontologies. IEEE Transactionson Knowledge and Data Engineering", Vol. 10, N°4, pp. 513-526.

[20] Jordão F., Mattosinho A.P., (2010), "Mining Product Opinions and Reviews on the Web", Mémoire de Master. Université de Dresden, Allemagne.

[21] Pang B., Lee L., (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, Vol. 2

[22] Li, Y., Zheng Z., Dai H., (2005), "KDD CUP-2005 report: Facing a great challenge",

SIGKDD Explorations, vol. 7, issue 2, pp. 91–99.

[23] Farek L., (2010), "Identification d'opinions dans les journaux arabes", mémoire de magistère, Université de Annaba, Algérie.

[24] Kim S.M, Hovy E., (2004), "Determining the sentiment of opinions", Proceedings of the International Conference on Computational Linguistics (COLING).

[25] Kobayashi N., Inui K., Matsumoto Y., (2007), "Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining". Proceedings of the 2007 Joint Conference on Empirical Methods in Natural.

[26] Snyder B., Barzilay R., (2007), "Multiple Aspect Ranking Using the Good Grief Algorithm", Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, HLT-NAACL 2007: pp. 300-307.

[27] Wiebe, J., Cardie C., (2005), "Annotating expressions of opinions and emotions in language". Language Resources and Evaluation (formerly Computers and the Humanities, 2005.

[28] Bethard, S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D., (2004), "Automatic extraction of opinion propositions and their holders", Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text.

[29] Choi Y., Cardie C., Riloff E., and Patwardhan S., (2005), "Identifying sources of opinions with conditional random fields and extraction patterns", Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).

[30] Liu B., (2008), « Opinion Mining & Summarization - Sentiment Analysis », Tutoriel donné à la conférence WWW-2008, Avril 21, Pékin, Chine.

[31] Hu M., Liu B., (2004), "Mining and summarizing customer reviews", Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168-177.

[32] Liu B., (2006), "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Database Management & Information Retrieval, 2ème edition, 2011, Springer Publisher.

[33] Liu B., Hu M., Cheng J., (2005), "Opinion observer: Analyzing and comparing opinions on the web", Proceedings of WWW, Chiba, Japan.

[34] Wilson T., Wiebe J., Hwa R., (2004), "Just how mad are you? Finding strong and weak opinion clauses", Proceedings of AAAI, pp. 761–769.

[35] Martin J., White P.R.R., (2005), "The Language of Evaluation: Appraisal in English", Palgrave Macmillan, ISBN-13: 978–1–4039–0409–6.

[36] Parrott W., (2001), "Emotions in Social Psychology", Psychology Press, Philadelphia.

[37] Wiebe J., Wilson T., (2002), "Learning to disambiguate potentially subjective expressions", Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 112–118.

[38] Wiebe, J., Wilson T., Cardie C., (2005), "Annotating expressions of opinions and emotions in language", Language Resources and Evaluation, Volume 39, Issue 2-3, pp 165-210. Kluwer Academic Publishers.

[39] Wiebe J., Wilson T., Bruce R., Bell M., Martin M., (2004), "Learning subjective language", Computational Linguistics, vol. 30, pp. 277–308.

[40] Wiebe J., Wilson T., Hwa R., (2004), "Just how mad are you? Finding strong and weak opinion clauses", Proceedings of AAAI, pp. 761–769.

[41] Wiebe J., Riloff E., Patwardhan S., (2006), "Feature subsumption for opinion analysis", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Sydney, Australie.

[42] Wiebe J., (2000), "Learning subjective adjectives from corpora", Proceedings of AAAI, Austin Texas, USA.

[43] Pang B., Lee L., Vaithyanathan S., (2002), "Thumbs up? Sentiment classification using machine learning techniques," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.

[44] Harb A., Dray G., Plantié M., Poncelet P., Roche M., Trousset F., (2008), « Détection d'Opinion : Apprenons les bons Adjectifs ! », Atelier FODOP'08, pp. 59-66. Fontainebleau (France).

[45] Dziczkowski G., (2008), « Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques », Thèse de doctorat, Ecole Supérieure des Mines de Paris.

[46] Poirier D., (2011), « Des textes communautaires à la recommandation », Thèse de doctorat, Université Pierre et Marie Curie - Paris 6.

[47] Gillot S., (2010), « Fouille d'opinions », Rapport de stage. Institut de Recherche en Informatique et Systèmes Aléatoires(IRISA).

[48] Reffin J., Taras E.Z, Ekaterina O.B, (2010), "Comparable Domain Dependency in Sentiment Analysis", Journal of Siberian Federal University. Humanities & Social Sciences, Vol. 3, Issue 5. pp.764-775.

[49] Charlet J., (2002), « L'ingénierie des connaissances développements, résultats et perspectives pour la gestion des connaissances médicales », Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris (France).

[50] Hernandez N., (2005), « Ontologies de domaine pour la modélisation du contexte en recherche d'information », Thèse de doctorat, Université Paul Sabatier de Toulouse (France).

[51] Mizoguchi R., (2004), « Le rôle de l'ingénierie ontologique dans le domaine des EIAH », Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF). Volume 11.

[52] Hubert G., (2010), « Recherche d'information et contexte », Habilitation à diriger des recherches, Université Toulouse 3 – Paul Sabatier.

[53] Verna G., (2006) « Les Défis de la Gestion des Connaissances en Contexte Interculturel », Essai de Maitrise en Administration des Affaires, Université de Laval, Québec.

[54] Allan Ed., (2003), "Challenges in information retrieval and language modeling", SIGIR Forum, Volume 37 Issue 1, pp 31-47.

[55] Aussenac-Gilles N., Gleizes M.P, Haemmerlé O., Mothe J., (2007), « OntoTextes (Ontologies et Textes) », Bilan du projet de recherche financé au titre du BQR 2007/AO1. Rapport de recherche, IRIT/RR--2007-24--FR, IRIT.

[56] Bachimont B., (2004), « Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle », Mémoire d'Habilitation à Diriger des Recherches, Université de Technologie de Compiègne.

[57] Charlet J., Kassel G., Zacklad M., Borigault D., (2000), « Ingénierie des connaissances : recherches et perspectives », In Ingénierie des connaissances, Évolutions récentes et nouveaux défis, Edition Eyrolles, Paris, ISBN 2-212-09110-9.

[58] Woolf H., (1990), "Websters New World Dictionary of the American Language", Warner Books, New York, 0446360260 (ISBN13: 9780446360265).

[59] Sowa, J.F., (1984), "Conceptual Structures: Information Processing in Mind and Machine", Addison-Wesley Publishing Company, USA.

[60] Lame G., (2002), « Construction d'ontologie à partir des textes, une ontologie du droit dédiée à la recherche d'information sur le Web », Thèse de doctorat, Ecole des Mines de Paris.

[61] Kayser D., (1997), « La représentation des connaissances », Hermes Publisher, ISBN 2-86601-647-5.

[62] Guarino N., Carrara M., Giaretta P., (1994), "Formalizing ontological commitments", In Proceedings of the AAAI conference.

[63] Guarino N., Oberle D., Staab S., (2009), "What is an Ontology?" ITSC-CNR, Laboratory for Applied Ontology, Université de Trento, Italie,

[64] Aussenac-Gilles N., Mothe J., (2004), « Ontologies as Background Knowledge to Explore Document Collections », In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), pp 129-142.

[65] Guarino N., (1998), "Formal Ontology and Information Systems, In Formal Ontology in Information Systems", Edition IOS Press. pp 3-15.

[66] Berners-Lee T, Hendler J., Lassila O., (2001), "The Semantic Web", Scientific American, pp 28–37.

[67] Boukhadra A., (2011), « La composition dynamique des services Web sémantiques à base d'alignement des ontologies owl-s », Mémoire de Magister, Ecole National Supérieur d'Informatique – Alger (Algérie).

[68] Bachimont B., (2000), « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances », In J. Charlet, M.

Zacklad, G. Kassel & D. Bourigault (Eds.), Ingénierie des connaissances, évolutions récentes et nouveaux défis. Paris: Eyrolles.

[69] Kent R., (2001), "The iff foundation ontology", CAI Ontology Workshop.

[70] Ghafour A.S., (2004), « Méthodes et outils pour l'intégration des ontologies », Rapport de Stage de DEA, Université Claude Bernard, Lyon1 (France).

[71] Gómez-Pérez A., Moreno A., Pazos J., Sierra-Alonso A., (2000), "Knowledge Maps: An essential technique for conceptualization", In Data & Knowledge Engineering, Volume 33, issue 2, pp 169-190.

[72] Gómez-Pérez A., Fernandez M., (1996), "Towards a Method to Conceptualize Domain Ontologies, In Proceedings of the European Conference on Artificial Intelligence (ECAI'96), pp 41–52.

[73] Hoffmann P., (2004), « Appariement contextuel d'ontologies », Rapport de Stage de DEA, Université Claude Bernard, Lyon1(France).

[74] Van Heijst G., Schreiber G., Wielinga B., (1997), "Using explicit ontologies for KBS development", International Journal of Human-Computer Studies, Volume 42, issue (2/3), pp 183-292.

[75] Mokhtari N., (2010), « Extraction et exploitation d'annotations sémantiques contextuelles à partir de texte », Thèse de doctorat, Université de Nice-Sophia Antipolis.

[76] Mondary T., Després S., Nazarenko A., Szulman S., « Construction d'ontologies à partir de textes : la phase de conceptualisation », Conférence Ingénierie des Connaissances (IC 2008), Loria, Nancy (France)

[78] Khalfi S., (2012), « Construction d'une ontologie pour la prise en charge des patients à domicile », Mémoire de Magister, Université Mantouri de Constantine (Algérie).

[79] Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A., (2003), « Methodologies, tools and languages for building ontologies. Where is their meeting point?", Data & Knowledge Engineering. Volume 46, issue 1, pp. 41–64. Elsevier.

[80] Ushold M., King M., (1995), "Towards a Methodology for Building Ontologies", Workshop on Basic Ontological Issues in Knowledge Sharing.

[81] Uschold M., Gruninger M., (1996), « Ontologies: Principles Methods and Applications », Knowledge Engineering Review Volume 11, pp. 93-136.

[82] Grüninger, M., Fox, M., (1995). "Methodology for the design and evaluation of ontologies", IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, pp. 6.1-6.10.

[83] Kactus (1996). "The KACTUS Booklet version 1.0", Esprit Project 8145.

[84] Swartout, B., Ramesh, P., Knight, K., Russ, T. (1997), "Toward Distributed Use of Large-Scale Ontologies", AAAI'97 Spring Symposium on Ontological Engineering, (pp. 138–148). Stanford University, California.

[85] Dahmani B. F., (2010), « Modélisation basée ontologie pour l'apprentissage interactif – Application à l'évaluation des connaissances de l'apprenant », Thèse de doctorat, Université Mouloud Mammri de Tizi-Ouzou (Algérie).

[86] Handschuh S., (2005), "Creating Ontology-based Metadata by Annotation for the Semantic Web", Thèse de doctorat, Université de Karlsruhe (Allemagne).

[87] Florence A., (2007), « Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle », Thèse de doctorat, Université Paris X – Nanterre (France).

[88] Carbonell J., (1997), "Subjective Understanding: Computer Models of Belief Systems". Thèse PhD, Université Yale (USA).

[89] Wilks Y., Bien J., (1984), "Beliefs, points of view and multiple environments". In Proc. of the international NATO symposium on Artificial and human intelligence, pp. 147-171, New York (USA). Elsevier North-Holland, Inc.

[90] Dave K., Lawrence S., Pennock D. M., (2003), "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", In WWW '03: Proceedings of the 12th international conference on World Wide Web, pp. 519-528, New York (USA).

[91] Ounis, I. M., Macdonald C., Mishne G., Soboroff I., (2006), "Overview of the Trec-2006 Blog Track". In Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006), pp. 17-31.

[92] Grouin C., Arnulphy B., Berthelin J.-B., El Ayari S., Garcia-Fernandez A., Grappy A., Hurault-Plantet M., Paroubek P., Robba I., Zweigenbaum P., (2009), « Présentation de l'édition 2009 du DEfi Fouille de Textes (DEFT'09) ». In Actes de l'atelier de clôture de la cinquième édition du DEfi Fouille de Textes, pp. 35–50, Paris (France).

[93] Liu B., (2010), "Sentiment analysis and subjectivity", In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL, ISBN 978-1420085921.

[94] Turney P. D., Littman M. L., (2003), "Measuring praise and criticism: Inference of semantic orientation from association". ACM Trans. Inf. Syst., Volume 21, issue 4, pp.315-346.

[95] Andreevskaia A., Bergler S., (2006), "Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses", Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics.

[96] Liu B., (2009), "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data Centric Systems and Applications)". Springer-Verlag New York, Inc. Secaucus, NJ, (USA), ISBN: 3540378812, 1st edition.

[97] Turney. P. D. (2002), "Thumbs up or thumbs down ?: semantic orientation applied to unsupervised classification of reviews". In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424, Morristown, NJ, (USA).

[98] Yu H., Hatzivassiloglou V., (2003), "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences". In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp.129-136, Morristown, NJ, (USA).

[99] Pereira F., Tishby N., Lee L., (1993), "Distributional clustering of english words". In Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 183-190, Morristown, NJ, (USA).

[100] Lin D., (1998), "Automatic retrieval and clustering of similar words". In Proceedings of the 17th international conference on Computational linguistics, pp. 768-774, Morristown, NJ, (USA).

[101] Hatzivassiloglou V., McKeown K. R., (1997), "Predicting the semantic orientation of adjectives". In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pp. 174–181, Morristown, NJ, (USA).

[102] Kanayama H., Nasukawa T., (2006), "Fully automatic lexicon expansion for domain-oriented sentiment analysis", EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 355-363

[103] Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K. J., (1990), "Introduction to wordnet : an on-line lexical database". Int J Lexicography, Volume 3, issue 4, pp. 235-244.

[104] Hu M., Liu B., (2004), "Mining opinion features in customer reviews", In Deborah L. Mcguinness, George Ferguson, Deborah L. Mcguinness, and George Ferguson, editors, AAAI,. AAAI Press, pp. 755–760

[105] Ounis I., Macdonald C., Soboroff I., (2008), "Overview of the trec-2008 blog track". In Proceedings of The seventeenth Text REtrieval Conference (TREC 2008). NIST.

[106] Salton G., Lesk M. E., (1995), "The smart automatic document retrieval systems—an illustration". Commun. ACM, Volume8, issue 6, New York (USA).

[107] Nigam K., Hurst M., (2006) "Towards a robust metric of polarity. In Computing Attitude and Affect in Text: Theory and Applications", The Information Retrieval Series Volume 20, 2006, pp. 265-279 Dordrecht, the Netherlands. Springer.

[108] Crestan E., Gigandet S., Vinot R., (2007), « Approches naïves à l'analyse d'opinion ». In Actes de l'atelier de clôture du 3ème DEfi Fouille de Textes, pp. 45–56, Grenoble, (France).

[109] Pang B., Lee L., (2004), "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts". In ACL'04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article N° 271, Morristown, NJ, (USA).

[110] Plantié M., Roche M., Dray G., Poncelet P., (2008), "Is a voting approach accurate for opinion mining ?", In DaWaK '08 : Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, pp. 413-422, Berlin, Heidelberg. Springer Verlag.

[111] Généreux M., Santini M., (2007), « Défi : Classification de textes français subjectifs ». In Actes de l'atelier de clôture du 3ème DEfi Fouille de Textes, pp. 83–93, Grenoble (France).

[112] Trinh A., (2007), « Classification de texte et estimation probabiliste par machine à vecteurs de support ». In Actes de l'atelier de clôture du 3ème DEfi Fouille de Textes, pages 69–82, Grenoble (France).

[113] Wilson T., Wiebe J., Hwa R., (2004), "Just how mad are you ? Finding strong and weak opinion clauses". In In Proceedings of AAAI, pp. 761-769.

[114] Schapire R. E. Singer Y., (2000), "Boostexter: A boosting-based system for text categorization. Machine Learning", Volume 39, Issue (2/3), pp.135–168.

[115] Cohen W., (1996), "Learning trees and rules with set-valued features". In Proceedings of the 13th National Conference on Artificial Intelligence, pp. 709-716. AAAI Press.

[116] Joachims T., (1999), "Making large-scale support vector machine learning practical". Advances in kernel methods: support vector learning, pp. 169–184. MIT Press Cambridge, MA, (USA).

[117] Nigam K., Hurst M., (2006), "Towards a robust metric of polarity". In Computing Attitude and Affect in Text: Theory and Applications, pp. 265–279.

[118] Riloff E., Wiebe J., (2003), "Learning extraction patterns for subjective expressions", In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 105–112, Morristown, NJ, (USA).

[119] Dziczkowski G., Wegrzyn-Wolska K, (2008), "An autonomous system designed for automatic detection and rating of film reviews". Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, Volume 1, pp. 847-850.

[120] Benamara F., Cesarano C., Picariello A., Reforgiato D., Subrahmanian V., (2007), "Sentiment analysis : Adjectives and adverbs are better than adjectives alone", In International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado, (U.S.A), pp. 203–206, AAAI Press.

[121] Das S., Chen M., (2001), "Yahoo! for amazon: Extracting market sentiment from stock message boards". Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

[122] Popescu A.M., Etzioni O., (2005), "Extracting Product Features and Opinions from Reviews". In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 339-346, Association for Computational Linguistics Stroudsburg, PA, (USA).

[123] Carenini G., Raymond T.N., Zwart E., (2005), "Extracting Knowledge from Evaluative Text". In Proceedings of the 3rd international conference on Knowledge capture, pp. 11-18, ACM New York (USA).

[124] Gamon M., Anthony A., Simon C.O., Ringger E., (2005), "Pulse: Mining Customer Opinions from Free Text", In Proceedings of International symposium on intelligent data analysis N°6, Madrid, pp.121-132, Springer-Verlag Berlin, Heidelberg.

[125] Goldensohn B., Kerry H., McDonald R., Tyler N., George A.R, Reynar J., (2008), "Building a Sentiment Summarizer for Local Service Reviews", WWW2008 Workshop: Natural Language Processing Challenges in the Information Explosion Era (NLPIX 2008).

[126] Zhao L., Chunping L., (2009), "Ontology Based Opinion Mining for Movie Reviews Knowledge Science, Engineering and Management Lecture Notes in Computer Science Volume 5914, pp 204-214.

[127] Feiguina O., (2006), « Résumé automatique des commentaires de Consommateurs ». Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc. en informatique, Département d'informatique et de recherche opérationnelle, Université de Montréal.

[128] Xiwen C., Xu F., (2008), "Fine-grained Opinion Topic and Polarity Identification". In Proceedings of the Sixth International Language Resources and Evaluation (LREC' 08), Marrakech, Morocco.

[129] Pimwadee C., Zhou L., (2005), "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches". Proceedings of HICSS-05, the 38th Hawaii International Conference on System Sciences.

[130] Vernier M., Monceaux L., Daille B., (2011), « Identifier la cible d'un passage d'opinion dans un corpus multithématique», TALN 2011, Montpellier (France) ;

[131] Hatzivassiloglou V., Wiebe J.M., (2000), « Effects of Adjective Orientation and Gradability on Sentence Subjectivity », Proceedings of the 18th International Conference on Computational Linguistics, Volume 1, issue 11-12, pp. 299-305.

[132] Wiebe J.M., (1999), « Development and Use of a Gold-Standard Data Set for Subjectivity Classifications », Proceedings of the Association for Computational Linguistics (ACL), Volume 37, p. 246-253.

[133] Xiaowen D., Liu B., (2007), « The Utility of Linguistic Rules in Opinion Mining SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 811-812, ACM New York (USA).

[134] Cadilhac A., Benamara F., Aussenac-Gilles N., « Ontolexical resources for feature based opinion mining: a case-study », Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pp. 77-86, Beijing (China).

[135] Ohana B., Tierney B., (2009), « Sentiment Classification of Reviews Using SentiWordNet », 9th IT & T Conference School of Computing. Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd.

[136] Baccianella S, Esuli A., Sebastiani F., (2010), "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", In Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT, 2010, pp. 2200–2204.

[137] Maurel, S., Curtoni, P., et Dini, L. (2008), « L'analyse des sentiments dans les forums », Dans Atelier Fodop**?**, pp. 9–22.

[138] Sanchez D., Moreno A. (2008), "Learning non-taxonomic relationships from web documents for ontology domain construction". In Data & Knowledge Engineering. Volume 64, Issue 3, pp. 600–623

[139] Yoshida Y., Hirao T., Iwata T., Nagata M., Matsumoto Y., (2011), "Transfer Learning for Multiple-Domain Sentiment Analysis — Identifying Domain Dependent/Independent Word Polarity", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

[140] Taysir H., Soliman, A. Ali M., Hedar A.R, Doss M.M., (2013) "Mining Social Networks' Arabic Slang Comments", In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13), Prague, Czech Republic.

[141] Abdul-Mageed M., Diab, M., (2012) "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis". In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA).

[142] Elarnaoty M., Abdel R.S., Fahmy, A., (2012), "A Machine Learning Approach for Opinion Holder Extraction Arabic Language". CoRR, abs/1206.1011.

[143] El-Halees A., (2011), "Arabic opinion mining using combined classification approach". In the proceeding of the International Arab Conference on Information Technology (ACIT'2011), Riyadh, Saudi Arabia.

[144] Motaz K.S, Wesam A., (2010), "OSAC: Open Source Arabic Corpora", International Conference on Electrical and Computer Systems (EECS'10), Lefke, North Cyprus.

[145] Marie-Catherine M., Christopher D.M., (2008), "Stanford typed dependencies manual", Revised for the Stanford Parser v. 3.3.

Annexes

Annexe 1. Apache Jena Framwork

Jena est une API Java pour les applications Web sémantique. Le paquet-clé RDF pour le développeur d'applications est **com.hp.hpl.jena.rdf.model**. L'API a été définie en termes d'interfaces donc le code de l'application peut fonctionner avec différentes implémentations sans changement. Ce paquet contient des interfaces pour représenter des modèles, des ressources, des propriétés, des littéraux, des déclarations et tous les autres concepts-clés de RDF, et une **ModelFactory** pour de classes spécifiques.

Les paquets **com.hp.hpl...jenaimpl** contiennent l'implémentation des classes qui peuvent être communes à plusieurs implémentations. Par exemple, elles définissent les classes **ResourceImpl**, **PropertyImpl** et **LiteralImpl**, qui peuvent être utilisées directement ou dérivées par différentes implémentations. Les applications doivent rarement, voire jamais, utiliser ces classes directement. Par exemple, plutôt que de créer une nouvelle instance de **ResourceImpl**, il vaut mieux utiliser la méthode **createResource** quel que soit le modèle utilisé. Ainsi, si l'implémentation du modèle a utilisé une implémentation optimisée de **Resource**, alors aucune conversion entre les deux types ne sera nécessaire.

- **Création de modèles d'ontologies en utilisant Jena**

Un modèle d'ontologie est celui qui présente le RDF comme une ontologie : des classes, des individus, des différents types de propriétés et ainsi de suite. Jena prend en charge les ontologies RDFS et OWL :

createOntologyModel() crée un modèle d'ontologie qui est en mémoire et présente des ontologies OWL.

createOntologyModel (OntModelSpec spec, Model base) crée un modèle d'ontologie selon les spécifications de ***OntModelSpec*** qui présente l'ontologie de base.

createOntologyModel (OntModelSpec spec, ModelMaker maker, Model base) crée un modèle d'ontologie OWL selon les spécifications du modèle de base. Si le modèle de

l'ontologie doit construire des modèles supplémentaires (pour les importations OWL), ont doit utiliser ***ModelMaker*** pour les créer.

- **SPARQL pour interroger une ontologie**

SPARQL est un protocole (W3C du 15 janvier 2008) et un langage de requêtes qui permet d'exploiter l'approche sémantique des données RDF.

Il est doté :

- d'un langage de requêtes avec syntaxe basée sur des triplets
- d'un protocole d'accès comme un service Web (SOAP)
- d'un langage de présentation des résultats (XML)

SPARQL cible donc l'interrogation de métadonnées RDF, structure de base du Web sémantique.

Il fonctionne en parfaite synergie avec les autres technologies Web sémantique du W3C :

- RDF (Resource Description Framework) pour la représentation des données,
- RDFS (schéma RDF),
- OWL (Web Ontology Language) pour la création de vocabulaires,
- GRDDL.

Exemple de la syntaxe en triplets simplifiée avec des points d'interrogation pour marquer les variables :

```
?x rdf:type ex:Personne
```

Langage de patterns à matcher :

```
select ?sujet ? propriete ?valeur where

{?sujet?propriete?valeur}
```

Le pattern est par défaut une conjonction de triplets

Il existe deux formes possibles pour la présentation des résultats :

- le binding i.e. la liste des valeurs sélectionnées pour chaque réponse rencontrée (format XML stable ; bien avec XSLT) ;

- les sous graphes des réponses rencontrées en RDF (format RDF/XML, bien pour applications utilisant RDF)

Dans le cas des services Web 2.0 actuels, les Web services sont certes disponibles, mais ne sont pas normalisés, il faut donc connaître les méthodes du Web services et la structure des données pour les interroger.

Avec SPARQL, nous n'avons pas besoin de connaître a priori la structure et le contenu des données pour pouvoir les interroger. En effet, Sparql permet d'interroger n'importe quel composant d'un triplet qui a la forme Sujet-Prédicat-Objet.

Exemple :

Si on veut connaître tous les triplets qui composent un fichier RDF, sans rien connaître a priori des ressources décrites, des propriétés utilisés ou du contenu, on effectue la requête suivante :

```
SELECT ?sujet ?predicat ?objet
WHERE {
   ?sujet ?predicat ?objet
}
```

Le nom des variables est précédé d'un point d'interrogation.

La ligne SELECT permet de sélectionner l'ensemble des tuples, ou lignes de variables (sujet, predicat, object) correspondant aux contraintes de la clause WHERE.

Annexe 2. Stanford Parser

L'analyseur de Stanford est un analyseur de langage naturel du groupe de traitement du langage naturel de Stanford. Utilisé pour analyser les données d'entrée écrites en plusieurs langues comme l'anglais, l'allemand, l'arabe et le chinois, il a été développé et maintenu depuis 2002, principalement par Dan Klein et Christopher Manning. L'application est sous licence GNU GPL, mais la licence commerciale est également disponible.

- **Installation et exigences**

L'analyseur fonctionne sous Windows et Unix/Linux/MacOSX et nécessite un Java Runtime Environment (JRE) (Java 1.5 ou supérieur). Le paquet contient également une base interface graphique (GUI) pour la visualisation des arbres de la structure.

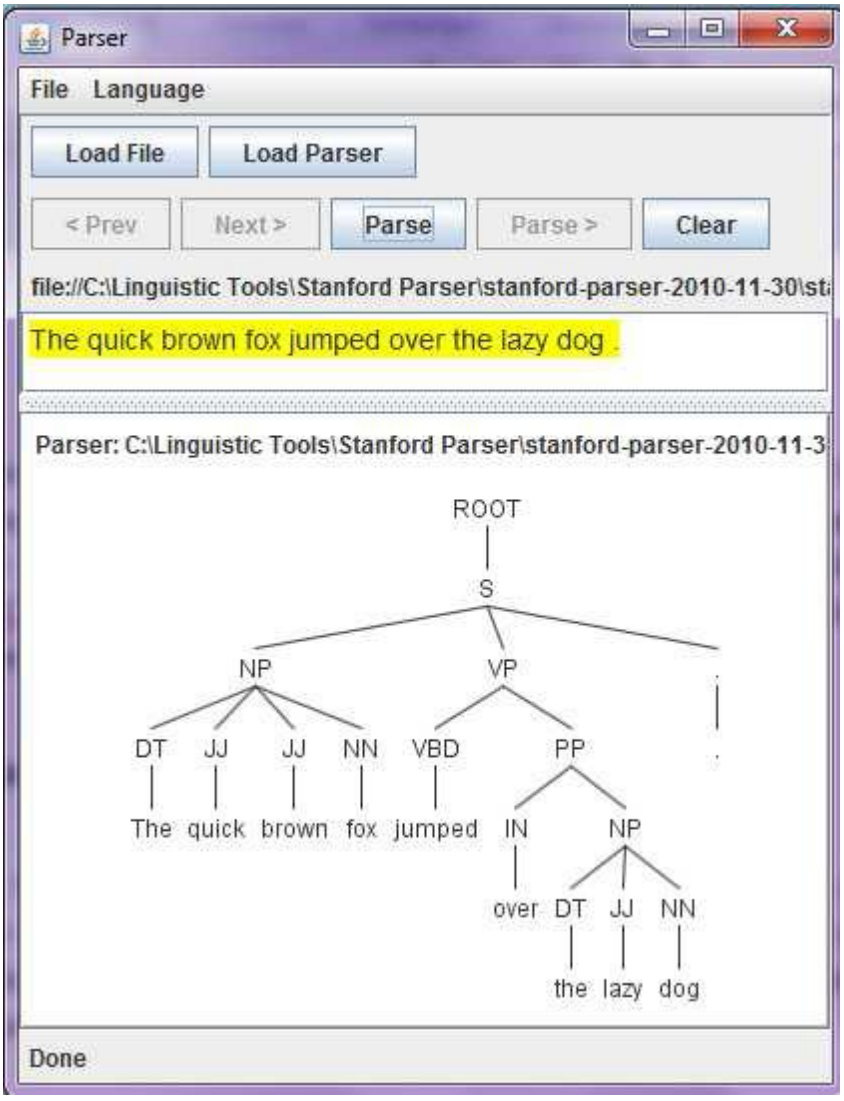
- **Exécution de l'analyseur**

Modèles de l'analyseur

Des analyseurs pour des différentes langues comme le chinois, l'arabe, l'anglais et l'allemand sont fournis. Le modèle d'analyseur appelé **FACTORED** est plus complexe et nécessite plus de mémoire, car il contient deux grammaires et conduit au système de fonctionner trois analyseurs. En outre, il y a deux analyseurs des fichiers **wsjPCFG.ser.gz** et **wsjFACTORED.ser.gz**.

Utilisation de l'interface graphique

L'utilisation de l'interface graphique est recommandée lorsque vous utilisez l'analyseur de Stanford pour la première fois. Afin de commencer l'application il suffit de lancer le fichier **lexparser-gui.bat** (sous Windows). Des phrases simples qui peuvent être saisies ou reçues en ouvrant un fichier texte peuvent être étiquetées après avoir sélectionné un fichier de l'analyseur. En outre, une visualisation de l'arbre de structure peut être vue comme dans la figure suivante :



Utilisation de la ligne de commandes

L'utilisation de la ligne de commandes est recommandée car il fournit un contrôle plus fin sur le processus d'analyse. Pour appliquer l'analyseur de Stanford, aller dans le répertoire où vous avez extrait l'analyseur et tapez les commandes suivantes sur une ligne de commande :

```
java -mx150m -cp stanford-parser.jar
edu.stanford.nlp.parser.lexparser.LexicalizedParser
OPTIONS
parserFile input1 input2 ...
```

Annexe 3. LIBSVM

LIBSVM est une bibliothèque populaire d'apprentissage automatique, open source, développée à l'Université nationale de Taiwan et écrit en C ++ mais avec un API C. LIBSVM implémente l'algorithme SMO pour Machines à Vecteurs de Support (SVM), supportant la classification et la régression. Son objectif est de laisser les utilisateurs peuvent facilement utiliser SVM comme un outil.

Les différentes formulations utilisées sont : C-support vector classification (C-SVC), ν-support vector classification (ν-SVC), distribution estimation (one-class SVM), ε-support vector regression (ε-SVR), and ν-support vector regression (ν-SVR).

Une utilisation typique de LIBSVM comporte deux étapes : d'abord, l'apprentissage d'un ensemble de données pour obtenir un modèle, ensuite l'utilisation de ce modèle pour prédire les classes d'un ensemble de données de test. Pour SVC et SVR, LIBSVM peut également estimer la probabilité de sortie. De nombreuses extensions de LIBSVM sont disponibles à libsvmtools[1].

Structure de package LIBSVM

Le package de LIBSVM est structuré comme suit :

1. le répertoire principal : programmes noyau C / C ++ et des exemples de données. En particulier **svm.cpp** met en œuvre des algorithmes d'apprentissage et de test.

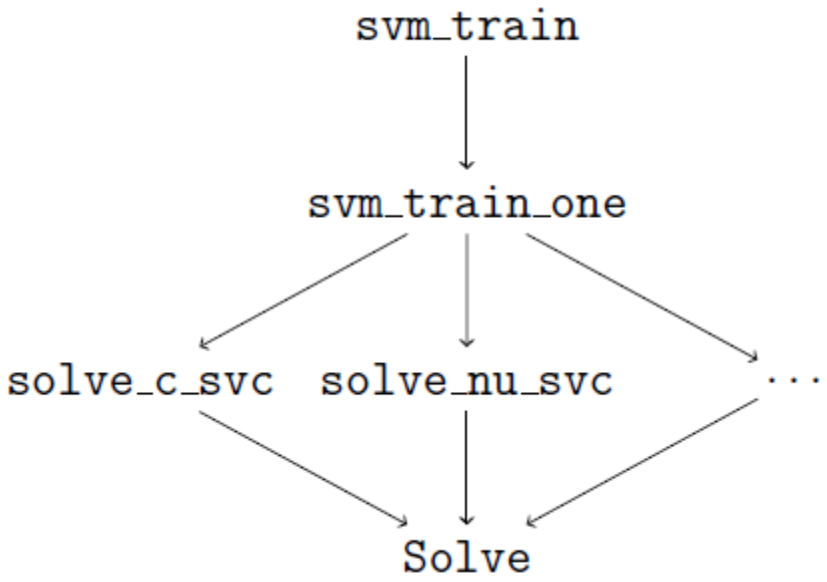
2. le sous-répertoire de l'outil : ce sous-répertoire inclut des outils pour la vérification des formats de données et la sélection des paramètres de SVM.

3. autres sous-répertoires contiennent des fichiers binaires préconstruits et des interfaces langages / logiciel.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

Organisation du code

Tous les algorithmes d'apprentissage et de test sont implémentés dans le fichier **svm.cpp**. Les deux principaux sous-programmes sont **svm_train** et **svm_predict**. La procédure d'apprentissage est plus sophistiquée, si nous donnons l'organisation du code dans la figure suivante :



Pour la classification, **svm_train** découple un problème multi-classe à un problème à deux classes et appelle **svm_train_one** plusieurs fois. Pour la régression et une seule classe SVM, il appelle directement **svm_train_one**.

Les sorties de probabilité pour la classification et la régression sont également traitées dans **svm_train**, ensuite, selon la formulation SVM, **svm_train _one** appelle une sous-routine correspondante comme **solve_c_svc** pour **v-SVC** et **solve_nu_svc** for **v-SVC**. Toutes les sous-routines **solve_*** appellent le solveur **Solve** après la préparation des valeurs d'entrée appropriées. La sous-routine **Solve** minimise une forme générale du problème d'optimisation.