ED n° 431 : Information, communication, modélisation et simulation.

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'ECOLE NATIONALE SUPERIEURE DES MINES DE PARIS

Spécialité "Automatique, Robotique, et Informatique Temps Réel"

présentée et soutenue publiquement par
Grzegorz DZICZKOWSKI

le 4 Décembre 2008

Analyse des sentiments :
système autonome d'exploration des opinions
exprimées dans les critiques cinématographiques

Directeur de thèse : Robert Mahl
Co-encadrant : Katarzyna Węgrzyn-Wolska

Jury

M.	Jean-Jacques Girardot	Président, Rapporteur
M.	Witold Kosiński	Rapporteur
M.	Gaël Dias	Examinateur
M.	Robert Mahl	Directeur, Examinateur
M.	Katarzyna Węgrzyn-Wolska	Directeur, Examinateur

Résumé

Analyse des sentiments - système autonome d'exploration des opinions exprimées dans les critiques cinématographiques.
Directeur de thèse : Robert Mahl, ENSMP,
Co-encadrement : Katarzyna Wegrzyn-Wolska, ESIGETEL,

Cette thèse décrit l'étude et le développement d'un système conçu pour l'évaluation des sentiments des critiques cinématographiques. Un tel système permet :
– la recherche automatique des critiques sur Internet,
– l'évaluation et la notation des opinions des critiques cinématographiques,
– la publication des résultats.
Afin d'améliorer les résultats d'application des algorithmes prédicatifs, l'objectif de ce système est de fournir un système de support pour les moteurs de prédiction analysant les profils des utilisateurs. Premièrement, le système recherche et récupère les probables critiques cinématographiques de l'Internet, en particulier celles exprimées par les commentateurs prolifiques. Par la suite, le système procède à une évaluation et à une notation de l'opinion exprimée dans ces critiques cinématographiques pour automatiquement associer une note numérique à chaque critique ; tel est l'objectif du système. La dernière étape est de regrouper les critiques (ainsi que les notes) avec l'utilisateur qui les a écrites afin de créer des profils complets, et de mettre à disposition ces profils pour les moteurs de prédictions.

Pour le développement de ce système, les travaux de recherche de cette thèse portaient essentiellement sur la notation des sentiments ; ces travaux s'insérant dans les domaines de *ang : Opinion Mining* et d'*Analyse des Sentiments*. Notre système utilise trois méthodes différentes pour le classement

des opinions. Nous présentons deux nouvelles méthodes ; une fondée sur les connaissances linguistiques et une fondée sur la limite de traitement statistique et linguistique. Les résultats obtenus sont ensuite comparés avec la méthode statistique basée sur le classificateur de Bayes, largement utilisée dans le domaine.

Il est nécessaire ensuite de combiner les résultats obtenus, afin de rendre l'évaluation finale aussi précise que possible. Pour cette tâche nous avons utilisé un quatrième classificateur basé sur les réseaux de neurones.

Notre notation des sentiments à savoir la notation des critiques est effectuée sur une échelle de 1 à 5. Cette notation demande une analyse linguistique plus profonde qu'une notation seulement binaire : positive ou négative, éventuellement subjective ou objective, habituellement utilisée.

Cette thèse présente de manière globale tous les modules du système conçu et de manière plus détaillée la partie de notation de l'opinion. En particulier, nous mettrons en évidence les avantages de l'analyse linguistique profonde moins utilisée dans le domaine de l'analyse des sentiments que l'analyse statistique.

Mots clefs : Opinion Mining, Analyse des Sentiments, Classification du Texte, Catégorisation du Texte, Traitement Automatique de la Langue Naturelle (TALN), Information Retrieval, Moteur de Prédiction.

A Anna

A toute ma famille

A tous ceux qui me sont chers...

Remerciement

J'adresse tout d'abord, ma profonde reconnaissance et mes vifs remerciements à M. Robert MAHL pour avoir accepté de diriger cette thèse. Je le remercie pour son soutien et la patience dont il a fait preuve à mon égard durant cette thèse.

J'exprime également ma profonde gratitude à Mme Katarzyna WEGRZYN-WOLSKA responsable du laboratoire LRIT-IR à l'ESIGETEL. Elle m'a initié aux techniques de catégorisation de texte, m'a encouragé à poursuivre dans cette voie et m'a accueilli dans son équipe.

Je remercie M. Jean-Jacques GIRARDOT professeur à l'École Nationale Supérieure des Mines de Saint-Étienne et M. Witold KOSINSKI professeur à Polish-Japanese Institute of Information Technology, pour m'avoir fait l'honneur d'accepter la lourde tâche d'être rapporteur de cette thèse.

Je tiens à remercier également M. Gaël DIAS chercheur à Université de Beira Interior qui m'a fait l'honneur d'être membre du jury.

Mes remerciements vont naturellement aux voisins de bureau, professeurs, amis et collègues, agents administratifs que j'ai côtoyé durant ces années à l'ESIGETEL. Ils m'ont aidé et ont tout simplement rendu le déroulement de cette thèse agréable. Sans oublier M. Blaise Collin, Mlle Catherine Bernard et M. Lamine Bougueroua pour leurs commentaires et leurs contributions à la correction de mon rapport.

Je remercie mes collègues et amis thésards. Je leur exprime ma profonde sympathie et leur souhaite une bonne continuation.

Finalement je remercie chaleureusement, tous les membres de ma famille pour la confiance qu'ils m'accordent, leur amour, et leurs encouragements et surtout je remercie infiniment ma chère Anna qui m'a soutenu et encouragé pendant ces années.

Table des matières

Table des figures

Liste des tableaux

LISTE DES TABLEAUX

Chapitre 1

Introduction

1.1 Présentation du sujet

De nos jours, l'Internet est un outil incontournable d'échange d'information tant au niveau personnel que professionnel. Le Web nous offre un monde de l'information prodigieux et a évolué des simples ensembles de pages statiques vers des services de plus en plus complexes. Ces services nous offrent l'achat de tous les produits, la lecture de son journal préféré en ligne, la rencontre de l'âme soeur, la discussion sur de multiples forums ou la possibilité de s'exprimer sur les blogs.

L'Internet contient un nombre énorme d'informations, et pour la plupart d'entre nous c'est le premiér lieu pour trouver ces informations, réserver l'avion ou l'hôtel, acheter des produits, consulter les avis d'autres utilisateurs sur les produits qui nous intéressent, lire les commentaires avant de choisir le film à voir au cinéma, voir des propositions d'autres personnes avant de choisir les cadeaux de mariage etc. Le problème principal n'est plus de savoir si l'information se trouve sur le Web mais comment la trouver car le flux informationnel est extrêmemment bruité. Un autre problème, non liée à Internet lui même mais plutôt à des considérations sociologiques, est que la globalisation nous envahit. Nous avons accès à beaucoup plus de produits que l'on ne peut en connaître. Ces produits peuvent être de différentes sortes : de la musique, des films, de la technologie, les transports, le commerce, le travail, l'école, etc. L'internet vient en aide aux utilisateurs et facilite énormément le référencement, la recherche et l'accès aux informations.

Les moteurs de prédictions ont été crées pour fournir à l'utilisateur des alternatives de produits et, bien sûr, pour des raisons commerciales. Les gens aiment généralement consulter les recommandations d'autres utilisateurs avant de se faire leurs propres opinions. Pour cette raison les prédictions en ligne sont devenues très utiles pour les clients. Les algorithmes des moteurs de prédiction sont basés sur l'expérience et l'avis des autres utilisateurs. Ces algorithmes sont basés sur les correspondances entre les produits et entre les utilisateurs. Les algorithmes donnent des résultats très intéressants dans le cas où ils arrivent à trouver des autres utilisateurs qui ont des goûts très similaires. Dans ce cas ils fournissent des alternatives de produits qui sont intéressantes pour un utilisateur. Mais pour trouver des utilisateurs *"co-frères"* les moteurs de prédiction ont besoin d'avoir une très grande base de profils d'utilisateurs. Le sujet de cette thèse tente de répondre à ces besoins.

Le domaine de recherche présenté dans cette thèse est *l'Analyse des Sentiments*. Le but générales est de génèrer des profils d'utilisateurs pour qu'ils puissent être utilisés par les algorithmes prédictif. Les profils concernent les films cinématographiques, et le but général est de créer un système autonome qui servira de support pour les moteurs de prédiction. Le rôle d'un tel système est de rechercher des critiques, d'effectuer une notation des sentiments automatiquement, et la création des profils finaux. L'activité de recherche concerne la notation de l'opinion, car c'est une tâche très ambigüe.

Dans la littérature, l'analyse des sentiments est connue sur le nom de *Opinion Mining* et elle est récemment devenu un domaine en plein développement en raison de ses nombreuses applications. Mais à part le support du moteur prédicatif nous pouvons citer des nombreuses utilisations comme : la recommandation (par exemple des voitures), l'explication des sondages des suffrages aux élections, la consultation des avis sur les produits, la détection de spam, l'analyse et la surveillance des opinions pour améliorer les produits (matériels ou intellectuels) ou l'étude de marché.

Dans cette thèse nous présentons un système entier pour la création de profils. Concernant la notation des sentiments nous présentons plusieurs approches pour attribuer une note mettant en relief l'intensité et la polarité de l'opinion de la critique

cinématographique. Nous trouvons généralement dans la littérature le traitement strictement statistique. Nous présentons deux approches basées sur le traitement linguistique. Nous comparons nos résultats avec l'approche statistique, en montrant l'intérêt d'une analyse linguistique profonde. Nous montrerons que l'architecture de la partie de notation de l'opinion est le résultat de nombreuses études, tests et améliorations apportées par les travaux de cette thèse.

L'objectif de cette thèse est de fournir un système autonome pour la recherche et notation de l'opinion décrite dans les critiques cinématographiques dans le but de générer des profiles pour les moteurs de prédictions. Pour atteindre cet objectif, nous avons procédé par étapes. Nous avons commencé par étudier les problématiques liées à la notation de l'opinion et les solutions existantes. A l'issue de cette étude, nous avons proposé des approches basées sur le traitement linguistique approfondi afin d'améliorer l'attribution d'une note à l'intensité de l'opinion. Les contributions de cette thèse concernent principalement les points suivants :

- la création du module de la récupération des critiques cinématographiques de l'Internet ;
- l'implémentation d'une nouvelle approche (le classificateur de comportement des groupes) ;
- l'implémentation d'une nouvelle approche (le classificateur linguistique) ;
- la comparaison des approches avec une approche statistique (le classificateur de Bayes),
- la combinaison des résultats de toutes ces approches pour améliorer la notation finale de la critique cinématographique ;
- la création d'une application rassemblant tous les modules du système proposé.

1.2 Organisation du rapport

Au *Chapitre 2* nous présenterons la technique de catégorisation de texte. Nous nous concenterons sur les techniques de classification de texte et sur l'apprentissage automatique. Ces techniques sont utilisées dans le système développé. Nous décrirons aussi la représentation du corpus documentaire et donc le prétraitement, l'indexation et la vectorisation du corpus. Nous finirons ce chapitre en montrent les mesures générales de

performance pour analyser et tester des méthodes de classification du texte.

Au *Chapitre 3* nous décrirons les techniques présentés dans le *Chapitre 1* pour l'analyse des sentiments. Nous montrerons différentes approches existantes et leur utiliation. Nous expliquerons les raisons pour lesquelles l'analyse de l'opinion est si complexe, et nous montrerons les recherches sur la subjectivité ou objectivité des phrases, sur le calcul de la polarité ou de l'intensité de l'opinion. Nous constaterons que la majorité des recherches dans le domaine de l'*Opinion Mining* concerne le traitement statistique.

Pour cette raison nous présenterons dans le *Chapitre 4* le traitement linguistique. Nous mettrons plus d'attention à décrire les ressources linguistiques que nous utilisons dans notre recherche via une application existante. Nous préciserons le rôle de l'ambiguïté dans la recherche linguistique, et nous présenterons les solutions existantes de l'analyse du texte. Malgré le fait que nous n'ayons pas retrouvé d'utilisation de cette ressource pour la notation et la détection des sentiments dans la littérature, nous émettrons l'hypothèse que c'est une voie importante dans le domaine.

Dans le *Chapitre 5* nous présenterons la fonctionnalité et l'architecture du système développé - système autonome pour la détection et notation automatiquement de la critique cinématographique. Nous décrirons la partie de recherche des critiques, nous donnerons et expliquerons l'architecture globale de la partie de la notation de l'opinion, et nous décrivons la partie de publication des résultats. Dans ce chapitre nous ne rentrerons pas dans les détails de la classification des sentiments.

Dans le *Chapitre 6*, nous décrirons quatre classificateurs implémentés durant notre recherche. Nous présenterons trois classificateurs de l'opinion et un classificateur final pour combiner les notes récupérées. Dans ce chapitre nous présenterons également en détail la chaîne de traitement nécessaire à la classification.

Dans le *Chapitre 7* nous présenterons les tests sur chaque classificateur de la notation de l'opinion, nous comparerons ces classificateurs en mettant en évidence les inconvénients et les avantages de chacun. Nous présenterons les résultats obtenus sur

chaque classification sur la même base d'apprentissage, et le même ensemble des critiques testées. Nous présenterons ensuite l'amélioration de nos résultats en appliquant la classification finale.

Enfin, nous conclurons cette thèse dans le *Chapitre 8* par un bilan des résultats obtenus, et par une présentation des perspectives de recherches qu'ouvrent nos travaux.

Chapitre 2

Le traitement du corpus documentaire par les approches statistiques

2.1 De la Recherche d'Information à l'Analyse des Sentiments

Dans les dix dernières années les tâches de gestion basées sur le contenu de documents (collectivement connu sous le nom de "Recherche d'Information" - *ang : Information Retrieval* - IR) ont acquis un statut important dans le domaine des systèmes d'information, en raison de l'augmentation de la disponibilité des documents sous forme numérique et de la nécessité d'y accéder en souplesse.

La Catégorisation de Texte (*ang : Text Categorization* - TC), l'activité de l'étiquetage des textes en langage naturel avec des catégories de thématiques prédéfinies, est une de ces tâches. Celle-ci remonte au début des années 60, mais elle n'est devenue l'un des principaux sous-domaines de la discipline des systèmes d'information qu'au début des années 90, grâce à un intérêt accru et à la disponibilité de matériels plus puissants. La catégorisation de texte est actuellement appliquée dans de nombreux et différents contextes : l'indexation de documents basée sur un lexique, le filtrage de documents, la génération automatique de métadonnées, la suppression de l'ambigüité du sens des mots, le peuplement des catalogues hiérarchique de ressources Web, et en général toutes les

applications nécessitant l'organisation de documents ou le traitement sélectif et l'adaptation de documents [Sebastiani (2002)].

Actuellement la "TC" est un domaine entre l'Apprentissage Automatique (*ang : Machine Learning* - ML) et la Recherche d'Information (IR). Elle partage un certain nombre de caractéristiques avec d'autres tâches telles que l'extraction de connaissances à partir de textes et la Fouille de Textes (*ang : Texte Mining*) [Knight (1999), Pazienza (1997)]. La "ML" décrit un processus inductif général qui construit automatiquement un classificateur de texte par l'apprentissage, à partir d'une série des documents préclassifiés ou de caractéristiques de catégories d'intérêts. La Fouille de Textes est un ensemble de traitements informatiques consistant à extraire des connaissances selon des critères de nouveauté ou de similarité dans des textes produits par des humains pour des humains [Joachims & Sebastiani (2002), Lewis & Haues (1994)].

Un domaine utilisant les techniques de IR, TC, ML ou Fouille de Texte est notamment le domaine de l'Analyse des Sentiments, connu sur le nom de (*ang : Opinion Mining*). La recherche dans ce domaine couvre plusieurs sujets, notamment l'apprentissage de l'orientation sémantique des mots ou des expressions, l'analyse sentimentale de documents et l'analyse des opinions et attitudes à l'égard de certains sujets ou produits.

2.2 La Catégorisation de Texte

La Catégorisation de Textes consiste en l'attribution d'une valeur booléenne à chaque paire $< d_j, c_i > \in D \times C$ où D est un domaine des documents et

$$C = c_1, ..., c_{|C|}$$

est un ensemble de catégories prédéfinies. Une valeur de T attribuée à la paire $< d_j, c_i >$ indique une décision de déposer d_j sous c_i, et une valeur de F indique une décision de ne pas déposer d_j sous c_i. Plus formellement, la tâche consiste à approximer une fonction inconnue d'une cible

$$\overline{\phi} : D \times C \rightarrow \{T, F\}$$

(qui décrit la façon dont les documents doivent être classifiés) par le biais d'une fonction

$$\phi : D \times C \rightarrow \{T, F\}$$

appelée le classificateur de telle sorte que $\overline{\phi}$ et ϕ coïncident autant que possible [Sebastiani (2002)].

En se fondant uniquement sur le caractère endogène des connaissances pour le classement d'un document fondé uniquement sur sa sémantique, et compte tenu du fait que la sémantique d'un document est une notion subjective, il s'ensuit que l'adhésion d'un document à une catégorie [Saracevic (1975)] ne peut être décidée de manière déterministe. Ceci est illustré par le phénomène d'inter-indexeur d'incohérence [Cleverdon (1984)] : lorsque deux des experts humains décident de classer un document d_j dans une catégorie c_i, il peut y avoir avoir désaccord ; ce qui se passe en fait fréquemment.

La Catégorisation de Texte a été utilisée dans un certain nombre d'applications différentes. Les premières applications concernées étaient l'indexation automatique pour les systèmes de Recherche d'Information (IR) booléens. Les premières recherches dans le domaine ont été effectuées par Borko et Bernick [Borko & Bernick (1963)], Gray et Harley [Gray & Harley (1971)], Heaps [Heaps (1973), Maron [Maron (1961)]. A chaque document est attribué un ou plusieurs mots ou expressions clés décrivant son contenu, ces mots et expressions clés appartiennent à un ensemble fini appelé dictionnaire contrôlé, souvent composé d'un thesaurus thématique hiérarchique (par exemple, le thesaurus de NASA pour la discipline aéronautique, ou le thésaurus de MESH pour la médecine) [Sebastiani (2002)]. Habituellement, cette attribution est effectuée par des indexeurs manuels, et c'est donc une activité coûteuse. Divers classificateurs de texte explicitement conçus pour l'indexation de documents ont été décrit dans la littérature, par exemple : Fuhr et Knorz [Fuhr & Knorz (1984)], Robertson et Harding [Robertson & Harding (1984)], et Tzeras et Hartmann [Tzeras & Hartmann (1993)].

L'indexation automatique utilisant les dictionnaires est étroitement liée à la génération automatique de métadonnées. Dans les bibliothèques numériques, nous sommes souvent plus intéressés par le marquage des documents par des métadonnées qui les décrivent sous différents aspects (par exemple, date de création, type de document ou le format, disponibilité, etc.). Le rôle de certaines de ces métadonnées est de décrire la

sémantique du document de la signification des codes bibliographiques, des mots-clés ou des phrases-clés.

L'indexation avec un vocabulaire contrôlé est un exemple de la problématique générale d'organisation du document. Le plus souvent, de nombreux autres problèmes relatifs à l'organisation et au classement du document, que ce soit pour des organisations personnelles ou la structuration d'un document de base d'entreprise, peuvent être réglées par les techniques de TC. Dans les bureaux d'un journal, par exemple, les annonces doivent être classées dans les catégories telles que les rencontres, voitures à vendre, immobilier, etc. avant les publications. Les journaux avec un grand nombre d'annonces bénéficieraient d'un système automatique qui pourrait choisir pour une annonce la catégorie donnée la plus appropriée. D'autres applications possibles sont les applications d'organisation des brevets en catégories pour rendre leur recherche plus facile [Larkey (1999)], le classement automatique des articles de journaux sous les sections appropriées (par exemple, la politique, événements, styles de vie, etc.), ou le regroupement automatique en sessions des papiers de conférence [Sebastiani (2002)].

Une autre application des techniques de TC est le Filtrage de Textes (*ang : Text Filtering* - TF). Le Filtrage de Textes est l'activité de classification d'un flux de documents expédiés de manière asynchrone par un producteur d'information à destination d'un consommateur d'information [Belkin & Croft (1992)]. Un cas typique est une situation dans laquelle le producteur est une agence de presse et le consommateur est un journal [Hayes *et al.* (1990)]. Dans ce cas, le système de filtrage doit empêcher la livraison de documents qui n'intéressent pas le consommateur. Le filtrage peut être considéré comme un cas de TC de l'étiquetage, c'est la classification des documents en deux catégories disjointes, la catégorie "pertinents" et la catégorie "non pertinents". En outre, un système de filtrage peut également classer les documents jugés pertinents pour le consommateur en catégories thématiques, en classant par exemple à part les articles de sport pour un journal de sport. Tous les articles de sports devraient être classés en fonction du sport qu'ils traitent, de manière à permettre aux journalistes spécialisés dans des sports individuels d'accéder uniquement aux documents les concernant. De même, un système de filtrage des mails peut filtrer les spam ainsi que classer les messages dans des catégories thématiques pour l'utilisateur [Androutsopoulos *et al.* (2000), Drucker *et al.*

(1999)]. Un système de filtrage peut être installé chez le producteur d'information, dans ce cas il doit envoyer les documents seulement à des consommateurs intéressés, ou chez tous les consommateurs. Dans ce cas il doit bloquer la livraison de documents jugés sans intérêt pour le consommateur. Dans le premier cas, le système construit et met à jour un "profil" pour chaque consommateur [Liddy *et al.* (1994)], alors que dans le dernier cas un seul profil est nécessaire. Le filtrage d'information en utilisant les techniques de ML est largement débattu dans la littérature : Amati et Crestani [Amati & Crestani (1999)], Iyer et al. [Iyer *et al.* (2000)], Kim et al. [Kim *et al.* (2000)], Tauritz et al. [Tauritz *et al.* (2000)], et Yu et Lam [Yu & Lam (1998)].

Les techniques de TC permettent également de lever l'ambigüité sur le sens des mots (*ang : Word Sense Disambiguation* - WSD). La WSD est l'activité de recherche dans un texte des sens des mots ambigus. Un seul mot peut avoir plusieurs significations. La tâche du système WSD est donc de décider de quel des sens il s'agit. La WSD est très importante pour de nombreuses applications, y compris le traitement du langage naturel et l'indexation des documents par le sens des mots. La WSD peut être considérée comme une tâche de TC [Gale *et al.* (1993), Escudero *et al.* (2000)] si nous considérons le contexte d'occurrence des mots comme un document et le sens du mot comme une catégorie. La WSD est juste un exemple du problème plus général consistant à lever les ambiguïtés du langage naturel, un des problèmes les plus importants en linguistique computationnelle.

Parmi d'autres applications qui sont basées sur les techniques de TC nous pouvons citer la catégorisation des discours par combinaison de la reconnaissance de la parole [Myers *et al.* (2000), Schapire & Singer (2000)], la catégorisation de documents multimédias à travers l'analyse de légendes [Sable & Hatzivassiloglou (2000)], l'identification d'auteur de textes littéraires d'auteur inconnu [Forsyth (1999)], l'identification de la langue pour les textes de langue inconnue [Cavnar & Trenkle (1994)], l'identification automatique du genre du texte [Kessler *et al.* (1997)], le classement automatisé des essais [Larkey (1998)] et la catégorisation hiérarchique des pages Web [Attardi *et al.* (1998), Furnkranz (1999), Oh *et al.* (2000), Yang *et al.* (2002)].

2.3 L'Apprentissage Automatique

Dans les années 80, l'approche la plus populaire pour la création des classificateurs automatique de documents a consisté à construire manuellement un système expert capable de prendre des décisions de TC. Un tel système d'expertise était composé généralement d'un ensemble de règles logiques définies manuellement par une catégorie, du type,

```
if <DNF formula> then <category>
```

DNF (*ang : disjonctive forme normale*) est une disjonction de propositions conjonctives. Le document est classé dans la <category> si et seulement s'il est en accord avec la formule, donc s'il est en accord avec au moins une des propositions. L'exemple le plus connu de cette approche est le système CONSTRUE [Hayes *et al.* (1990)], construit par le Carnegie Groupe pour l'agence de presse Reuter. L'inconvénient de cette approche est que les règles doivent être définies manuellement par un ingénieur des connaissances à l'aide d'un expert du domaine. Si l'ensemble des catégories est mis à jour, ces deux professionnels doivent intervenir à nouveau, et si le classificateur est adapté à un tout autre domaine (c'est-à-dire, ensemble de catégories), des experts d'un domaine différent doivent intervenir et le travail doit être repris à partir de zéro.

Depuis le début des années 90, l'approche de ML pour le besoin de TC a gagné en popularité et a fini par devenir l'approche dominante [Mitchell (1996)]. Dans cette approche, un processus inductif (également appelé l'apprentissage) construit automatiquement un classificateur pour une catégorie c_i en observant les caractéristiques d'un ensemble de documents classés manuellement pour c_i ou $\overline{c}_i$ par un expert du domaine. De ces caractéristiques le processus inductif tire les caractéristiques que doit avoir le nouveau document pour être classé dans la catégorie c_i. Les avantages de l'approche de ML sont évidents. L'effort d'ingénierie va à la construction, non pas d'un classificateur, mais d'un constructeur automatique de classificateurs (l'apprenant).

Cela signifie que tout ce qui est nécessaire est que l'apprenant subisse une construction inductive et automatique d'un classificateur à partir d'une série de documents classifiés manuellement. Dans ce cas, nous n'avons plus besoin de traiter à nouveau des classificateurs qui existent déjà et la série initiale des catégories est mise à jour si le

classificateur est porté à un tout autre domaine pour définir les règles manuellement.

Dans l'approche de ML, les documents pré-classifiés sont alors les ressources clés. Dans le cas le plus favorable, ils sont déjà disponibles, ce qui se passe généralement pour les organisations qui ont déjà effectué la catégorisation manuellement de même activité et ont décidé d'automatiser le processus. Les cas le moins favorable est le cas où les documents classés manuellement ne sont pas disponibles, ce qui se passe généralement pour les organisations qui commencent une activité de catégorisation et optent pour un mode automatique. L'approche ML est plus pratique également dans ce dernier cas. Il est en fait plus facile de classer manuellement un ensemble de documents que de construire et de modifier un ensemble de règles, car il est plus facile de caractériser des cas de "celui-ci" que de décrire ce concept en mots, ou de décrire une procédure de reconnaissance des cas.

L'approche de ML repose sur la disponibilité d'un corpus initial

$$\Omega = d_1, ..., d_{|\Omega|} \subset D$$

de documents pré-classifiés sous

$$C = c_1, ..., c_{|C|}.$$

En d'autres termes, les valeurs de la fonction $\breve{\Phi} : D \times C \rightarrow T, F$ sont connues pour chaque paire $< d_j, c_i > \in \Omega \times C$. Un document d_j est un exemple positif de c_i si

$$\breve{\Phi}(d_j, c_i) = T,$$

un exemple négatif de c_i si

$$\breve{\Phi}(d_j, c_i) = F.$$

Dans les paramètres de recherche, une fois qu'un classificateur Φ a été construit il est souhaitable d'évaluer son efficacité. Dans ce cas, avant la construction du classificateur, le corpus initial est divisé en deux séries, pas nécessairement de taille égale : un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage est :

$$EA = \{d1, ..., d_{|EA|}\}.$$

2. LE TRAITEMENT DU CORPUS DOCUMENTAIRE PAR LES APPROCHES STATISTIQUES

Le classificateur Φ pour les catégories $C = \{c_1, ..., c_{|C|}\}$ est construit en observant les caractéristiques de ces documents. L'ensemble de test

$$ET = \{d_{|EA|+1}, ..., d_{|\Omega|}\}$$

est utilisé pour tester l'efficacité des classificateurs. Chaque $d_j \in ET$ est donné au classificateur, et les décisions du classificateur $\Phi(d_j, c_i)$ sont comparées avec les décisions d'expert $\breve{\Phi}(d_j, c_i)$. Une mesure d'efficacité de la classification est basée sur la fréquence des valeurs de $\Phi(d_j, c_i)$ correspondant aux valeurs de $\breve{\Phi}(d_j, c_i)$.
Les documents de ET ne peuvent pas participer d'une façon quelconque à la construction d'induction du classement. Si cette condition n'était satisfaite, les résultats expérimentaux obtenus seraient probablement trop bons, et l'évaluation n'aurait donc pas de caractère scientifique [Mitchell (1996)]. La validation est une phase indispensable à tout processus d'apprentissage. Elle consiste à vérifier que le modèle construit sur l'ensemble d'apprentissage permet de classer tout individu avec le minimum d'erreurs possible. Nous citerons trois méthodes de validation généralement utilisées :

- validation par le test,
- validation croisée,
- validation "bootstrap".

Dans le cas de la validation par le test, les résultats de l'évaluation précédente seraient une estimation pessimiste de la performance réelle, la dernière classification ayant été formée sur plus de données que le classificateur évalué. L'ensemble d'apprentissage permet de générer le modèle, l'ensemble de test permet d'évaluer l'erreur réelle du modèle sur un ensemble indépendant évitant ainsi un biais d'apprentissage. S'il s'agit de tester plusieurs modèles et de les comparer, nous pouvons sélectionner le meilleur modèle selon ses performances sur l'ensemble de validation et ensuite évaluer l'erreur réelle sur l'ensemble de test [Figure 2.1].

Une alternative est la validation croisée [Mitchell (1996)], dans laquelle k différents classificateurs $\Phi_1, ..., \Phi_k$ sont construits par le partitionnement initial du corpus en k ensembles disjoints $ET_1, ..., ET_k$ et la validation par test est ensuite appliquée de façon itérative sur les paires $EA_i = \Omega - ET_i, ET_i$. L'efficacité finale est obtenue par le calcul individuel de l'efficacité de $\Phi_1, ...\Phi_k$. La validation croisée ne construit pas de modèle

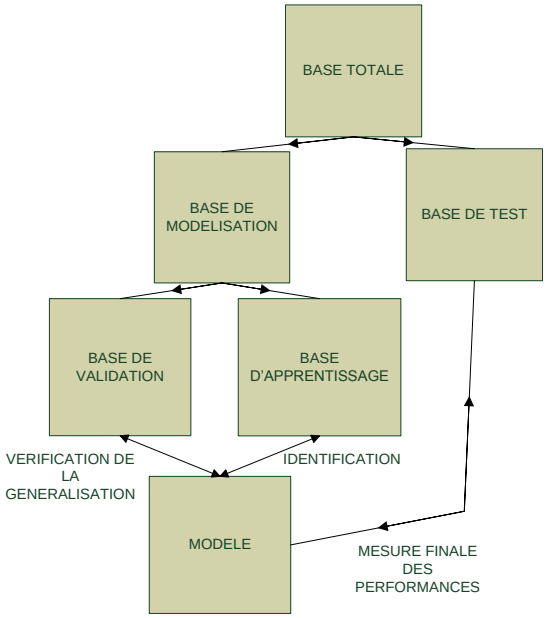


Figure 2.1: Processus de validation par le test -

utilisable, elle estime juste l'erreur réelle [Figure 2.2]. En général le nombre k de parties est fixé a 10. L'algorithme est le suivant :

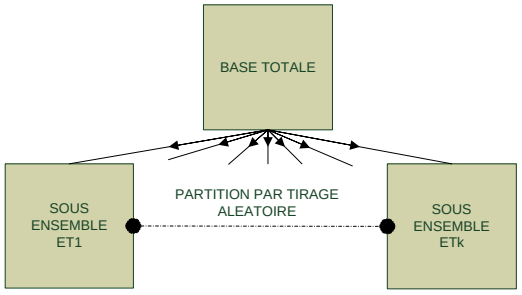


Figure 2.2: Processus de validation croisée -

```
ET est un ensemble, k un entier
Découper ET en k parties égales ET1,...ETk
Pour i de 1 a k
  Construire un modèle M avec l'ensemble ET-ETi
  Evaluer une mesure d'erreur ei de M avec ETi
Fin pour
Retourner l'espérance mathématique de la mesure des erreurs
```

Un autre moyen est l'utilisation de la "validation bootstrap". Etant donné un échantillon S de taille n, nous tirons avec remise un ensemble d'apprentissage de taille n (un élément de S peut ne pas appartenir a l'ensemble d'apprentissage, ou y figurer plusieurs fois), l'ensemble de test est S [Figure 2.3].

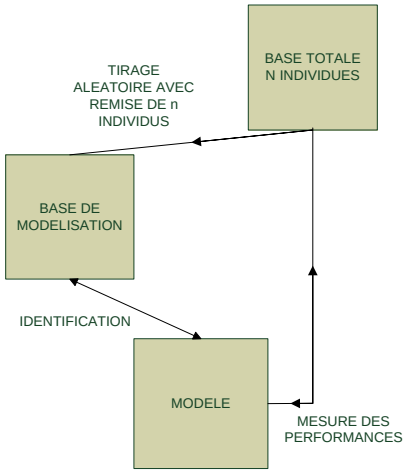


Figure 2.3: Processus de Bootstrap -

Ces deux dernieres méthodes fournissent de bons estimateurs de l'erreur réelle mais sont très couteuses en temps de calcul. Elles sont utiles pour les petits échantillons. Dans ces approches, il arrive souvent que les paramètres internes des classificateurs doivent être réglés par des essais, pour obtenir les valeurs des paramètres rendant la meilleure efficacité. Afin de rendre possible cette optimisation, dans la validation par test, l'ensemble $\{d_1, ..., d_{|EA|}\}$ est en outre divisé en un sous ensemble d'apprentissage $EP = \{d_1, ...d_{|EP|}\}$ à partir duquel le classificateur est construit, et un sous ensemble de validation $EV = \{d_{|EP|+1}, ..., d_{|EA|}\}$, sur lequel la répétition des tests du classificateur pour l'optimisation des paramètres est effectuée. La variante peut être utilisée dans la validation croisée. Evidement, pour la même raison que nous n'avons pas testé un classificateur sur les documents sur lesquels il a été formé, nous ne pouvons pas le tester sur les documents sur lesquels il a été optimisé. L'ensemble de test et de validation doivent être séparés.

2.4 Représentation des corpus documentaires

2.4.1 L'unité linguistique

Les textes en langage naturel ne peuvent pas être directement interprétés par un classificateur ou par les algorithmes de classification. Le sens d'un document peut être porté par un ensemble d'unités linguistiques particulières, aux caractéristiques plus ou moins élaborées issues de l'analyse du corpus documentaire. Les premières unités linguistiques qui représentent du sens sont les lemmes des mots. La reconnaissance de ces unités linguistiques nécessite d'effectuer un prétraitement linguistique des mots du texte. L'unité linguistique peut être représentée par le mot ou la phrase. Dans le premier cas l'unité linguistique est le mot tel qu'il apparait dans le document. Chaque mot est extrait du texte en considérant des séparateurs comme l'espace, la tabulation, et la ponctuation. Le nombre de mots caractérisant un corpus de documents peut être très grand, il est donc nécessaire de conserver un sous ensemble de ces mots. Ce filtrage repose à la base sur les fréquences d'occurrences des mots dans le corpus.

D'autres approches utilisent non pas des mots, mais des groupes de mots, voir des phrases comme l'unité linguistique décrivant le sens. Ce type d'unité linguistique décrit le sens plus complètement qu'un simple mot [Lewis (1992b)]. De plus, grâce à cette approche nous avons une relation d'ordre entre les mots, les co-occurrences de mots. L'inconvénient est que la fréquence d'apparition des groupes de mots ne permet pas d'offrir des statistiques fiables, car le grand nombre de combinaisons entre les mots engendre des fréquences trop faibles pour être exploitables.

Une autre approche pour représenter le corpus documentaires est l'utilisation de la technique des n-grammes [Shannon (1948)]. En général le n-gramme se définit comme étant une séquence de n caractères consécutifs. Le principe des n-gramme est que pour une chaine de k caractères entourée de blancs, nous génèrons $k+1$ n-grammes [Cavnar & Trenkle (1994)]. Un exemple de découpage pour le mot *porte* en bi-gramme ($n = 2$) est le suivant :

```
_porte_ : _p, po, or, rt, te, e_
```

Dès qu'on extrait tous les n-gramme d'un document on défini la liste des n-grammes triés par ordre décroissant de leur fréquence d'apparition. Ces méthodes sont indépendantes de la langue et ni la segmentation en unités linguistiques, ni des prétraitements comme le filtrage et la lemmatisation ne sont nécessaires.

2.4.2 Prétraitement du texte

Si nous utilisons de mots comme unité linguistique nous remarquons que plusieurs mots ont des sens communs ou forment simplement une autre forme de conjugaison. Attribuer un sens différent à ces mots relèverait d'une redondance sans pertinence sémantique. Pour cette raison un traitement appelé stemmatisation (*ang : stemming*) est à effectuer. Elle consiste à retrouver la racine de chaque mot. C'est un traitement qui procède à une analyse morphologique du texte [Porter (1980)]. Ce traitement est basé sur un dictionnaire de suffixes qui permet d'extraire le radical du mot grâce à l'étude morphologique des mots.

Le traitement qui demande une analyse plus complexe que la stemmatisation est la lemmatisation qui est fondée sur un lexique. Un lexique est un ensemble de lemmes avec lequel nous pouvons faire référence au dictionnaire. L'objectif de la lemmatisation est d'associer à chaque mot une entrée dans le lexique. Comme de nombreux mots de même graphie peuvent provenir de différents lemmes, l'analyse morphologique est insuffisante. La lemmatisation nécessite donc de réaliser en plus une analyse syntaxique pour résoudre les ambigüités, elle effectue donc une analyse morphosyntaxique [Schmid (1994)].

Avant d'effectuer un des prétraitements précédents, il est usuel d'utiliser une "stop-list" pour éliminer tout les mots qui ne participent pas activement au sens du document. Elle contient les pronoms, les articles et les mots trop fréquents pour être discriminants, nous éliminons donc toutes les unités linguistiques non discriminantes. Le risque de la stop-list est que l'on peut éliminer des mots qui pourraient être utiles pour la classification.

2.4.3 L'indexation des documents et la réduction de dimension

Une procédure d'indexation du texte d_j en une représentation compacte de son contenu doit être appliquée uniformément aux documents d'apprentissage, de validation et de tests. Le choix d'une représentation du texte dépend de ce que l'on considère comme étant les unités linguistique du sens du texte (le problème de sémantique lexicale). Les approches de l'indexation sont partagées en :

- celles qui étudient les différents moyens de comprendre ce qu'est une unité linguistique,
- celles qui se basent sur différentes manières de calculer les poids des unités.

Aptťe et al. [Apt'e *et al.* (1994)], Dumais et al. [Dumais *et al.* (1998)], Lewis [Lewis (1992a)] ont démontré que l'utilisation comme unité linguistique de représentations plus sophistiquée que le mot ne donne pas des résultats beaucoup plus fiables. Lewis [Lewis (1992a)] argumente que la raison de ces résultats est probablement que, bien que l'indexation basée sur les phrases montre une sémantique de qualité supérieure, le traitement statistique est moins intéressent que l'indexation basée sur les mots. Car l'indexation des mots composés a plus d'unités, plus de synonymes, une plus faible cohérence de la correspondance (comme les synonymes ne sont pas affectés aux mêmes documents), et une fréquence inférieure d'unités par document [Lewis (1992a)]. L'amélioration des résultats est possible en faisant la combinaison des ces deux approches.

Le poids des unités varie le plus souvent entre 0 et 1. Dans un cas spécial le poids peut être binaire (1 indique la présence et 0 l'absence du terme dans le document). Dans le cas d'indexation non binaire, pour déterminer le poids de w_{kj} de l'unité t_k dans le document d_j, toutes les techniques d'indexation de IR qui représentent un document comme un vecteur de termes pondérés peuvent être utilisées. La plupart du temps, la fonction *tfidf* est utilisée [Salton & Buckley (1988)], définie comme suit :

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot log\frac{|T_r|}{\#_{T_r}(t_k)} \tag{2.1}$$

où $\#(t_k, d_j)$ désigne le nombre de fois que t_k se produit en d_j, et $\#_{T_r}(t_k)$ indique la fréquence du document de terme t_k - c'est à dire : le nombre de documents d'ensemble d'apprentissage (EP) dans lesquels t_k se produit. Cette fonction montre que, plus un terme apparait souvent dans un document, plus il est représentatif ; plus le nombre

de documents contenant un terme est important moins ce terme est discriminatoire. La sémantique d'un document est réduite à la sémantique lexicale collective des unités lexicales, sans tenir compte de la sémantique compositionnelle. Pour que le poids prenne des valeurs dans l'intervalle [0,1] le poids résultant de tfidf est souvent normalisé :

$$w_{k,j} = \frac{tfidf(t_k, d_j)}{\sqrt{\Sigma_{s=1}^{|T|}(tfidf(t_s, d_j))^2}} \tag{2.2}$$

Le rôle de la représentation textuelle est représenté mathématiquement de façon à ce que l'on puisse effectuer le traitement analytique, tout en conservant au maximum la sémantique. La représentation mathématique généralement utilisée est l'utilisation d'un espace vectoriel comme espace de représentation cible. La caractéristique principale de la représentation vectorielle est que chaque unité linguistique est associée à une dimension propre au sein de l'espace vectoriel. Deux textes utilisant les mêmes segments textuels seront donc projetés sur des vecteurs identiques. Le formalisme le plus utilisé pour représenter les textes est le formalisme vectoriel [Salton *et al.* (1975), Salton (1983)]. Le texte d_j est représenté comme vecteur des unités linguistique $\vec{d_j} = [w_{1,j}, ..., w_{|T|,j}]$, où T est l'ensemble des unités qui se trouvent au moins une fois dans au moins un document d'apprentissage (EP), et $0 \leq w_{k,j} \leq 1$ représente combien de fois une unité est présente dans le document d_j. Dans cette approche, chaque dimension de l'espace vectoriel correspond à un élément textuel, nommé terme d'indexation, préalablement extrait par calcul selon plusieurs méthodes proposées par Salton [Salton (1983), Salton & Buckley (1988)]. Cette action de sélection fait appel à un premier processus d'indexation. Salton a proposé plusieurs versions du processus de prétraitement des unités linguistiques décrites précédemment. Dans un premier temps ce traitement était une stemmatisation. Ensuite, avec l'évolution des techniques informatiques, de nombreuses solutions sont apparues pour l'analyse morphosyntaxique des textes. Le processus d'indexation lui-même consiste à effectuer un simple inventaire complet de tous les lemmes du corpus. Ensuite, vient le processus de sélection des lemmes qui vont constituer les unités linguistiques du domaine ou les dimensions de l'espace vectoriel de représentation du corpus documentaire. Dans ce formalisme, chaque dimension de l'espace vectoriel correspond à un segment textuel, préalablement extrait du corpus documentaire. Une dimension est un mot, un paragraphe, une lettre ou un assemblage de lettres. Les unités linguistiques sont choisies dans l'ensemble des lemmes en fonction de leur pouvoir de discrimination.

Elles constituent les termes d'indexation. La sélection des termes d'indexation correspond donc à une réduction de la dimension de l'espace effectuée en fonction de critères de discrimination. Le critère de sélection des termes d'indexation le plus utilisé est la fréquence en documents IDF (*ang : Inverse Document Frequency*). Il consiste à calculer le nombre de documents dans lesquels apparaît un lemme puis de prendre l'inverse de ce nombre. Pour une collection de documents D, la sélection des unités linguistiques qui ont une fréquence en documents entre $D/100$ et $D/10$ génère le plus souvent un ensemble d'unités linguistiques ayant un pouvoir de discrimination satisfaisant pour la recherche documentaire [Salton *et al.* (1975)]. Un lemme l_i constitue une dimension de l'espace vectoriel de représentation du corpus de textes. Un texte ou document D est représenté par un vecteur $D = (ft_1..., ft_i, ..., ft_{|V|})$ dans lequel V représente le vocabulaire ou l'ensemble des unités linguistiques sélectionnées et ft_i le nombre d'occurrences du lemme l_i dans le document D.

Une variante du modèle vectoriel standard est le modèle LSI (*ang : Latent Semantic Indexing*) qui prend en compte la structure sémantique des unités linguistiques [Deerwester *et al.* (1990)]. LSI utilise la matrice du modèle vectoriel standard, dans laquelle chaque élément x_{ij}, où j - est le document et i - l'unité linguistique, est le nombre d'occurrences de l'unité linguistique u_i dans le document D_j. Une décomposition en valeurs singulières SVD (*ang : Singular Value Decomposition*) de cette matrice est effectuée et seuls les premiers vecteurs propres sont pris en compte.
Nous appelons M la matrice où l'élément (i,j) décrit l'occurrence d'unités linguistiques i dans le document j. La ligne de cette matrice représente le vecteur correspondant à l'unité linguistique en précisant sa relation avec chaque document :

$$t_i^T = [x_{i,1}, ..., x_{i,n}].$$

La colonne de cette matrice représente le vecteur correspondent au document en précisant sa relation avec chaque unité linguistique :

$$d_j = \begin{vmatrix} x_{1,j} \\ \vdots \\ x_{m,j} \end{vmatrix}$$

Le produit scalaire $t_i^T t_p$ entre deux vecteurs de l'unité linguistique montre la corrélation entre les unités dans les documents. Le produit scalaire des matrices MM^T contient tous ces produits scalaires des unités linguistiques. La décomposition de la matrice M est répresentée par les matrices U et V - les matrices orthogonales et par la matrice Σ - la matrice diagonale est de la forme :

$$M = U\Sigma V^T$$

Nous approximons la matrice Σ par la matrice réduite $\widehat{\Sigma}$, pour représenter les documents dans un espace réduit de dimension. Chacune des dimensions de l'espace de représentation final correspond à une combinaison linéaire des unités linguistiques. L'espace de représentation n'a donc pas pour support un ensemble de termes d'indexation, ce qui rend les dimensions relativement difficiles à interpréter directement. Ce modèle permet de représenter les termes par des vecteurs qui sont une indication du profil d'occurrence du terme dans les documents. Cette propriété peut donc être utilisée pour établir une notion de similarité entre termes, ou représenter un document comme la moyenne des vecteurs représentant les termes qu'il contient. LSI a été utilisé dans plusieurs travaux de recherche comme : Hull [Hull (1994)], Shutze [Schutze (1998)], Weigend et al. [Weigend *et al.* (1999)], et Yang [Yang (1995)].

Un autre modèle a été proposée par Besançon [Besancon & Rajman (2000)] - le modèle DSIR. Dans cette approche les documents sont représentés sous forme de vecteurs obtenus par un calcul de cooccurrence entre les termes d'indexation d'un corpus documentaire. Une matrice de cooccurrence M est calculée en prenant en compte l'apparition conjointe de 2 termes présents dans la même phrase. Les vecteurs directeurs de Besançon sont la somme pondérée de 2 vecteurs : $\alpha V + (1 - \alpha)M.V$, où V est un vecteur mot-clé comme proposé par Salton, et $M.V$ est le produit du vecteur V avec la matrice de cooccurrences M construite sur le corpus. Le résultat est une représentation des documents sous forme de vecteurs qui portent à la fois la représentation statistique des mots clés et la contribution des autres termes du corpus pour chaque mot clé. La représentation de chaque texte exprime à la fois des fréquences d'apparition de mots clés mais également les contributions des liens de cooccurrence du corpus documentaire dont le texte fait partie. Les termes d'indexation peuvent être choisis dans l'ensemble de

toutes les unités linguistiques en fonction de leur fréquence dans les documents [Plantie (2006)].

2.5 Les techniques de classification

La construction des classificateurs du texte a été abordée de diverses manières. Ici, nous présentons seulement les méthodes qui ont été les plus populaires dans le domaine de la catégorisation du texte - TC. La classification, appelée également induction supervisée, consiste à analyser de nouveaux candidats et à les affecter, en fonction de leurs caractéristiques ou attributs, à telle ou telle classe prédéfinie. Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains traits descriptifs. La classification fournit de l'aide à la prise de décision comme par exemple pour établir un diagnostic médical à partir de la description clinique d'un patient, pour donner une réponse à la demande de prêt bancaire de la part d'un client sur la base de sa situation personnelle ou pour déclencher un processus d'alerte en fonction de signaux reçus par des capteurs. La construction inductive d'un classificateur de classement pour une catégorie $c_i \in C$ consiste habituellement en la définition d'une fonction telle pour un document d_j, qu'elle retourne une valeur indiquant l'état de catégorisation qui représente le fait que $d_j \in c_i$. Cette fonction de l'état de catégorisation (*ang : Categorization Status Value - CSV*) prend des valeurs entre 0 et 1. Les documents sont ensuite classés en fonction de leur valeur CSV_i. La fonction CSV_i prend différentes formes selon la méthode d'apprentissage utilisée : par exemple, dans l'approche de "naif Bayes" $CSV_i(d_j)$ est définie en termes de probabilité.

La construction d'un classificateur peut consister en la définition d'une fonction $CSV_i : D \rightarrow \{T, F\}$ ou d'une fonction $CSV_i : D \rightarrow [0, 1]$. Un paramètre de seuil (*ang : threshold* τ_i) est défini tel que $CSV_i(d_j) \geq \tau_i$ est interprété comme *Vrai - T* et $CSV_i(d_j) \leq \tau_i$ est interprété comme *Faux - F*.

Le paramètre de seuil peut être calculé *analytiquement* ou *expérimentalement*. La méthode analytique est possible uniquement en présence de résultat théorique qui indique comment calculer le seuil pour maximiser l'efficacité [Lewis (1995)]. C'est le cas

pour les classificateurs de probabilité.

Quand un tel résultat théorique n'est pas connu, il faut revenir à la deuxième méthode - expérimentale, qui consiste à tester différentes valeurs de τ_i sur un ensemble de documents de validation (EV) et de choisir la valeur qui maximise l'efficacité [Cohen & Singer (1999), Wiener *et al.* (1995), Yang (1999)].

La procédure de classification est générée automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante. Nous disposons, par exemple, d'une base des symptômes des patients avec la situation de leur état de santé respectif, et le diagnostic de maladie. Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification qui, au vu des symptômes d'un patient, devra établir un diagnostic médical. Il s'agit donc d'induire une procédure de classification générale à partir d'exemples. Le problème est donc un problème inductif, il s'agit d'extraire une règle générale à partir de données observées. La procédure générée devra classifier correctement les exemples de l'échantillon mais surtout avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions. Les méthodes statistiques supposent que les descriptions des objets d'une même classe se répartissent en respectant une structure spécifique à la classe. Les méthodes d'apprentissage à partir d'exemples sont très utilisées dans la recherche d'informations dans de grands ensembles de données.

2.5.1 Classificateur de Bayes

Les classificateurs probabilistes interprètent la fonction $CSV_i(d_j)$ en termes de $P(c_i|\vec{d_j})$, ce qui représente la probabilité qu'un document représenté par un vecteur $\vec{d_j} =< w_1, j, ..., w_{|T|j} >$ de termes qui appartient à c_i, et calcule cette probabilité en utilisant le théorème de Bayes, définie par :

$$P(c_i|\vec{d_j}) = \frac{P(c_i)P(\vec{d_j}|c_i)}{P(\vec{d_j})}. \tag{2.3}$$

où $P(\vec{d_j})$ est la probabilité qu'un document choisi au hasard ait le vecteur $\vec{d_j}$ comme représentation, et $P(c_i)$ est la probabilité qu'un document choisi au hasard appartienne à c_i. L'estimation de probabilité $P(c_i|\vec{d_j})$ est problématique, puisque le nombre de vecteurs

$\vec{d_j}$ possibles est trop élevé. Pour cette raison il est courant de faire l'hypothèse que toutes les coordonnées du vecteur du document sont statistiquement indépendantes. Donc :

$$P(\vec{d_j}|c_i) = \Pi_{k=1}^{|T|} P(w_{kj}|c_i)). \tag{2.4}$$

Les classificateurs probabilistes qui utilisent cette hypothèse sont appelés *les classificateurs de "naïf Bayes"*, et trouvent utilisation dans la plupart des approches probabilistes dans le domaine de catégorisation du texte [Wang *et al.* (2005), Lewis & Gale (1994)]. Le caractère *naïve* du classificateur est dû au fait qu'en général, cette hypothèse n'est pas vérifiée dans la pratique.

2.5.2 Calcul d'un classificateur par la méthode des SVM

Les méthodes des SVM *ang : Support Vector Machine* a été introduite par Joachims [Joachims (1998), Joachims (1999)], puis utilisée par Drucker [Drucker *et al.* (1999)], Taira et Haruno [Taira & Haruno (1999)], et Yang et Liu [Yang & Liu (1999)]. La méthode des SVM géométriques peut être considérée comme la tentative de trouver, parmi toutes les surfaces $\sigma_1, \sigma_2, ...$ d'un espace de dimensions $|T|$ ce qui sépare les exemples d'apprentissage positifs des négatifs. L'ensemble d'apprentissage est donné par un ensemble de vecteurs associés à leur classe d'appartenance : $(X_1, y_1), ..., (X_u, y_u)$, $X_j \in R^n, y_j \in \{+1, -1\}$ avec

- y_j représente la classe d'appartenance. Dans un problème à deux classes la première classe correspond à une réponse positive ($y_j = +1$) et la deuxième classe correspond à une réponse négative (y_j = -1)
- X_j représente le vecteur du texte numéro j de l'ensemble d'apprentissage.

La méthode SVM sépare les vecteurs à classe positive des vecteurs à classe négative par un hyperplan défini par l'équation suivante : $W \otimes X + b = 0, W \in R^n, b \in R$ [Figure 2.4].

En général, un tel hyperplan n'est pas unique. La méthode SVM détermine l'hyperplan optimal en maximisant la marge : la marge est la distance entre les vecteurs étiquetés positifs et les vecteurs étiquetés négatifs. L'ensemble d'apprentissage n'est pas nécessairement séparable linéairement, des variables d'écart ξ_j sont introduites pour

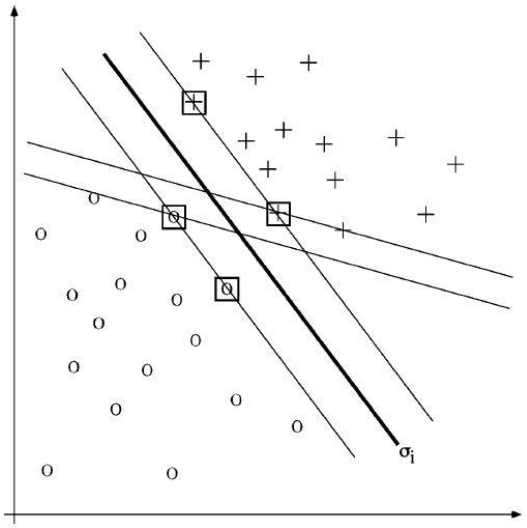


Figure 2.4: Apprentissage du classificateur SVM - Les cercles et les croix représentent respectivement les réponse positives et négatives, les lignes représentent les surfaces de décision. La surface de décision σ_i montre le meilleur cas.

tous les X_j . Ces ξ_j prennent en compte l'erreur de classification, et doivent satisfaire les inégalités suivantes :

- $W \otimes X + b \geq 1 - \xi_j$,
- $W \otimes X + b \leq 1 + \xi_j$.

En prenant en compte ces contraintes, nous devons minimiser la fonction d'objectif suivante : $\frac{1}{2} \|W\|^2 + C\Sigma_{j=1}^{u} \xi_j$. Le premier terme de cette fonction correspond à la taille de la marge et le second terme représente l'erreur de classification, avec u représentant le nombre de vecteurs de l'ensemble d'apprentissage. Trouver la fonction objective précédente revient à résoudre le problème quadratique suivant : trouver la fonction de décision f telle que : f(X) = signe(g(X)) dans laquelle la fonction g(X) est :

$$g(X) = \Sigma_{i=1}^{m} \lambda_i y_i X_i \otimes X + b \tag{2.5}$$

avec :

- Signe(x) représente la fonction suivante :
  - Si x>0 alors Signe(x)=1
  - Si x<0 alors Signe(x)=-1

  - Si x=0 alors Signe(x)=0
- y_j représente la classe d'appartenance,
- λ_i représente les paramètres à trouver,
- $X_i \otimes X$ représente le produit scalaire du vecteur X_i avec le vecteur X.

2.5.3 Calcul d'un classificateur par la méthode des arbres de décision

Un classificateur de texte basé sur la méthode d'arbre de décision est un arbre de noeuds internes qui sont marqués par des termes, les branches qui sortent des noeuds sont des tests sur les termes, et les feuilles sont marquées par catégories [Mitchell (1996)]. Ce classificateur classe un document du test d_j en testant récursivement les poids des noeuds internes de vecteur $\vec{d_j}$, jusqu'à ce qu'une feuille soit atteinte. L'étiquette de ce noeud est alors attribuée à d_j. La plupart de ces classificateurs utilise une représentation du document binaire, et sont donc crée par des arbres binaires.

Il existe un certain nombre d'approches pour l'apprentissage de l'arbre de décision. Les plus populaires sont ID3 (utilisé par Fuhr [Fuhr *et al.* (1991)]), C4.5 (Cohen et Singer [Cohen & Singer (1999)], Joachims [Joachims (1998)]) et C5 (utilisé par Li et Jain [Li & Jain (1998)]).

Une méthode pour effectuer l'apprentissage d'un arbre de décision pour la catégorie c_i consiste à vérifier si tous les exemples d'apprentissage ont la même étiquette (c_i ou $\bar{c_i}$), dans le cas contraire nous sélectionnons un terme t_k, et nous partitionnons l'ensemble d'apprentissage en classes de documents qui ont la même valeur pour t_k, et à la fin l'on crée les sous-arbres pour chacune de ces classes. Ce processus est répété récursivement sur les sous-arbres jusqu'à ce que chaque feuille de l'arbre généré de cette façon contienne des exemples d'apprentissage attribués à la même catégorie c_i, qui est alors choisie comme l'étiquette de la feuille. L'étape le plus important est le choix du terme de t_k pour effectuer la partition. Toutefois, une telle méthode de construction d'arbre peut faire l'objet de surapprentissage, comme certaines branches peuvent être trop spécifiques pour les données d'apprentissage. La plupart des méthodes d'apprentissage des arbres incluent une méthode pour la construction d'arbre et pour élaguer les branches trop spécifiques [Mitchell (1996)].

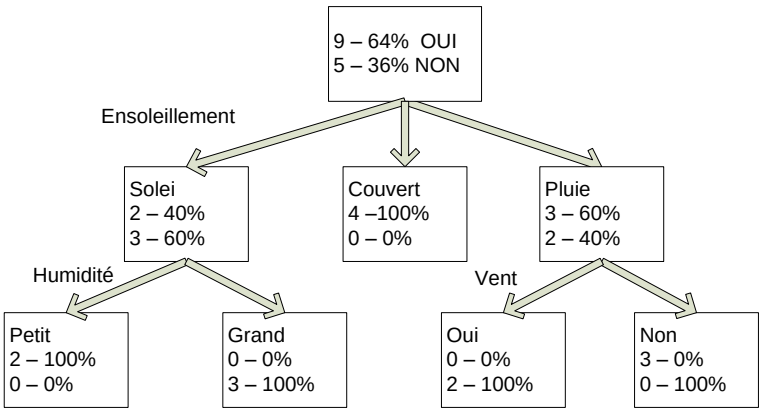


FIGURE 2.5: L'exemple d'un arbre de décision - L'algorithme d'apprentissage cherche à produire des groupes d'individus les plus homogènes possible du point de vue de la variable à prédire à partir des variables de météo.

L'exemple d'un arbre de décision décrit dans l'ouvrage de Quinlan [Quinlan (1993)] est montrée sur [Figure 2.5]. Il s'agit de prédire le comportement de sportifs (Jouer ; variable à prédire) en fonction de données météo (Ensoleillement, Température, Humidité, Vent ; variables prédictives). Sur chaque sommet de l'arbre est décrite la distribution de la variable à prédire. Dans le cas du premier sommet, la racine de l'arbre, nous constatons qu'il y a 14 observations dans notre fichier, 9 d'entre eux ont décidé de jouer (Jouer = oui), 5 ont décidé le contraire (Jouer = non). Ce premier sommet est segmenté à l'aide de la variable Ensoleillement, 3 sous-groupes ont été produits. Le premier groupe à gauche (Ensoleillement = Soleil) comporte 5 observations, 2 d'entre eux correspondent à Jouer = oui, 3 à Jouer = non. Chaque sommet est ainsi itérativement traité jusqu'à ce que l'on obtienne des groupes suffisamment homogènes. Elles correspondent aux feuilles de l'arbre, des sommets qui ne sont plus segmentés.

2.5.4 Réseau de neurones

Un classificateur de texte basé sur les réseaux de neurones (*ang : neural networks NN*) est un réseau d'unités, où les unités d'entrée représentent les termes, l'unité(s)

de sortie représentent la catégorie ou les catégories d'intérêts, et le poids sur les bords reliant les unités représentent les relations de dépendance. Pour classer un document de test d_j, ses poids w_{kj} sont chargés dans les unités d'entrée ; l'activation de ces unités se propage à travers le réseau, et la valeur de l'unité de sortie(s) détermine la décision du classement. Une manière typique d'apprentissage de réseau de neurones est la retro propagation, qui consiste à rétro propager l'erreur commise par un neurone à ses synapses et aux neurones qui y sont reliés. Pour les réseaux de neurones, on utilise habituellement la rétro propagation du gradient de l'erreur, qui consiste à corriger les erreurs selon l'importance des éléments qui ont justement participé à la réalisation de ces erreurs.

Le type le plus simple de classificateur de NN est le perceptron [Dagan *et al.* (1997), Ng *et al.* (1997)], qui est un classificateur linéaire. Dans cet algorithme, le classificateur de c_i est d'abord initialisé par la mise de tous les poids w_{kj} à la même valeur positive. Quand un exemple d'apprentissage d_j (représenté par un vecteur $\vec{d_j}$ de poids binaire) est examiné, et si le résultat de la classification est correct, rien n'est fait, alors que si le résultat est faux, les poids du classificateur sont modifiés :

- si d_j est un exemple positif de c_i, le poids de $w_k i$ des termes actifs (c'est-à-dire, les termes t_k tels que $w_{kj} = 1$) sont "promus" par augmentation d'une valeur $\alpha > 0$ (appelé taux d'apprentissage),
- si d_j est un exemple négatif de c_i les mêmes poids sont "rétrogradés" en diminuant leur valeur par α.

Lorsque le classificateur a atteint un niveau raisonnable d'efficacité, le fait qu'un poids w_{ki} est très faible signifie que t_k a contribué négativement à la procédure de classement, il peut donc être éliminé de la représentation. Le classificateur perceptron a montré une bonne efficacité [Dagan *et al.* (1997)].

Un réseau de neurones est en général composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Chaque couche est composée de n_i neurones, prenant leurs entrées sur les n_{i-1} neurones de la couche précédente. À chaque synapse est associé un poids synaptique, de sorte que les n_{i-1} sont multipliés par ce poids, puis additionnés par les neurones de niveau i, ce qui est équivalent à multiplier le vecteur d'entrée par une matrice de transformation. Le réseau de neurones peut

également contenir des boucles qui en changent radicalement les possibilités mais aussi la complexité. De la même façon que des boucles peuvent transformer une logique combinatoire en logique séquentielle, les boucles dans un réseau de neurones transforment un simple dispositif de reconnaissance d'entrées en une machine complexe capable de toutes sortes de comportements. La représentation schématique d'un neurone artificiel avec un index j est montre sur [Figure 2.6].

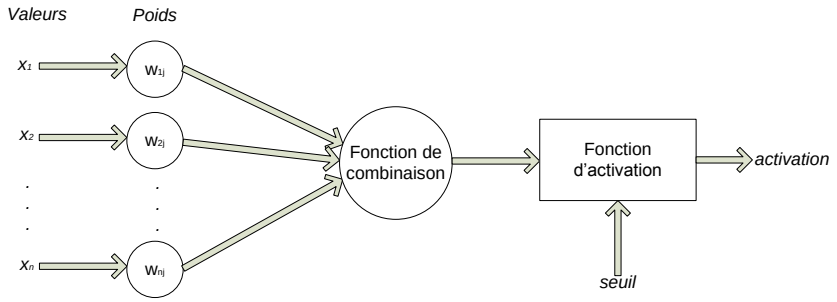


Figure 2.6: Structure d'un neurone artificiel - Le neurone calcule la somme de ses entrées puis cette valeur passe à travers la fonction d'activation pour produire sa sortie.

2.5.5 Mesure de performance

L'évaluation du classificateur des documents est effectuée de façon expérimentale, plutôt que de façon analytique. Le traitement est expérimental car pour évaluer un système de façon analytique, donc pour prouver que le système est correct et complet, nous aurions besoin d'une spécification formelle du problème que le système tente de résoudre. La notion de systèmes de catégorisation de texte est, en raison de son caractère subjectif, par nature non-formalizable comme il s'agissait de déterminer l'appartenance d'un document à une catégorie.

L'efficacité de classification se mesure généralement par les paramètres classiques du domaine de la recherche d'information :

- la précision,
- le rappel.

La *Précision* π_i est définie comme la probabilité conditionnelle $P(\breve{\Phi}(d_x, c_i) = T|\Phi(d_x, c_i) = T)$ qui signifie la probabilité que, si d_x est classé dans c_i, cette décision est correcte. La précision c'est un ratio entre le nombre de documents pertinents trouvés et le nombre

total de documents trouvés. Elle mesure le bruit, et plus elle est proche de 100%, moins il y a de bruit, et donc meilleure est la réponse.

De même, le *rappel* ρ_i est défini comme $P(\Phi(d_x, c_i) = T|\breve{\Phi}(d_x, c_i) = T)$, qui signifie la probabilité que, si un document quelconque d_x devait être classé sous c_i , cette décision était prise. Le rappel est un ratio entre le nombre de documents pertinents trouvés et le nombre de documents pertinents présents dans la base. Plus il est proche de 100%, moins il y a de silence, et meilleure est la réponse.

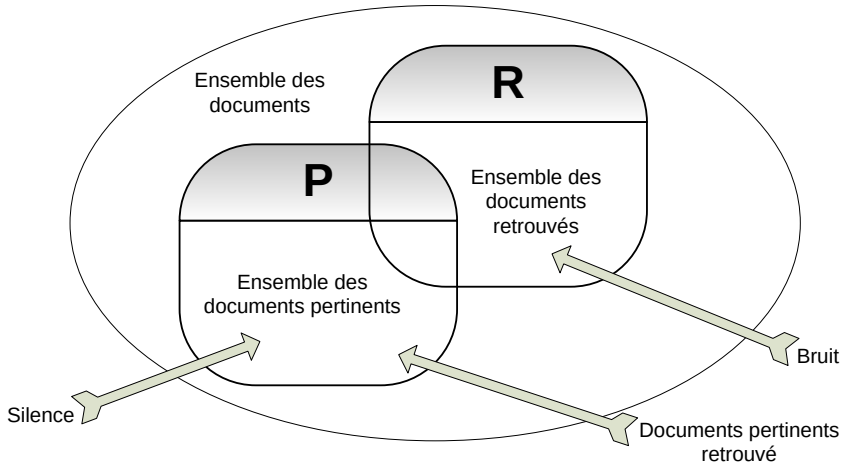


Figure 2.7: Exemple de représentation du bruit et du silence en recherche de l'information. - Le rôle de la pertinence et du rappel dans la définition du bruit et silence

La précision peut être considérée comme le "degré de solidité" du classificateur, tandis que le rappel peut être considéré comme son "degré d'exhaustivité". L'évaluation de la pertinence dans les recherches documentaires essaie de quantifier le rappel et la précision, avec les indices de bruit et de silence [Figure 2.7].

Le bruit est un ratio entre le nombre de documents ramenés en réponse, mais qui ne sont pas pertinents par rapport à la question posée, et le nombre total de documents. Donc :

$$Bruit = 1 - Précision$$

Le silence est un ratio entre le nombre de documents pertinents qui n'apparaissent pas

dans le résultat de la recherche et le nombre total de documents. Donc :

$$Silence = 1 - Rappel$$

Selon cette définition, la précision et le rappel peuvent être définis comme les probabilités subjectives, qui mesurent l'attente de l'utilisateur d'un système se comportant correctement lors de la classification d'un document inconnu pour une catégorie c_i. Ces probabilités peuvent être estimées par un tableau qui montre toutes les possibilités d'un classificateur [Tableau 2.1].

Tableau 2.1: Possibilité de résultat du classificateur pour une catégorie ci

Categorie c_i		Classement de l'Expert	
		Vrai	Faux
Jugement de	Positif	TP_i	FP_i
classificateur	Négatif	FN_i	TN_i

FP_i (erreurs de commission) est le nombre de documents de test mal classés dans la catégorie c_i ; TN_i est l'ensemble de documents bien classés qui n'appartiennent pas à la catégorie c_i ; TP_i est l'ensemble de documents bien classés qui appartiennent à la catégorie c_i), et FN_i (erreurs d'omission) est l'ensemble des documents de catégorie c_i non classés par le classificateur. La précision et le rappel peuvent être exprimés de la façon suivante :

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \tag{2.6}$$

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \tag{2.7}$$

Les mesures alternatives de précision et de rappel couramment utilisées dans la littérature, telles que la pertinence (*ang : accuracy*) :

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

et erreur :

$$E = \frac{FP + FN}{TP + TN + FP + FN} = 1 - A \tag{2.9}$$

ne sont pas trop utilisées dans les techniques du domaine de catégorisation du texte, [Yang (1999)] car leur dénominateur a des grandes valeurs, ce qui les rend beaucoup plus insensibles aux variations de bonnes décisions : les valeurs des nominateurs TP+TN et FP+FN des formules du calcul de la pertinence et de l'erreur respectivement n'auront pas d'influence sur les résultats de ces dérniers.

Une mesure d'efficacité non standard a été proposée par Sable et Hatzivassiloglou [Sable & Hatzivassiloglou (2000)], qui ont suggéré de baser la mesure de précision et de rappel non pas sur les valeurs "absolue" de réussite et d'échec - c'est-à-dire :

- 1 si $\Phi(d_j, c_i) = \breve{\Phi}(d_j, c_i)$
- 0 si $\Phi(d_j, c_i) \neq \breve{\Phi}(d_j, c_i)$,

mais sur les valeurs de succès relatif - c'est-à-dire :

- $CSV_i(d_j)$ si $\breve{\Phi}(d_j, c_i) = T$
- $1 - CSV_i(d_j)$ si $\breve{\Phi}(d_j, c_i) = F$.

Cela signifie que, pour une décision correcte (respectivement mauvaise), le classificateur est récompensé (respectivement pénalisé) proportionnellement à la confiance que l'on peut accorder à la décision. Cette mesure ne récompense pas le choix d'un bon seuil, et est donc impropre à l'autonomie des systèmes de classification.

Ni le rappel ni la précision ne donnent du sens séparément l'un de l'autre. En fait, le classificateur Φ tel que $\Phi(d_j, c_i) = T$ pour tous les d_j et c_i (l'accepteur trivial) a le rappel $\rho = 1$. Lorsque la fonction CSV_i a une valeur dans l'intervalle [0, 1], il suffit de mettre chaque seuil à 0 pour obtenir l'accepteur trivial. Dans ce cas la précision est généralement très faible. En revanche il est bien connu que des niveaux plus élevés de la précision peuvent être obtenus au prix de faibles valeurs du rappel. Donc une classification doit être évalué en combinant des valeurs de précision et de rappel. Plusieurs mesures ont été proposées, parmi lesquelles les plus fréquentes sont les suivants :

- seuil de rentabilité (*ang : breakeven*),
- function F_β,
- function Fscore.

La première mesure - le seuil de rentabilité (*ang : breakeven*), est la valeur à laquelle la précision est égale au rappel [Joachims (1999), Yang (1999)]. Ce résultat est obtenu

par le tracé de la précision en fonction du rappel pour divers seuils ; le seuil de rentabilité est la valeur de la précision ou de rappel pour lesquelles la courbe coupe la ligne $\rho = \pi$. Cette idée repose sur le fait que, en diminuant le seuil de 1 à 0, le rappel augmente toujours de 0 à 1 et la précision diminue monotonement d'une valeur proche de 1.

La deuxième mesure est la fonction F_β [Rijsbergen (1979)], pour certains $\beta \in (0, \infty)$ qui est telle que :

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho} \tag{2.10}$$

β peut être considéré comme le degré relatif d'importance attribué à la précision et au rappel. Ce coefficient indique le poids que l'on souhaite affecter à la précision par rapport au rappel. Habituellement, le coefficient β prend la valeur 1. Dans ce cas la fonction F_β est appelée Fscore et prend la forme :

$$F_1 = \frac{2\pi\rho}{\pi + \rho} \tag{2.11}$$

2.6 Conclusion

Dans ce chapitre nous avons présenté les techniques du domaine de Categorisation du Texte. Ce domaine fait partie du domaine de la Recherche d'Information. Nous avons présenté les techniques d'Apprentissage Automatique en précisant les techniques de classification, ainsi que les différents moyens de représentation de corpus documentaires.

L'une des raisons pour lesquelles depuis le début des années 90 l'efficacité des classificateurs de texte a été améliorée de façon spectaculaire est l'arrivée des méthodes d'Apprentissage Automatique, qui ont significativement augmenté l'utilisation de systèmes de ML.

Dans le chapitre qui suit nous monterons l'utilisation des techniques présentées dans ce chapitre pour l'analyse des sentiments - le domaine de l'*Opinion Mining*. Ce domaine fait partie de Catégorisation de Texte et utilise ces techniques pour analyser l'opinion et les sentiments décrits dans le texte.

Chapitre 3

Analyse des sentiments

3.1 Opinion Mining, Analyse des Sentiments

Dans le chapitre précédent nous avons introduit le terme *Opinion Mining*. C'est le domaine qui s'occupe de traitement d'opinion, du sentiment, et de la subjectivité dans le texte et nous avons précisé que c'est un sous domaine de la catégorisation de texte. Les principales tâches de l'*Opinion Mining* sont l'analyse de l'opinion et l'analyse de la subjectivité. Cette dernière est utilisée pour reconnaitre le langage décrit l'opinion afin de distinguer les langues objectives.

Le terme *Opinion Mining* apparaît dans un article de Dave [Dave *et al.* (2003)] qui a été publié dans l'acte de conférence WWW 2003. Selon Dave, l'*Opinion Mining* devrait "traiter un ensemble de résultats de recherche pour un cas donné, générer une liste des attributs (qualité, caractéristiques, etc.) et agréger des avis sur chacun d'entre eux (mauvais, modéré, de bonne qualité). Toutefois, l'*Opinion Mining* a récemment été interprétée de manière plus générale pour inclure de nombreux types d'analyse d'évaluation de texte [Liu (2006)].

Le terme *"Analyse des Sentiments"* est utilisé pour décrire l'analyse automatique de texte évaluatif et pour la recherche de valeur prédictive des jugements. Elle a été introduite dans les travaux de Das et Chen [Das & Chen (2001)] et Tong [Tong (2001)] en 2001 afin d'analyser des sentiments dans le cadre de l'économie de marché. Ensuite d'autres travaux sur l'analyse des sentiments ont été proposés par Turney [Turney

(2002)] et Pang et al [Pang *et al.* (2002)]. Depuis 2002, un nombre important d'articles citant l'*Analyse des Sentiments* ont vus le jour, ces travaux se concentrent sur la classification des commentaires et à leur polarité (positif ou négatif). Aujourd'hui, l'*Opinion Mining* et l'*Analyse des Sentiments* font partie du même domaine de recherche.

3.2 Les besoins de connaitre des sentiments des autres

Connaitre l'opinion des autres personnes a toujours été un élément d'information important durant le processus de décision. Les gens très souvent demandent à d'autres de leur recommander un mécanicien d'automobiles ou d'expliquer leur choix de votes aux élections par exemple. Avant de prendre des décisions, les gens s'intéressent énormément aux avis des autres personnes dans différents domaines. Ils consultent les avis des autres consommateurs avant d'effectuer un achat, ou regardent les avis des autres personnes avant de voir un film au cinéma ou avant d'acheter un disque. Grace à l'Internet nous pouvons découvrir les opinions et les expériences de très grand nombre de personnes qui ne sont ni nos amis, ni les experts de domaines, mais des gens qui peuvent avoir les mêmes goûts que nous, et donc leurs opinions peuvent être très utiles pour nous avant de faire notre choix et d'avoir notre propre idée sur un sujet donné. Aujourd'hui, de plus en plus de personnes donnent leur avis sur différents sujets, ces avis sont à la disposition de tout le monde sur internet.

Selon les sondages [comScore/the Kelsey group (2007), Horrigan (2008)], 81% des utilisateurs de l'Internet ont fait au moins une fois la recherche en ligne sur un produit et environ 80% parmi eux déclarent que les opinons des autres personnes ont une influence significative sur leur décision d'achat, ce qui représente un très grand nombre de personnes. Environ 30% ont fourni un avis sur un produit, sur un service ou sur une personne en ligne via un système de notation, ce qui n'est pas insignifiant comme nombre. Pour cette raison, c'est à dire grâce a l'intérêt que les utilisateurs montrent pour les opinions sur les produits et les services, ainsi que l'influence potentielle qu'exercent de tels avis, les fournisseurs des articles montrent une très grande attention au développement des systèmes de notations [Hoffman (2008)]. Avec l'explosion du Web 2.0, des plates-formes comme les blogs, des forums de discussion, de réseau Peer-to-Peer, et divers autres types de moyens de communication sociale, les consommateurs ont à

leur disposition une tribune sans précédent, de portée et de puissance, permettant de partager leurs expériences et de marquer leur avis (positif ou négatif) sur n'importe quel produit ou service. Les entreprises peuvent répondre aux besoins des consommateurs en effectuant de la surveillance et de l'analyse des opinions pour améliorer leur produit [Zabin & Jefferies (2008)]. Malheureusement le risque de modification des opinions est important. De ce fait, il est nécessaire d'avoir un système capable d'analyser automatiquement les comportements généraux liés à la consommation, afin de mieux comprendre comment les différents produits et les services sont perçus par les clients.

Un tel système devra premièrement collecter des opinions des consommateurs et des utilisateurs dans des documents qui montrent les opinions et les phrases subjectives. Parfois, cela est relativement facile, comme dans les cas de grands sites où les opinions des utilisateurs sont bien structurées comme par exemple Epinions.com, Imdb.com, Amazon.com. Le problème devient plus complexe dans le cas des blogs, qui contiennent aussi des parties de texte subjectives, mais les documents souhaités dans les blogs peuvent varier assez largement dans le contenu, le style, la présentation et même en niveau de grammaticalité. Il est très intéressant de travailler sur des commentaires venant des blogs car ils sont plus pertinents que les sites de vente, et généralement ils expriment mieux l'intensité des opinons.

Une fois que les documents intéressants sont collectés, nous sommes confrontés au problème d'identification de l'ensemble des avis et sentiments exprimés par ces documents. Pour résoudre cette tâche, il faut préciser le domaine d'intérêt, car si par exemple nous notons les opinions de film nous remarquons que la langue est spécifique, les caractéristiques de films peuvent être groupés dans des ensembles prédéfinies ce qui facilitera l'analyse automatique.

La dernière étape du système est de présenter les résultats de sa notation en précisant l'intensité de chaque opinion. Un tel système a été créé durant cette thèse. La présentation de tout le système ainsi que les différentes approches de notation des opinions seront détaillées dans les chapitres suivants.

3.3 La complexité de notation d'opinion

Pour démontrer la complexité de la notation de l'opinion nous allons nous baser sur un exemple d'une critique cinématographique retrouvée sur le site IMDB.com. L'exemple est le suivant :

> It's A Wonderful Life. I've only met 2 people in real life and 1 person on the IMDB who hates this one. My favorite film ever !

Comme nous pouvons le constater, la critique est composée de trois phrases qui ont une polarité opposée. Même si nous arrivons à déduire facilement que la première phrase étant le titre de film, *It's a Wonderful Life*, nous aurons deux phrases subjectives mais difficile à noter correctement. La dernière phrase est plutôt facile à noter : *Mon film préféré de tout les temps.* Mais le problème se pose pour la notation de la phrase : *J'ai connu seulement ... qui ont détesté ce film.* Car une étude statistique nous montre que la polarité est négative pour cette phrase pourtant la polarité est réellement positive et avec une très grande intensité.

Les résultats d'une étude de Pang et al. [Pang *et al.* (2002)], sur les critiques cinématographiques montrent que l'utilisation de mots clés corrects peut être moins triviale que l'on pourrait penser initialement. Le but de Pang et al. était de mieux comprendre la difficulté de classification de polarité des sentiments. Deux personnes ont été invitées à choisir des mots clés qu'ils considèrent comme de bons indicateurs des opinions positives et négatives. Les deux listes de mots clés réalisent environ 60% de précision. En revanche, les listes de mots de la même taille, mais choisis en fonction de traitement statistiques sur un document d'apprentissage réalisent près de 70% de précision. En effet, l'application de techniques d'apprentissage automatique (ML) basée sur les modèles d'unigramme peut atteindre plus de 80% de précision [Pang *et al.* (2002)], qui est beaucoup mieux que la performance basée sur les mots-clés des experts.

Les sentiments peuvent souvent être exprimés d'une manière très subtile, ce qui rend difficile l'identification par les unités du document quand nous les considérons séparément. Si nous considèrons une phrase qui indique une très forte opinion, il est difficile d'associer cette opinion avec les mots-clés ou les expressions dans cette phrase. En général, les sentiments et la subjectivité sont très sensibles au contexte et dépendent de

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

De plus sur l'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 film, avec les meilleurs acteurs, meilleurs metteurs en scène, mais la dernière phrase peut être *Malgré tout ça je suis sortie du cinéma avant la fin.* Même si nous arrivons à attribuer l'intensité de cette opinion nous nous retrouverons avec une seule opinion négative contre plusieurs 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 ça ne respecte aucune règle et il est impossible de prévoir tout les cas possibles, en plus très souvent la même phrase peut être considérée comme positive pour une personne et négative pour une autre.

3.4 Détection de phrases subjectives

Pour de nombreuses applications, nous devons décider si un document contient des informations subjectives ou non et d'identifier quelles parties du document sont subjectives, pour pouvoir ensuite traiter seulement la partie subjective.

Les travaux de Hatzivassiloglou et Wiebe [Hatzivassiloglou & Wiebe (2000)] ont démontré l'orientation des phrases en se basant sur l'orientation des adjectifs. L'objectif était de dire si une phrase donnée est subjective ou non en jugeant les adjectifs figurant dans cette phrase [Beineke *et al.* (2004), Wiebe *et al.* (2001), Wilson *et al.* (2005),]. Wiebe et al. [Wiebe *et al.* (2004)] présentent une étude complète sur la reconnaissance de la subjectivité en utilisant différents indices et caractéristiques (la comparaison des résultats en utilisant les adjectifs, les adverbes et les verbes en prenant en compte la structure syntaxique comme par exemple l'emplacement des mots).

Une autre approche a été proposée par Wilson et al. [Wilson *et al.* (2004)] qui ont proposé une classification des opinions selon leur intensités (la force de l'opinion) et selon d'autres éléments subjectifs. Lorsque d'autres recherches ont porté sur la distinction entre subjectivité et objectivité ou sur la distinction des phrases positives et négatives, Wilson et al. ont classé la force des opinions et des émotions exprimées dans des clauses individuelles. La force est dite *neutre* lorsqu'elle correspond à l'absence d'opinion et de subjectivité.

Des travaux récents considèrent également les relations entre l'ambigüité du sens des mots et de la subjectivité [Wiebe & Mihalcea (2006)]. La détection de subjectivité peut aussi être effectuée grâce aux techniques de la classification. Par exemple, Yu et Hatzivassiloglou [Yu & Hatzivassiloglou (2003)] atteignant une précision élevée de 97% en utilisant un classificateur "naïf Bayes" sur un corpus spécifique composé des articles de Wall Street Journal. La tâche consistait en une séparation entre les faits (articles d'actualités et d'affaires) et les opinions (articles de réponses de la rédaction aux lettres des lecteurs).

3.5 La polarité et l'intensité de l'opinion

La classification de polarité de l'opinion consiste à la classification d'un document comme positif ou négatif. Beaucoup de travaux sur la détection de polarité ont été effectués pour la critique cinématographique. Dans ce contexte l'opinion positive et l'opinion négative sont souvent évaluatives.

Une valeur appelée orientation sémantique a été créée pour démontrer la polarité des mots. Elle varie en deux grandeur : positive et négative et peut avoir différents niveaux d'intensité. Il existe plusieurs méthodes de calcul de l'orientation sémantique pour des mots. En général la méthode d'orientation sémantique des associations *SO-A* est calculée comme une mesure de l'association des mots positifs moins la mesure de l'association des mots négatifs :

$$SO - A(mot) = \Sigma_{pmot \in Pmots} A(mot, pmot) - \Sigma_{nmot \in Nmots} A(mot, nmot) \qquad (3.1)$$

où :
- A(mot,pmot) correspond à l'association de mot étudie avec le mot positive.

- A(mot,nmot) correspond à l'association de mot étudie avec le mot négative.

Si la somme est positive, le mot est orienté positivement, et si la somme est négative, l'orientation est négative. La valeur absolue de la somme indique l'intensité de l'orientation. Pour calculer la mesure de l'association entre les mots - A, il existe plusieurs possibilités. L'une d'elle est appelée *The Pointwise Mutual Information - SO-PMI* (proposée par Church et Hanks).

$$PMI(mot_1, mot_2) = log_2 \frac{p(word_1 \& word_2)}{p(word_1)p(word_2)} \tag{3.2}$$

Le $p(word_1 \& word_2)$ définit la probabilité que les deux mots coexistent ensemble. A titre d'exemple, le moteur d'AltaVista Advance search utilise la technique SO-PMI. Il se base sur l'operateur *near* pour le calcul de l'éloignement de deux mots, cet operateur prend une distance de 10 comme voisinage du mot.

Une autre possibilité pour analyser la relation statistique entre les mots dans le corpus est l'utilisation de la technique : *the Singular Value Decomposition (SVD)*. La méthode qui utilise SVD utilisée par Laundauer et Dumais [Dumais *et al.* (1998)] est appelée *Latent Semantic Analysis - SO-LSA*. Elle est basée sur la décomposition de la matrice qui contient en ligne et en colonnes les pondérations des mots et des parties du texte comme par exemple les phrases ou les paragraphes. Généralement la pondération d'un mot dans la chaine est calculée par rapport au *Term Frequency Inverse Document Frequency - tf-idf* [*Section 2.4.3*].

3.6 Différents approches pour l'analyse des sentiments

3.6.1 Le rôle de n-grammes dans la classification

La position des unités linguistiques dans une unité de texte par exemple, au milieu ou à la fin d'un document peut changer le niveau des sentiments ou de subjectivité de texte. Les informations de position sont parfois codées dans des vecteurs. Dans la littérature, il y a une discussion sur l'utilisation des n-grammes. Pang et al. [Pang *et al.* (2002)] décrivent que lors du classement des critiques cinématographiques par la polarité en utilisant la méthode d'unigramme, les résultats sont meilleurs que ceux obtenus par l'utilisation de bi-grammes. Pourtant Dave et al. [Dave *et al.* (2003)] démontrent que

dans certains situation, les bi-grammes et les tri-grammes donnent de meilleurs produits de classification de polarité. Riloff et al. [Riloff *et al.* (2006)] explorent l'utilisation d'une hiérarchie pour définir différents types de caractéristiques lexicales et les relations entre eux afin d'identifier les caractéristiques complexes utiles pour l'analyse d'opinion.

3.6.2 L'importance des adjectifs

L'orientation sémantique des mots a été élaboré premièrement pour les adjectifs [Hatzivassiloglou & McKeown (1997), Mullen & Collier (2004), Whitelaw *et al.* (2005)]. Les travaux sur la détection de subjectivité ont révélé une forte corrélation entre la présence d'adjectifs et la subjectivité de phrase [Hatzivassiloglou & Wiebe (2000)]. Ce constat a souvent été considéré comme la preuve que certains adjectifs sont de bons indicateurs de sentiment. Un certain nombre d'approches axées sur la présence ou la polarité des adjectifs ont été élaborées pour déduire la subjectivité ou la polarité des textes. L'une des premières approches était proposée par Turney [Turney (2002)] : plutôt que de se concentrer sur les adjectifs isolés, Turney a proposé de détecter des sentiments du document en se basant sur des expressions qui contiennent un adjectif ou un adverbe. L'approche de Turney peut être présentée en 4 phases :

- tout d'abord une décomposition de phrases (part-of-spech) est effectuée,
- ensuite nous regroupons les adjectifs et les adverbes en chaines de deux mots comme par exemple *romantic ambience*,
- nous appliquons ensuite SO-PMI pour calculer l'orientation sémantique de chaque chaine détectée,
- à la fin, nous effectuons une classification de critiques comme positive ou négative en calculant la moyenne de toutes les orientations retrouvées.

Les résultats obtenus par cette approche sont différents par rapport au domaine : pour les automobiles = 84%, pour les documents bancaires = 80% et pour les critiques cinématographiques = 65%. Le fait que les adjectifs sont de bons prédicateurs de l'opinion ne diminue pas la signification des autres éléments. Pang et al. [Pang *et al.* (2002)], dans l'étude de polarité de critiques cinématographique, ont démontré qu'utiliser seulement les adjectifs comme caractéristiques donne des résultats bien plus mauvais qu'en utilisant le même nombre d'unigrammes.

3.6.3 Traitement de la négation

Une autre caractéristique importante pour déterminer la polarité de l'opinion est la négation. Le seul lemme décrivant la négation peut changer complètement la polarité de la phrase. Das et Chen [Das & Chen (2001)] proposent de rajouter une indication de négation "*NON*" à des mots qui se trouvent près de la négation, de sorte que dans la phrase "Je ne comprend pas", le lemme "comprendre" est converti en un nouveau lemme "comprendre-NON".

Cependant, certaines apparences de la négation n'inversent pas la polarité de la phrase. Na et al. [Na *et al.* (2004)] améliorent de 3% la précision résultante de la modélisation de la négation. Ils analysent le texte en effectuant une décomposition spécifique d'une phrase en recherchant des occurrences de négation. Si ces dernières ne sont pas étiquetées comme des mots de négation prédéfinis, ils classent la phrase entière comme étant une phrase avec négation, et non le mot séparé.

Une autre difficulté avec la négation est qu'elle peut être décrite de manière très subtile, ainsi le sarcasme et l'ironie sont très difficiles à détecter.

3.6.4 Utilisation des méthodes d'apprentissage automatique

Pang et all [Pang *et al.* (2002)] décrivent les travaux qui utilisent les techniques de classification en se basant sur l'apprentissage automatique. Ils appliquent trois différentes techniques de classification pour effectuer la classification de critiques cinématographiques. Les classificateur sont basées sur :

- classificateur "naïf Bayes",
- entropie maximale,
- classificateur SVM (Support Vector Machine).

Pang obtient les meilleurs résultats avec la méthode de SVM - la pertinence de 83% en utilisant les unigrammes.

Pang et Lee [Pang & Lee (2004)] proposent une autre approche pour la classification de polarité des critiques cinématographiques. L'approche est composée de deux étapes [Figure 3.1]. Leur premier objectif est de détecter les parties des documents qui

sont subjectives. Ensuite ils utilisent le même classificateur statistique pour détecter la polarité seulement sur les fragments subjectifs détectés précédemment. Au lieu de faire la classification de subjectivité pour chaque phrase individuellement, ils admettent qu'il pourrait y avoir un certain degré de continuité dans la subjectivité des phrases - un auteur en général, ne change pas souvent entre le fait d'être subjectif ou objectif. Ils attribuent des préférences afin que les phrases à proximité aient le même niveau de subjectivité. Toutes les phrases dans le document sont ensuite étiquetées comme étant subjectives ou objectives dans le processus de classification collective.

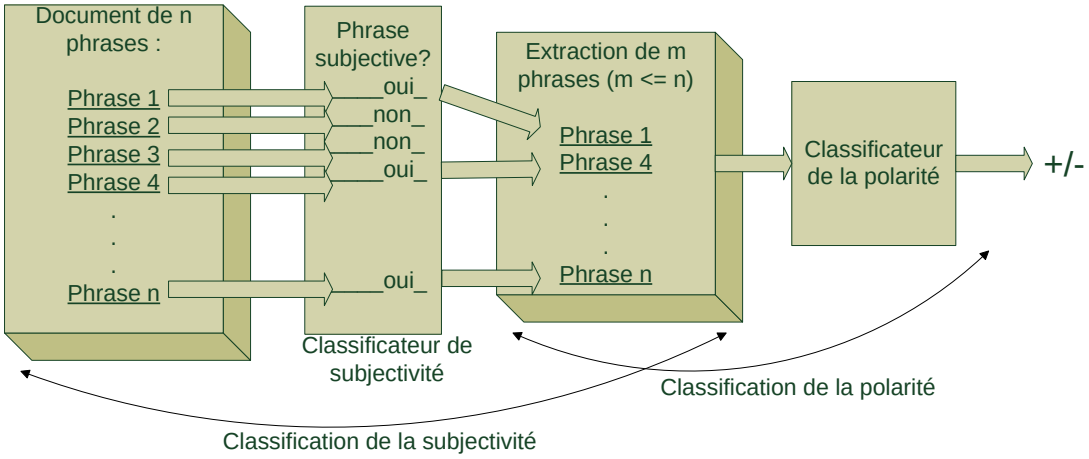


Figure 3.1: L'approche de Pang - Utilisation de la même technique de classification pour la détection de la subjectivité et ensuite de la polarité des phrases étiquetées comme subjectives

3.6.5 Approche de Dave

Une autre approche pour affirmer si la critique est positive ou négative a été proposée par Dave et al [Dave *et al.* (2003)]. Tout d'abord ils ont sélectionné un ensemble de caractéristiques $f_1, ..., f_n$, ils ont attribué ensuite les notes aux caractéristiques pour pouvoir placer les documents de test dans l'ensemble des critiques positives - C ou négatives - C'. Ensuite ils déterminent la fréquence d'occurrence normalisée - $p(f_i|C)$, en prenant le nombre d'occurences d'une caractéristique f_i dans C et en divisant par le nombre total de tokens dans C. La note attribuée aux caractéristiques est une valeur

allant de -1 à 1.

$$score(f_i) = \frac{p(f_i|C) - p(fi|C')}{p(f_i|C) + p(fi|C')} \tag{3.3}$$

Une fois que chaque caractéristique est notée, nous pouvons additionner les notes des mots d'un document inconnu et utiliser le signe de cette somme pour déterminer la classe C ou C'. Donc pour un document $d_j = f_1, ..., f_n$

$$class(d_i) = (C \ si \ eval(d_j) \ > \ 0 \ \ ou \ \ C' \ si \ eval(d_j) \ < \ 0) \tag{3.4}$$

ou $eval(d_j) = \Sigma_i score(f_i)$. La pertinence obtenue par cette approche est égale à 76% d'après les auteurs.

3.6.6 Utilisation de bootsraping

Une des approches pour la détection de la subjectivité des phrases est basée sur le *bootstrapping*, l'idée étant d'utiliser la sortie d'un classificateur initial pour avoir les données étiquetées sur lesquelles un algorithme d'apprentissage peut être appliqué.
Riloff et Wiebe [Riloff & Wiebe (2003)] ont utilisé cette méthode avec un classificateur initial de haute précision pour préparer la phase d'apprentissage. Cette dernière consistait en l'extraction des occurrences pour les expressions subjectives. Ils ont retrouvé des comportements intéressants ; par exemple le mot "*fact*" dans le contexte : *The fact is...* a une forte corrélation avec la subjectivité. Dans cette approche, Riloff and Wiebe utilisent deux classificateurs de haute précision mais avec un faible rappel : un classificateur de subjectivité et un classificateur d'objectivité.
Les classificateurs sont basés sur l'ensemble des mots uniques, des n-grammes ou des unités lexicales qui sont retrouvés manuellement et qui montrent une forte association de subjectivité. Ensuite ils utilisent les phrases retrouvées à l'issue de l'apprentissage pour les réintroduire dans le classificateur pour l'amélioration de l'étiquetage des phrases subjectives et objectives. Ce processus peut être répété plusieurs fois, ce qui augmentera à chaque fois la précision mais par conséquent réduira la valeur de rappel. La précision obtenue par la méthode de Riloff et Wiebe est inférieure à 90%, mais le rappel est de 40%.

3.7 Conclusion

Dans ce chapitre nous avons présenté l'utilisation des techniques de domaines de catégorisation du texte et d'apprentissage automatique pour les besoins d'analyse des sentiments. Comme nous pouvons le constater, il y a deux grands axes dans le domaine de l'*Opinion Mining* : le premier se base sur la détection de la subjectivité, tandis que le deuxième repose sur la détection de la polarité des phrases.

Dans ce chapitre nous avons présenté sommairement les différents travaux de recherches dans le domaine d'analyse des sentiments. Ce domaine de recherche est utilisable dans d'autres problématiques comme la détection du spam, l'analyse des dépêches de presse, l'analyse des dépêches politiques, l'analyse des dépêches médicales, la génération automatique de réponses ou la production d'un résumé à partir d'un texte. Nous nous sommes focalisés sur l'utilisation de l'analyse des sentiments pour la notation des critiques cinématographiques.

Dans ce chapitre nous nous sommes concentrés sur l'approche statistique pour l'analyse des sentiments, l'approche linguistique sera exposée dans le chapitre suivant.

Chapitre 4

Analyse linguistique

4.1 Les systèmes de compréhension de textes

La détection et la notation des sentiments peuvent être aussi effectuées par les techniques de *Traitement du Langage Naturel, ang : NLP - Natural Language Processing.* L'extraction d'information consiste à identifier de l'information bien précise d'un texte en langage naturel et à la représenter sous forme structurée [Pazienza (1997)]. C'est une recherche documentaire qui vise à retrouver dans un corpus un ensemble de documents pertinents au regard d'une question [Voorhess (1999)]. Elle consiste à constituer automatiquement une banque de données à partir de textes écrits en langage naturel. Il ne s'agit pas de donner du texte brut à l'utilisateur, mais d'apporter des réponses précises aux questions qu'il pose, par le remplissage d'un formulaire ou d'une base de données.

L'extraction nécessite des lexiques et des grammaires spécialisées. La mise au point de telles ressources est une tâche longue et fastidieuse qui demande, le plus souvent, une expertise du domaine abordé et des connaissances en linguistique informatique. Parmi ces connaissances, nous pouvons citer les techniques de filtrage, de catégorisation de documents et d'extraction d'information.

Au départ, le développement du domaine linguistique concerne les systèmes de compréhension traditionnels. La compréhension de textes est un domaine qui est exploré depuis le début du Traitement Automatique des Langues [Sabach (2001)]. Dans les années 70, sont apparus les systèmes " KWIC " qui effectuent la recherche statistique des

mots les plus significatifs [Salton (1983)]. Dans les années 80, des systèmes plus perfectionnés pour l'interrogation de bases de données en langage naturel ont vu le jour. L'exemple d'un de ces systèmes est le système "Lunar". Grâce à ce dernier, les géologues pouvaient interroger en anglais la base des minéraux collectés sur la lune après le retour des missions Apollo [Woods (1973)].

Les systèmes de compréhension de texte ont, pour la plupart, été conçus comme des systèmes génériques de compréhension, mais ils se sont révélés peu utilisables dans des applications réelles. La compréhension est vue comme une transduction qui transforme une structure linéaire. Cela signifie que le texte (i.e. la structure linéaire) est transformé en une représentation logico-conceptuelle intermédiaire. L'objectif final est ensuite de réaliser des inférences sur ces représentations dans le but d'effectuer différents traitements, par exemple répondre à des questions.

Pour comprendre l'ensemble du texte il faut effectuer l'analyse syntaxique et l'analyse sémantique. L'analyse syntaxique est la plus large possible à cause des ambiguïtés. L'analyse sémantique vise à produire une structure représentant le plus fidèlement possible l'ensemble de la phrase, avec ses nuances et sa complexité, puis à intégrer l'ensemble des structures produites en une structure textuelle. A la fin, nous obtenons une représentation logico-conceptuelle du texte. La représentation sémantique varie d'un système à l'autre. Nous pouvons voir dans le système " Core Language Engine " des formes dites logiques inspirées en partie de la grammaire de Montague [Alshawi (1992)]. Dans le système " Kalipsos ", la représentation sémantique est effectuée par les graphes conceptuels [Sowa (1984)] alors que le système " Acord " possède des structures de représentation discursive [Kamp (1981)]. Les structures sémantico-conceptuelles peuvent être plus ou moins larges, riches et complexes, plus ou moins ambiguës.

L'adaptation de ces systèmes pose le problème classique de la réutilisation des systèmes et des bases de connaissances qu'ils intègrent. L'adaptation d'une nouvelle tâche à un nouveau domaine nécessite la reconstruction d'une grande partie des bases de connaissances notamment le lexique sémantique.

4.1.1 Solutions proposées

L'échec relatif des systèmes de compréhension générique est aujourd'hui bien connu. Il faut cependant rappeler que ces systèmes issus des travaux de traitement automatique des langues des années 1980 ont réellement permis d'explorer cette approche générique de la compréhension de texte. Les chercheurs essayent d'avoir des dictionnaires électroniques relativement complets avec la syntaxe et la sémantique.

Ceci a poussé un grand nombre des chercheurs à décrire les langages naturels de la même façon que les langages formels. Maurice Gross entreprit avec son équipe du LADL l'examen exhaustif des phrases simples du français, afin de disposer de données fiables et chiffrées sur lesquelles il serait possible de faire des expériences scientifiques rigoureuses. Pour cela, chaque verbe fut étudié de manière à tester s'il vérifie ou non des propriétés syntaxiques comme le fait d'admettre une proposition complétive en position sujet. 6000 verbes ont été examinés à l'aide d'environ 300 propriétés. Le résultat est que pour 6000 verbes, nous avons environ 15000 emplois différents, qui présentaient un comportement syntaxique différent. Nous nous apercevons que nous ne pouvons pas décrire le français avec des règles générales. La même situation vaut pour toutes les autres langues. Les résultats de cette étude ont été codés dans des matrices appelées tables de lexique-grammaire. La table montre une description précise du comportement syntaxique de chaque verbe du français. L'objectif est d'utiliser toutes les ressources des tables lexique-grammaire pour obtenir un système capable d'analyser n'importe quelle structure de phrase simple. L'unité minimale de sens, d'après Maurice Gross, est la phrase, et non le mot. Le principe est donc d'étudier les transformations que les phrases simples peuvent subir. Les phrases simples ont été indexées par leurs verbes. Pour un verbe nous pouvons avoir plusieurs emplois différents. C'est grâce à des propriétés syntaxiques que nous pouvons distinguer les emplois d'un verbe. Il n'existe pas deux verbes possédant exactement le même comportement syntaxique. Nous ne pouvons donc pas formuler des règles générales qui pourraient expliquer la langue.

4.1.2 Le système UNITEX

Le travail linguistique a été réalisé grâce à l'application Unitex. L'application Unitex a été créée au Laboratoire d'Automatique Documentaire et Linguistique (LADL) sous la direction de M. Maurice Gross. L'auteur de cette application est M. Sébastien Paumier.

L'application Unitex est basée sur les outils linguistiques comme AGLAE [Paumier (2000)] et INTEX [Silberztein (1993)]. Unitex [Paumier (2004)], [Paumier (2003)] est un environnement de développement utilisé pour construire des descriptions formalisées à large couverture des langages naturels et appliqué à des textes de taille importante en temps réel. Les descriptions des langages naturels sont formalisées sous la forme de dictionnaires électroniques, de grammaires représentées par des graphes à nombre fini d'états et de lexiques-grammaires. Il fournit des outils pour décrire la morphologie flexionnelle et dérivationnelle, la variation orthographique et terminologique, le vocabulaire (les mots simples, les mots composés et les expressions figées), les phénomènes semi-figés à la limite entre lexique et syntaxe (grammaires locales, description des accords) et la syntaxe. Unitex permet de traiter en temps réel des textes de plusieurs méga-octets pour l'indexation de motifs morphosyntaxiques, la recherche d'expressions figées ou semi-figées, la production de concordances et l'étude statistique des résultats [Paumier (2000), Dziczkowski (2005)].

Unitex représente tous les objets traités dont les textes, les dictionnaires et les grammaires par les transducteurs à nombre fini d'états [*Voir Annexe*]. Un transducteur à nombre fini d'états est un graphe qui represente un ensemble de séquences en entrée, et leur associe des séquences produites en sortie. Par exemple le dictionnaire représente des séquences de lettres et produit les informations lexicales associées ; le transducteur d'un texte représente les séquences de mots qui forment chaque phrase et leur associe des informations lexicales ou syntaxiques (les marques linguistique produites par les différentes analyses). Unitex ramène toute l'opération à un ensemble limite d'opérations sur des transducteurs. Par exemple, appliquer des dictionnaires à un texte consiste à construire l'union des transducteurs de chaque dictionnaire et à construire l'union de

ce transducteur avec le transducteur du texte.

Les automates à états finis sont un cas particulier de transducteurs à nombre fini d'états, ils produisent l'information binaire à la sortie : séquence reconnue ou séquence non reconnue. Ils sont utilisés pour rechercher des séquences dans le texte, ils créent donc la liste de toute la séquence reconnue.

4.1.3 Les dictionnaires

Les dictionnaires électroniques d'*Unitex* utilisent le formalisme des DELA (Dictionnaires Electronique du LADL). Les dictionnaires électroniques décrivent les mots simples et les mots composés d'une langue en leur associant un lemme avec une série de codes grammaticaux, sémantiques et flexionnels. Les dictionnaires ont été élaborés par des équipes de linguistes pour plusieurs langues comme le français, l'anglais, le grec, l'italien, l'espagnol, l'allemand, le thaïlandais, le coréen, le norvégien, le portugais. Nous pouvons distinguer deux sortes de dictionnaires électroniques. Les premiers, les dictionnaires de formes fléchies, les types les plus utilisés, sont le DELAF (DELA de forme fléchie) et le DELACF (DELA de forme composée fléchie). Les programmes d'Unitex ne font pas de distinction entre les dictionnaires de forme simple et composée. Les deuxièmes, les dictionnaires de forme non fléchie, sont le DELAS (DELA de forme simple) et le DELAC (DELA de forme composée).

Pour associer un lemme et une information linguistique à un dictionnaire de formes non fléchies, nous utilisons des transducteurs lexicaux. Par exemple, dans le cas où il faut reconnaître 5000 chiffres romains, il est impossible de construire un dictionnaire DELAF. Il est plus simple de construire le transducteur correspondant à l'aide de 5 graphes simples [Silberztein (1993)].
Maintenant, nous allons présenter le format des dictionnaires. Le dictionnaire DELAF d'une langue contient toutes les formes fléchies de la langue et les associe au lemme. En plus il existe le code morphosyntaxique et éventuellement les codes syntaxiques, sémantiques et flexionnels. Voici l'exemple d'entrée du DELAF :

avions,avion.N+CONC :mp

La première ligne représente le fait que la forme "avions" est associée au lemme "avion". Puis la lettre N signifie que c'est un nom. CONC nous apporte l'information que la classe distributionnelle est Concret. A la fin, nous avons un code flexionnel :mp qui représente le masculin pluriel. Nous pouvons trouver plus de détails dans [Paumier (2004)]. Le dictionnaire DELAF contient toutes les formes fléchies du français, donc environ 680.000 formes fléchies différentes. Pour le dictionnaire DELACF, la seule différence est que la forme fléchie et le lemme peuvent contenir des séquences de lettres et de séparateurs. Voici un exemple d'entrée du dictionnaire DELACF :

Pomme de terre, pomme de terre,N+NDN+Conc :fs

Le DELACF contient 250.000 formes de noms composés, 8.000 adverbes figés, 15.000 formes figées utilisées avec le verbe être et 1 600 conjonctions de subordination.

Le format des DELAS est similaire à celui des DELAF. La différence est qu'il ne donne qu'une seule forme canonique suivie de codes grammaticaux et sémantiques. Voici un exemple :

Cheval,N4+Anl

Le premier code est interprété par le programme de flexion comme le nom de la grammaire à utiliser pour fléchir l'entrée. L'entrée de l'exemple ci-dessus indique que le mot cheval doit être fléchi avec une grammaire nommée N4 [Paumier (2004)].

4.1.4 Le réseaux des transitions récursives

Les grammaires sont des représentations de phénomènes linguistiques par des transitions récursives (RTN), un formalisme proche de celui des automates à états finis. De nombreuses études ont mis en évidence l'adéquation des automates aux problèmes linguistiques. Généralement une grammaire représente des séquences de mots et produit des informations linguistiques comme par exemple des informations sur la structure

syntaxique. Un dictionnaire représente des séquences de lettres et produit les informations lexicales associées. Le transducteur d'un texte représente les séquences de mots qui forment chaque phrase et leur associe des informations lexicales ou syntaxiques - des résultats produits par les différentes analyses. Les grammaires sont représentées au moyen de graphes que l'utilisateur peut créer et mettre à jour. L'application des dictionnaires à un texte consiste à construire l'union des transducteurs de chaque dictionnaire, et à construire l'union de ce transducteur avec le transducteur du texte.

Les corpus de texte sont représentés par des automates, dans lesquels chaque chemin correspond à une analyse lexicale. Les phénomènes linguistiques sont représentés par la grammaire locale qui est traduite en automates à états finis afin d'être aisément confrontés avec les corpus de texte.

Une grammaire locale [Gross (1997)] est une représentation par automate de structures linguistique difficilement formalisables dans des tables de lexique-grammaire ou dans des dictionnaires électroniques. Les grammaires locales, représentées sous forme de graphes, décrivent des éléments qui relèvent d'un même domaine syntaxique ou sémantique. Les descriptions linguistiques décrites sous la forme de grammaires locales sont utilisées pour une grande variété de traitements automatiques appliqués sur les corpus de texte. Ainsi, différentes méthodes de désambiguïsation lexicale ont été développées pour mettre en oeuvre des contraintes grammaticales décrites à l'aide de ce type de graphe.

Voici l'exemple de grammaire locale [Figure 4.1] qui a trouvé une utilisation dans le système de filtrage d'informations CORAIL [Balvet (2001)]. CORAIL est un moteur de filtrage d'informations intégrant des contraintes d'ordre linguistique par le biais de dictionnaires, de grammaires locales et de table du lexique-grammaire. Les grammaires locales présentent des propriétés intéressantes pour une application au filtrage d'informations.

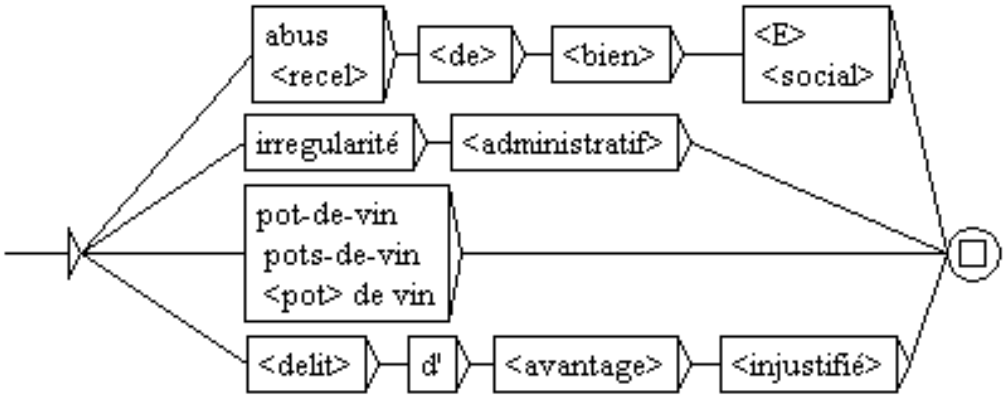


Figure 4.1: L'exemple d'une grammaire locale. -

4.1.5 Les tables de lexique-grammaire

Cette table de lexique-grammaire, que nous avons précédemment définie est une matrice décrivant les propriétés syntaxiques de tous les verbes simples du français. Chaque mot ayant un comportement quasi unique, les tables donnent la grammaire de chaque élément de lexique, d'où le nom de lexique-grammaire. Nous pouvons grâce à Unitex construire des grammaires à partir de telles tables. Le lexique-grammaire de Maurice Gross est une description systématique des propriétés syntaxiques et sémantiques des éléments syntaxiques du français, c'est à dire les verbes, les noms prédicatifs et les adjectifs. Il est organisé en groupes de tables, qui sont associés à une catégorie syntaxique donnée comme verbes pleins, verbes supports, noms, etc. Une table correspond à une construction syntaxique particulière et rassemble tous les mots qui entrent dans cette construction. Par exemple, la table des verbes contient tous les verbes qui admettent en plus d'un sujet un complément infinitif mais pas un complément qui soit une complétive finie ou non. Une table [Figure 4.2] est divisée en ligne selon les mots qu'elle contient et en colonnes selon les propriétés syntaxiques ou sémantiques qui s'appliquent à ces mots et leurs arguments. A l'intersection d'une ligne et d'une colonne, un signe + ou - indique que la propriété indiquée en entête de la colonne s'applique positivement ou négativement au mot placé en entête de la ligne. Cette propriété est soit un ajout d'information sur le mot ou un de ses arguments, soit une transformation du cadre de sous-catégorisation de base associée à la table. Actuellement le lexique-grammaire

est surtout développé pour les verbes et les locutions prédicatives. Ce lexique contient 15.000 entrées de verbes simples. En outre, 25.000 locutions prédicatives ont été décrites, de même que 20.000 locutions construites avec être ou avoir [Gardent *et al.* (2005)].

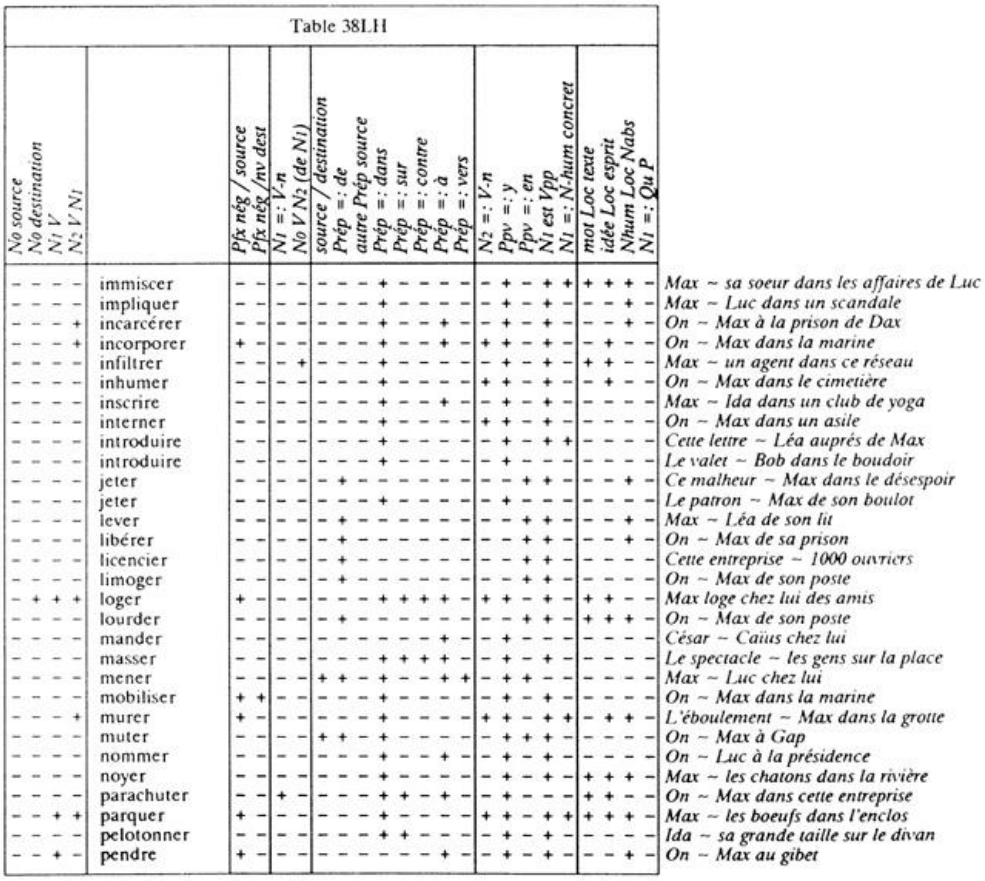


Table 38LH

N0 source	N0 destination	N1 V	N2 V N1		Pfx nég / source	Pfx nég /nv dest	N1 = V-n	N0 V N2 (de N1)	source / destination	Prép = : de	autre Prép source	Prép = : dans	Prép = : sur	Prép = : contre	Prép = : à	Prép = : vers	N2 = : V-n	Ppv = : y	Ppv = : en	N1 est Vpp	N1 = : N-hum concret	mat Loc texte	idée Loc esprit	Nhum Loc Nabs	N1 = : Qu P	
–	–	–	–	immiscer	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	+	+	+	+	+	–	*Max ~ sa soeur dans les affaires de Luc*
–	–	–	–	impliquer	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	+	–	–	–	+	–	*Max ~ Luc dans un scandale*
–	–	–	+	incarcérer	–	–	–	–	–	–	–	+	–	–	+	–	–	+	–	+	–	–	–	+	–	*On ~ Max à la prison de Dax*
–	–	–	+	incorporer	+	–	–	–	–	–	–	+	–	–	+	–	+	+	–	+	–	–	+	–	–	*On ~ Max dans la marine*
–	–	–	–	infiltrer	–	–	–	+	–	–	–	+	–	–	–	–	–	+	–	+	–	+	+	–	–	*Max ~ un agent dans ce réseau*
–	–	–	–	inhumer	–	–	–	–	–	–	–	+	–	–	–	–	+	+	–	+	–	–	+	–	–	*On ~ Max dans le cimetière*
–	–	–	–	inscrire	–	–	–	–	–	–	–	+	–	–	+	–	–	+	–	+	–	–	–	–	–	*Max ~ Ida dans un club de yoga*
–	–	–	–	interner	–	–	–	–	–	–	–	+	–	–	–	–	+	+	–	+	–	–	–	–	–	*On ~ Max dans un asile*
–	–	–	–	introduire	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	+	+	–	–	–	–	*Cette lettre ~ Léa auprés de Max*
–	–	–	–	introduire	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	–	–	–	–	–	–	*Le valet ~ Bob dans le boudoir*
–	–	–	–	jeter	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	–	–	+	–	*Ce malheur ~ Max dans le désespoir*
–	–	–	–	jeter	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	–	–	–	–	–	–	*Le patron ~ Max de son boulot*
–	–	–	–	lever	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	–	–	+	–	*Max ~ Léa de son lit*
–	–	–	–	libérer	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	–	–	+	–	*On ~ Max de sa prison*
–	–	–	–	licencier	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	–	–	–	–	*Cette entreprise ~ 1000 ouvriers*
–	–	–	–	limoger	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	–	–	–	–	*On ~ Max de son poste*
–	+	+	+	loger	+	–	–	–	–	–	–	+	+	+	+	–	+	+	–	+	–	+	+	–	–	*Max loge chez lui des amis*
–	–	–	–	lourder	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	+	–	+	+	+	–	*On ~ Max de son poste*
–	–	–	–	mander	–	–	–	–	–	–	–	–	–	–	+	–	–	+	–	–	–	–	–	–	–	*César ~ Caius chez lui*
–	–	–	–	masser	–	–	–	–	–	–	–	+	+	+	+	–	–	+	–	+	–	–	–	–	–	*Le spectacle ~ les gens sur la place*
–	–	–	–	mener	–	–	–	–	+	+	–	+	–	–	+	+	–	+	+	–	–	–	–	–	–	*Max ~ Luc chez lui*
–	–	–	–	mobiliser	+	+	–	–	–	–	–	+	–	–	–	–	–	+	–	+	–	–	–	–	–	*On ~ Max dans la marine*
–	–	–	+	murer	+	–	–	–	–	–	–	+	–	–	–	–	+	+	–	+	+	–	+	+	–	*L'éboulement ~ Max dans la grotte*
–	–	–	–	muter	–	–	–	–	+	+	–	+	–	–	–	–	–	+	+	+	–	–	–	–	–	*On ~ Max à Gap*
–	–	–	–	nommer	–	–	–	–	–	–	–	+	–	–	+	–	–	+	–	+	–	–	–	–	–	*On ~ Luc à la présidence*
–	–	–	–	noyer	–	–	–	–	–	–	–	+	–	–	–	–	–	+	–	+	–	+	+	+	–	*Max ~ les chatons dans la rivière*
–	–	–	–	parachuter	–	–	+	–	–	–	–	+	+	–	+	–	–	+	–	–	–	+	+	–	–	*On ~ Max dans cette entreprise*
–	–	+	+	parquer	+	–	–	–	–	–	–	+	–	–	–	–	+	+	–	+	+	+	+	+	–	*Max ~ les boeufs dans l'enclos*
–	–	–	–	pelotonner	–	–	–	–	–	–	–	+	+	–	–	–	–	+	–	+	–	–	–	–	–	*Ida ~ sa grande taille sur le divan*
–	–	+	–	pendre	+	–	–	–	–	–	–	–	–	–	+	–	–	+	–	+	–	–	–	+	–	*On ~ Max au gibet*

Figure 4.2: Echantillon de la table 38LH du lexique grammaire -

Bien qu'il soit aujourd'hui clair que le lexique est une composante essentielle des systèmes TAL, les ressources disponibles sont rares. Pour l'anglais, COMLEX Syntax [17] contient une information détaillée de 38.000 mots dont 6.000 verbes. VerbNet, lui, décrit 4.000 sens verbaux à partir de 191 classes sémantiques et 52 cadres syntaxiques.

Pour le français, plusieurs lexiques sont disponibles mais la plupart concernent la morphologie plutôt que la syntaxe. Ainsi le lexique LEFFF (Lexique des Formes Fléchies du Français) contient 5.000 verbes et 200.000 formes fléchies mais l'information associée est purement flexionnelle. Comme nous l'avons précisé, le lexique-grammaire de Gross contient une information détaillée et exhaustive et a été numérisé par le Laboratoire d'Automatique Documentaire et Linguistique (LADL). Il est maintenant partiellement disponible sous une licence LGPL-LR. Tout ceci facilite la constitution d'une ressource lexicale appropriée au TAL.

4.2 Extraction automatique d'information

Les premiérs essais pour effectuer de manière semi-automatique l'extraction se fondaient le plus souvent sur un large corpus annoté qui servait de base de connaissances pour l'apprentissage. C'est en particulier la stratégie adoptée par Riloff [Riloff (1993)] dans le système *Autoslog* qui constitue une des premières références dans le domaine. Autoslog fonde son analyse sur un corpus où les entités du domaine ont été préalablement annotées. Un analyseur syntaxique est ensuite appliqué sur le texte. Le résultat est un texte annoté où les principaux syntagmes (ensemble de lemmes) ont été identifiés, ainsi que des relations fonctionnelles entre eux (sujet-verbe, verbe-objet). Des schémas très généraux permettent alors d'extraire automatiquement un certain nombre de séquences où les entités sont mises en relation avec un élément discriminant, afin de typer de manière adéquate l'entité concernée. Des erreurs peuvent survenir à cause d'une mauvaise analyse syntaxique. L'apprentissage a été effectué sur un corpus *MUC-4* composée de 772 textes annotés et l'évaluation sur 100 autre textes.

Une autre approche était proposée par Soderland *et al.* [Soderland *et al.* (1995)]. Cette approche décrit un outil d'apprentissage fondé sur un corpus de 335 textes annotés. L'outil offre des processus de généralisation permettant de regrouper des entités en simplifiant la syntaxe. Un mot peut être remplacé par son type syntaxique ou sémantique ou même être éliminé. L'objectif était de limiter le temps de validation pour l'utilisateur et d'offrir une meilleure couverture du système. L'outil d'apprentissage est

nommé *Crystal*.

Une approche était proposée par Freitag [Freitag (1998)], qui propose de combiner des méthodes "naïf Bayes" avec une analyse relationnelle. Il utilise la méthode "naïf Bayes" pour calculer un score pour des mots susceptibles de remplir un champ donné, ensuite par analyse relationnelle il repère les relations entre ces mots. Le système commence avec un ensemble de règles vides, puis induit des règles à partir du corpus à chaque nouvel exemple traité. L'utilisation de ressources linguistiques entraine une meilleure précision mais diminue le rappel.

4.3 Conclusion

Dans ce chapitre nous avons présenté les techniques linguistiques que nous avons utilisées dans nos travaux. Nous verrons que le traitement linguistique donne des très bons résultats et cette approche ne devra pas être négligée dans le domaine d'analyse des sentiments [Dziczkowski & Wegrzyn-Wolska (2007b), Dziczkowski & Wegrzyn-Wolska (2008a), Dziczkowski & Wegrzyn-Wolska (2008b)]. C'est une approche novatrice dans le domaine de l'Analyse des Sentiments puisque l'utilisation de traitement purement linguistique n'a pas été abordée dans la littérature du domaine de l'*Opinion Mining*.

Afin d'effectuer une analyse linguistique, en plus des ressources linguistiques, le traitement demande une analyse manuelle coûteuse au niveau temps (par exemple pour la création des classes sémantiques ou des grammaires locales). C'est un des plus grands inconvénients du traitement linguistique. Pourtant ce traitement a un grand nombre d'avantages qui seront être précisés dans les chapitres suivants (*Chapitre 6* et *Chapitre 7*).

Nous avons également présenté les ressources linguistiques utilisées dans le domaine du Traitement Automatique du Langage Naturel *(TALN)*. Nous avons utilisé ces ressources linguistiques pour nos travaux de recherche, afin de pouvoir exprimer l'intensité de l'opinion. Parmi ces ressources, nous pouvant citer les grammaires locales, les tables de lexique-grammaires et les dictionnaires. Dans notre recherche nous avons utilisé une

application *Unitex* qui permet d'introduire toutes ces ressources linguistiques et qui permet de les appliquer au corpus des textes. L'implémentation des ressources linguistiques et l'utilisation des techniques linguistiques dans l'application *Unitex* sont présentées en détail dans la référence suivante : [Dziczkowski (2005), Dziczkowski & Wegrzyn-Wolska (2007a)].

Avec ce chapitre, nous avons finalisé la présentation des techniques de catégorisation et de compréhension du texte existantes, ainsi que l'utilisation de ces techniques dans le domaine de l'Analyse des Sentiments. Dans le *Chapitre 5*, nous allons décrire l'architecture du système développé. Notre système est un système autonome d'exploration des opinions exprimées dans les critiques cinématographiques. Dans les chapitres suivants, nous décrirons les classificateurs développés pour la notation des opinions et nous présenterons les tests effectués et la comparaison entre les classificateurs basés sur les techniques statistiques et linguistiques.

Chapitre 5

Système mis en oeuvre pour la notation d'opinion

5.1 Les besoins commerciaux

Avec la croissance du Web, le e-commerce est devenu très populaire. Beaucoup de sites Web offrent la possibilité de faire de la vente en ligne et donnent également la possibilité de mettre son propre avis en ligne sur des objets, des personnes, des produits et notamment des films. Les gens aiment généralement vérifier les recommandations des autres utilisateurs avant de se faire leurs propres opinions ou de faire leur choix. Les prédictions en ligne sont donc devenues très utiles pour les clients. Pour prédire le choix potentiel, des systèmes de recommandation *ang : Recommender System, RS* ont été créés. Un RS permet de prédire un choix sans aucune connaissance personnelle des alternatives. Les Algorithmes des moteurs de prédiction sont basés sur l'expérience et l'avis des autres utilisateurs. Il est utile de trouver des recommandations de personnes qui ont les mêmes goûts que nous, qui sont familiers avec le problème, ou qui sont des experts reconnus [Tarveen & Littman (2001)].

Un RS fournit des correspondances entre les utilisateurs qui ont le même profil. Un nouvel utilisateur doit donc créer son profil. Le moteur de prédiction proposera alors un nouveau choix limité basé sur le goût des autres utilisateurs qui ont le même profil. La crédibilité du résultat du RS ne peut pas reposer sur des raisons commerciales, car cela pourrait rendre les gens méfiants. L'efficacité d'un tel système dépend de la qualité et

de la quantité des données. Pour cette raison, le système présenté dans cette thèse fournit aux utilisateurs des profils qui sont nécessaires aux algorithmes des moteurs cognitifs.

L'objectif principal du système développé est de recueillir une énorme base de critiques cinématographiques avec leurs auteurs, et d'associer automatiquement les notes qui expriment des sentiments de la personne qui a écrit la critique. Le résultat de ce traitement est la création de la base de données qui contient les profils des utilisateurs. Notre système est basé sur la représentation statistique et sémantique des documents. Notre travail est composé en partie de l'extraction et du filtrage de l'opinion du texte, et en partie de la notation des sentiments des phrases subjectives.

Le sujet de cette thèse m'a été proposé par l'équipe de chercheurs de l'entreprise *Criteo* qui développe un moteur de prédiction pour les critiques cinématographiques. Leurs besoins concernent la création d'un système autonome pour la détection et la notation automatique de la critique cinématographique. Nous avons étudié et développé trois différentes méthodes de notation de l'opinion, nous avons effectué une étude comparative des trois méthodes et nous avons présenté les avantages et les inconvénients de chacune d'elles. Nous présentons aussi un classificateur final pour combiner les différents résultats obtenus. Le système développé prépare la base de données d'entrées pour le système de prédictions. Notre système réalise les tâches suivantes :

- Recherche automatique d'une critique cinématographique via internet,
- Attribution automatique d'une note, allant de 1 à 5, par rapport aux sentiments décrits dans la critique,
- Publication des résultats en générant les profils complets des utilisateurs.

5.2 Architecture du système

Notre système possède une architecture modulaire. Ses tâches principales sont les suivantes : recherche et collecte des critiques sur Internet, attribution d'une note aux critiques et présentation des résultats.

Chaque tâche est réalisée par un module spécialisé [Figure 5.1].

En premier lien, pour la partie de la notation de l'opinion, nous avons développé trois méthodes différentes pour l'attribution d'une note à une critique. Ces méthodes sont basées sur les différentes approches de la classification du document. En deuxième lien, nous avons développé pour chaque méthode un classificateur qui assigne séparément la note [Dziczkowski & Wegrzyn-Wolska (2008b)]. Nous avons, par conséquent obtenu trois notes pour chaque critique peuvant être différentes. Nous avons finalement utilisé un autre classificateur qui assigne la note finale à la critique cinématographique, fondée uniquement sur les trois notes attribuées antérieurement pendant le processus de classification [Dziczkowski & Wegrzyn-Wolska (2008a)]. Pour le calcul de la note finale, nous avons utilisé les valeurs des trois notes obtenues précédemment avec leurs probabilités.

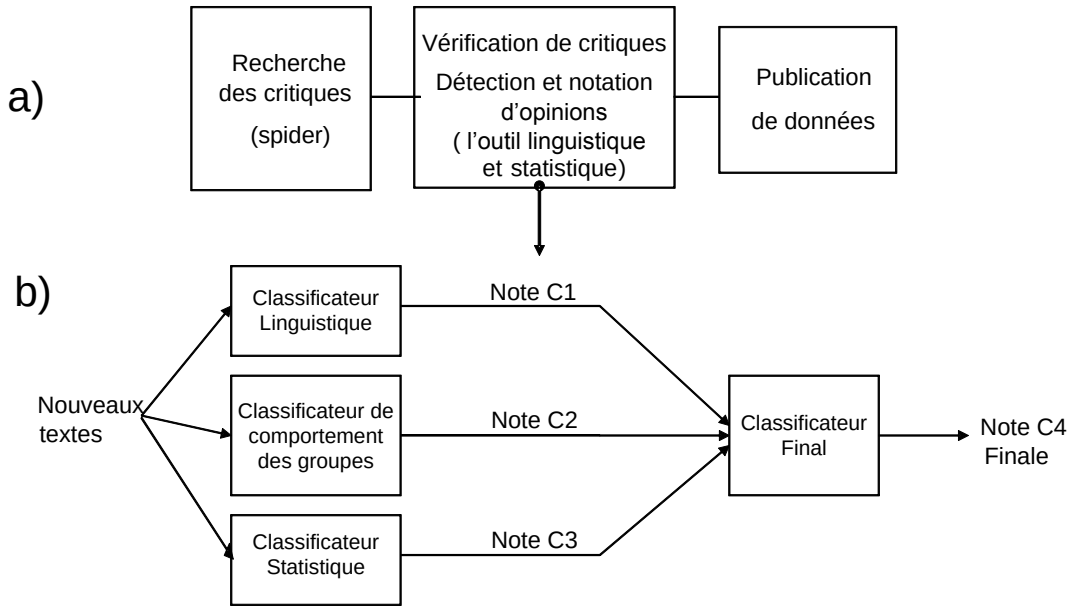


Figure 5.1: Architecture générale du système - a. Les trois modules principaux, b. Notation des critiques cinématographiques

5.3 Recherche des critiques

Internet est une importante source d'informations en ce qui concerne les avis sur les films et les critiques cinématographique. Nous pouvons trouver des fiches descriptives pour tous les films, et grâce au développement de nombreuse sites concernant la cinématographie, la communauté des internautes peut exprimer ses points de vue sur toutes les oeuvres. Ainsi, de nombreux sites fournissent des moyens d'expressions et de notation des films à leurs membres. Ces avis sont par principe subjectifs. Au final, nous obtenons une base de données des opinions subjectives très volumineuse. Cependant, les sites et les forums sont créés de manière différente, ces sources d'informations sont donc rarement interoperable et un traitement global de l'information pour un film donné est difficile.

L'outil développé pour la recherche des critiques cinématographiques a pour but de fusionner en une seule et même base de données toutes ces sources d'informations concernant les critiques des films. Ces informations alors regroupées et ordonnées sont accessibles de manière simple via un moteur de recherche classique ou via un accès direct à la base de données, sans l'obligation de parcourir un site extérieur, ni de faire une manipulation humaine sur les données.

Il n'y a actuellement aucun système connu qui permet de récupérer des critiques de films venant de plusieurs sites. La seule solution pour le faire serait de parcourir manuellement les différentes pages, de les indexer en cherchant les différentes informations et de parcourir ensuite les différentes sources d'informations une à une. Rare sont les entreprises mettant à disposition leurs bases de données, c'est pour cela que ce type d'application n'existe pas sur ce domaine particulier.

L'avantage de l'application développée est de permetre la création d'une base de données unique à partir de sources multiples, permettant ainsi une recherche facile mais aussi plus détaillée car plus complète. Le processus de la collection des critiques est effectué de manière automatique et ne nécessite aucune modification humaine qui serait coûteuse en temps.

L'inconvénient de cette application se résume à la nécessité d'avoir un grand espace de stockage car, celle-ci récupère de nombreuses critiques par film (la base de donnée a une taille importante, de plusieurs centaines de mégaoctets). De ce fait, le traitement, bien qu'automatique, prend beaucoup du temps.

5.3.1 Les étapes de collecte des critiques

Nous avons développé, pour le premier module de notre système, une application web basée sur une base de données et composée d'un formulaire de recherche permettant d'obtenir une liste de critique en fonction d'un titre de film saisi par l'utilisateur. Pour toute recherche, nous devons constituer l'ensemble des informations telles que les critiques des films, les informations sur les auteurs des critiques ainsi que les informations supplémentaires comme la note de film, si elle est attribuée ou les détails concernant les films et leur auteurs.

Nous avons effectué la recherche des critiques, sur le WEB, de trois façons générales. La recherche est effectuée sur :

- La liste des URL prédéfinies,
- Le format structuré RSS,
- Les forums de discussion.

Le choix des méthodes pour la collection des critiques est déterminé par les services existants. Il est effectué de manière à récupérer le plus grand nombre de critiques de différentes sources en récupérant les informations sur les auteurs. La première méthode consiste en une recherche sur les sites populaires et connus. Il existe plusieurs sites sur lesquels les internautes peuvent exprimer leur avis et effectuer des notations de films, à titre d'exemple : Amazon(www.amazon.com) et IMDb(www.imdb.com). Puisque nous avons la possibilité de connaitre le format du site nous pouvons facilement extraire des informations (la critique, l'auteur, la note ...). La deuxième méthode de la recherche est effectuée grâce au RSS. Dans ce cas, les informations sont structurées, ce qui facilite la recherche des textes des opinions. La dernière méthode parcourt les forums en regardant si le moteur qui a créé le forum est connu (par exemple php forum, phpfb). Si l'analyse du moteur a dejà été effectuée auparavant, nous aurons la possibilité d'en extraire les informations plus facilement. Dans notre recherche nous nous intéressons aux critiques

issues des forums de cinématographiques.

Après ce traitement, nous obtenons la liste des films indexés dans la base de données avec les critiques correspondantes. L'utilisateur se rend sur la page principale située à la racine de l'application et peut obtenir le listing des films enregistrés classés par ordre alphabétique avec la date d'indexation et le nombre de critiques correspondantes. Pour ajouter les critiques d'un nouveau film, il suffit de saisir le titre et le processus est lancé. Les résultats sont renouvelés à partir des différentes sources.

5.3.2 Fonctionnement de l'application

Le traitement doit se faire suivant plusieurs étapes :

1. Vérification des titres : l'application doit pouvoir vérifier l'intégrité du titre et vérifier que celui-ci est valide. Comme l'application effectue une recherche sur plusieurs sites, il faut être sûr que le titre demandé soit le titre original pour pouvoir effectuer une recherche cohérente sur l'ensemble des sources.
2. Détermination des sources d'informations prédéfinies : l'application récupère les informations à partir de sites web : URL prédéfinis (Amazon, IMDb) et les forums de discussion (Nntp).
3. Consultation des sites prédéfinis (URL) : l'application effectue une recherche d'items dans plusieurs sites :
   (a) La base de données d'Amazon : nous pouvons alors obtenir tout d'abord les commentaires de l'équipe Amazon (*editorial reviews*). Ensuite nous passons à la recherche de critiques par les consommateurs et les clients Amazon (*customer review*). Nous obtenons enfin pour chaque critique tous les détails tels que les ratings, les notes des autres utilisateurs et bien sûr les commentaires et titres [Figure 5.2].
   (b) IMDb : l'indexation des critiques du site IMDb se fait en utilisant le "parsing" des données. Ainsi, le seul accès disponible pour avoir les commentaires est d'utiliser le site IMDb, d'en télécharger le contenu et de récupérer les différentes informations. Cette méthode comporte des incertitudes sur la fiabilité des données récupérées, celles-ci étant très dépendantes de la structure de

la page, chose que nous ne pouvons contrôler. Nous utilisons un moteur de recherche pour obtenir l'id IMDb correspondant au titre de film. Une fois l'id de la fiche récupéré, nous parcourons les différentes pages de commentaires pour télécharger le code html de chaque page [Figure 5.3].

(c) Nntp : les newsgroups sont des forums reliés en réseau. Nous avons utilisé le site IMDb qui les enregistre et les ordonne par films ou par utilisateurs. L'application récupère l'id IMDb et nous obtenons ensuite la liste des différentes critiques des newsgroups pour le film concerné. Les pages sont alors téléchargées puis "parsées" pour récupérer les différentes informations [Figure 5.4].

4. Enregistrement des données : l'enregistrement de toutes les informations est fait dans une base de données. Pour le besoin de notre recherche nous nous sommes arrêté à une base de données de 300MB, ce qui correspond à deux cent mille entrées [Figure 5.5].

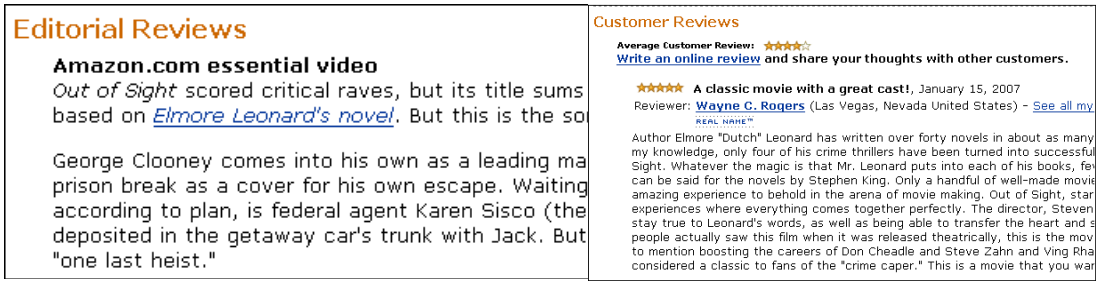


Figure 5.2: Indexation d'amazon - les commentaires de l'équipe d'Amazon et les critiques des utilisateurs

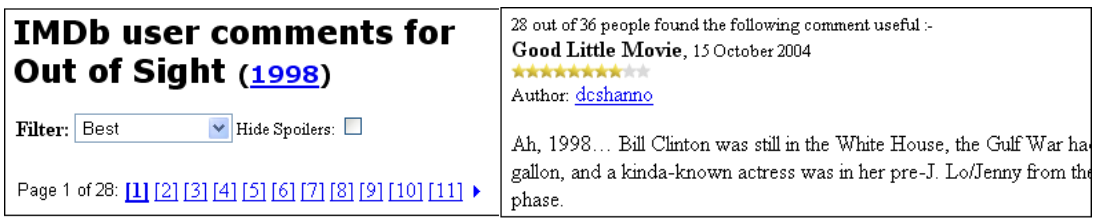
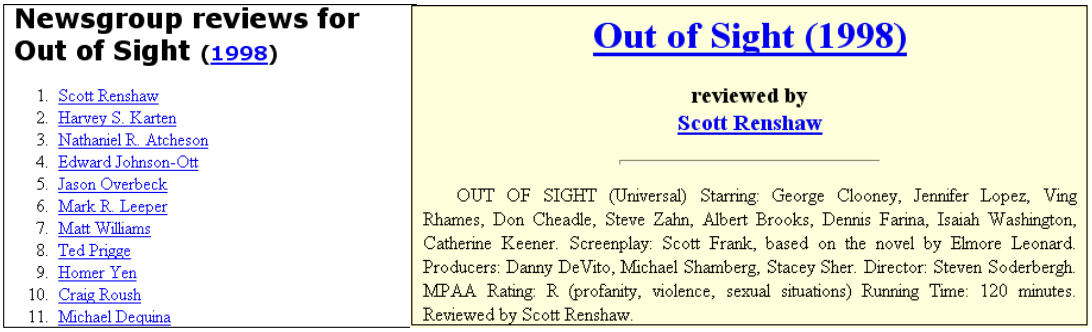


Figure 5.3: Indexation IMDb - le format du site de l'IMDb



Newsgroup reviews for Out of Sight (1998)

1. Scott Renshaw
2. Harvey S. Karten
3. Nathaniel R. Atcheson
4. Edward Johnson-Ott
5. Jason Overbeck
6. Mark R. Leeper
7. Matt Williams
8. Ted Prigge
9. Homer Yen
10. Craig Roush
11. Michael Dequina

Out of Sight (1998)

reviewed by
Scott Renshaw

OUT OF SIGHT (Universal) Starring: George Clooney, Jennifer Lopez, Ving Rhames, Don Cheadle, Steve Zahn, Albert Brooks, Dennis Farina, Isaiah Washington, Catherine Keener. Screenplay: Scott Frank, based on the novel by Elmore Leonard. Producers: Danny DeVito, Michael Shamberg, Stacey Sher. Director: Steven Soderbergh. MPAA Rating: R (profanity, violence, sexual situations) Running Time: 120 minutes. Reviewed by Scott Renshaw.

FIGURE 5.4: Indexation Nntp - le format du site du nntp

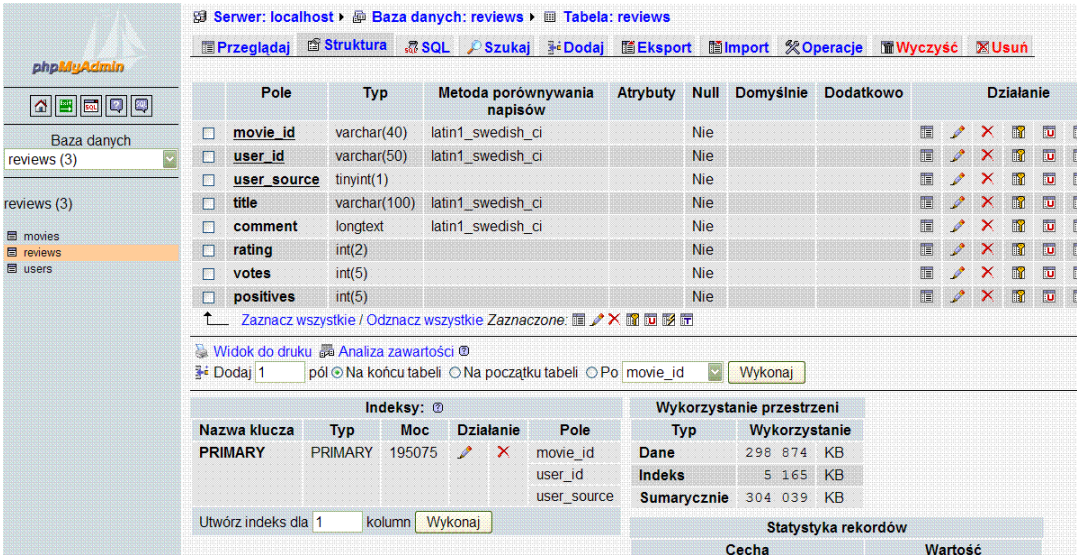


FIGURE 5.5: La base de données - le format de la base de données

Nous avons développé une interface pour l'application qui va permettre de faire l'affichage des données à partir d'un titre de film [Figure 5.6].

FIGURE 5.6: L'interface de l'application - trois fenêtres montrant : la recherche basée sur le titre du film, le sommaire d'un film et la critique cinématographique

5.4 La détection et la notation de l'opinion

Le module de détection et de notation de l'opinion est le principal module de notre système. L'objectif est d'effectuer automatiquement une notation des sentiments exprimés dans les critiques cinématographiques. Pour classer les opinions nous avons utilisé une notation suivant une échelle variant de 1 à 5. Nous avons utilisé trois différentes approches pour effectuer la notation. Nous présentons les classificateurs linguistiques [Dziczkowski & Wegrzyn-Wolska (2007a)] et le classificateur basé sur le comportement des groupes que nous avons proposés. Ces deux approches sont ensuite comparées avec le classificateur "naïf Bayes" et SVM [Figure 5.7].

Le dernier classificateur, appelé classificateur final, est utilisé pour gérer les trois notes récupérées des différentes approches et pour améliorer les résultats finaux.
Nous présentons la description détaillée de ces méthodes dans le *Chapitre 6*.

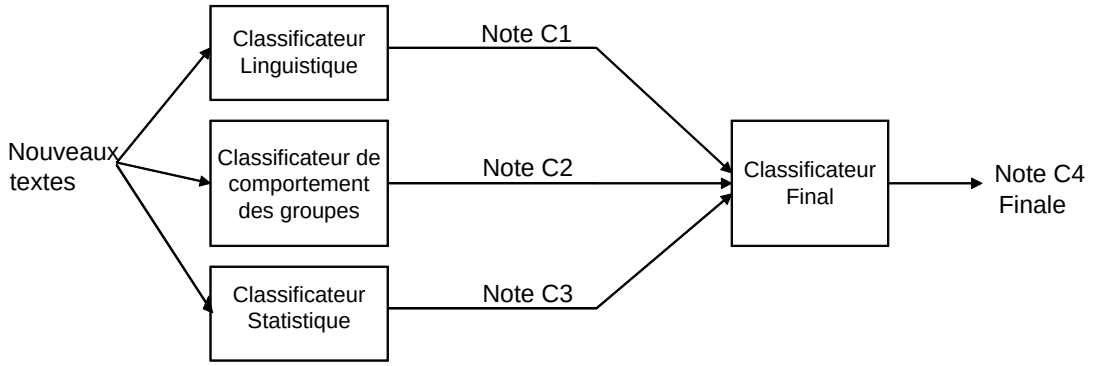


Figure 5.7: Notation de l'opinion - Les trois différents approches pour la notation de l'opinion et un classificateur pour combiner les résultats

5.4.1 Pourquoi une telle architecture ?

Nous avons présenté une architecture composée de plusieurs classificateurs pour effectuer séparément la notation des nouvelles critiques. Nous donnons dans la suite de cette section les raisons qui nous ont poussé à choisir une telle architecture.

Au début de notre recherche, nous avons proposé une première architecture pour la notation des sentiments [Figure 5.8].

Le but général d'une telle architecture est d'estimer une note grâce à un classificateur linguistique et ensuite de l'approuver par un classificateur statistique. Le classificateur linguistique est basé sur une étude linguistique profonde. Il analyse les groupes de critiques qui ont la même note associée (de 1 a 5). Le résultat de ce classificateur est donné par une note. Cette dernière est ensuite utilisée pour choisir l'un des cinq classificateurs statistiques permettant d'approuver la note estimée. L'approche est composée de deux étapes : la première pour l'estimation d'une note effectuée par le classificateur linguistique, et la deuxième pour l'attribution de la note finale approuvée par le classificateur statistique correspondant à la note estimée. Le classificateur statistique choisi par le classificateur linguistique favorise la note estimée par ce dernier [Dziczkowski & Wegrzyn-Wolska (2007b)].

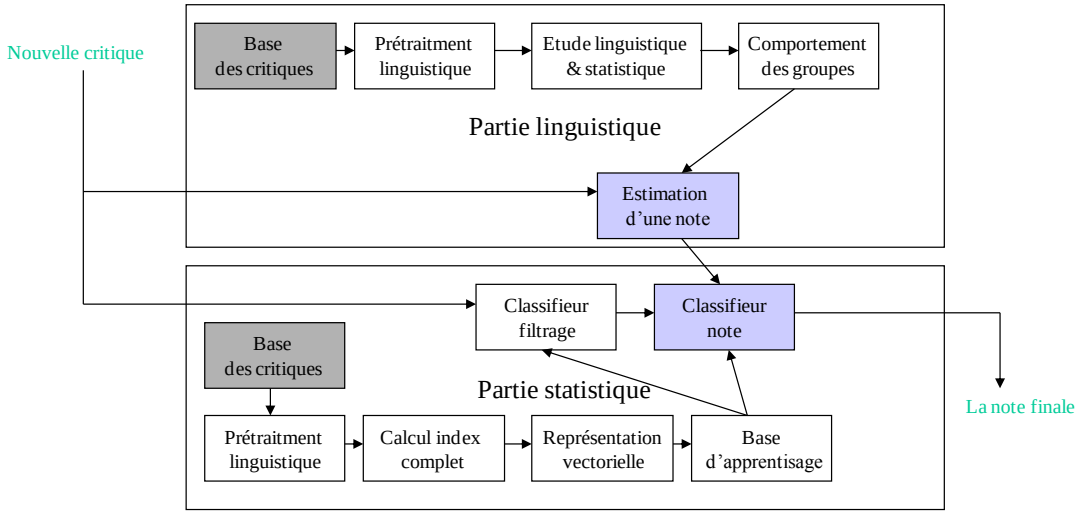


FIGURE 5.8: Architecture séquentielle - Le premier classificateur linguistique est utilisé pour estimer la note qui sera ensuite approuvée par le classificateur statistique

Pour cette première architecture nous avons effectué plusieurs tests. Nous avons remarqué que le passage du premier au deuxième classificateur provoque une perte significative d'informations. Nous avons également constaté que l'architecture n'est pas bien adaptée à nos besoin, car nous ne pouvons pas approuver une note obtenue par les techniques linguistiques avec les techniques statistiques et vice versa.

Nous avons de plus remarqué, à cette étape, que les groupes de critiques ayant la même note associée montrent une forte particularité, ce qui nous ramène à une étude des caractéristiques de comportements des groupes. Nous avons donc créé un nouveau classificateur - classificateur de comportement des groupes [*Section 6.2*].

A cause des inconvénients de la première architecture, nous avons décidé de passer d'un traitement séquentiel à un traitement parallèle. Nous avons, de cette façon obtenu une nouvelle architecture basée sur les trois classificateurs indépendants de la notation de l'opinion : classificateur linguistique, classificateur statistique et classificateur de comportement des groupes [5.7].

A ce stade du processus, nous avons obtenu trois notes pour une seule critique. Pour l'attribution de la note finale nous ajoutons une nouvelle étape qui sera décisive pour la

détermination de la note. L'attribution de la note finale est effectuée par le dernier classificateur grâce à son utilisation du réseau de neurones. L'utilisation de ce classificateur est justifiée par la présence d'une grande base d'apprentissage contenant les critiques cinématographiques annotées. En vue de ces conclusions et remarques, nous pensons détenir une bonne architecture finale responsable de la notation d'opinion de notre system.

5.5 Publication du résultat

La dernière étape de notre système est de partager notre base de données pour qu'elle puisse être utilisée par les moteurs prédicatifs. Le principal objectif est de fournir des profils d'utilisateurs pour améliorer les résultats des algorithmes des moteurs prédicatifs. Un profil correspond à un utilisateur ayant voté sur plusieurs films. Du point de vue du système de recommandation, un profil utilisable est un profil qui contient un nombre important de films notés, car les algorithmes prédicatif ne peuvent pas faire des correspondances avec des profils d'utilisateurs auant votés sur un seul film. Plus nous sélectionnons d'utilisateurs ayant effectués de multiples critiques, plus le profil est complété et l'utilisation efficace.

Dans notre base de données, nous ajoutons toutes les informations possibles extraites pendant le processus de la recherche et de la collection des critiques cinématographiques. Le format de la base de données est présenté dans la [Figure 5.9]. Pour la publication des données, nous avons développé un Service Web. La structure du Service Web ne sera pas détaillée dans ce manuscrit car elle ne rentre pas dans le cadre de cette thèse.

Comme nous l'avons déjà précisé, nous avons eu la possibilité de retirer des informations intéressantes sur les utilisateurs qui ont noté les films. A partir du site "*Movieslens*" par exemple - *www.grouplens.org*, étude menée par "University of Minnesota" et "the GroupLens Research Group", nous pouvons retirer des informations comme l'âge, le sexe, le statut socioprofessionnel ou le code postal (zip code) des utilisateurs. Ces informations sont intéressantes dans le but d'effectuer des analyses statistiques des utilisateurs.

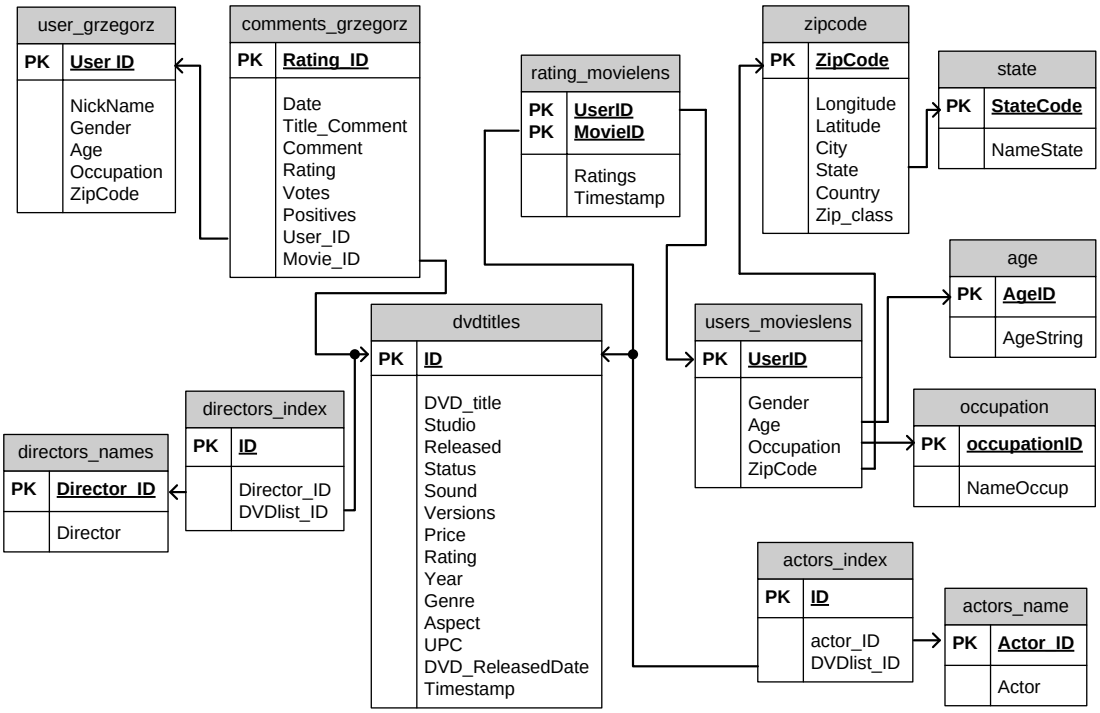


Figure 5.9: Le format de la base de données -

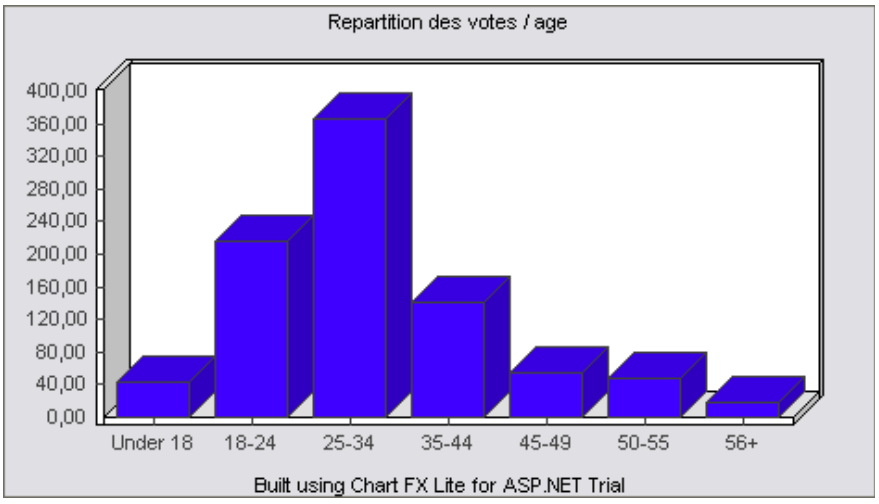


Figure 5.10: Exemple d'une étude statistiques - Répartition des votes par rapport à l'âge des utilisateurs pour le film *Goldeneye* qui a une note moyenne de 3,54 et a reçu 888 votes

La figure [Figure 5.10] par exemple, donne une répartition des votes par âge des utilisateurs pour un film donné. Nous remarquons que la catégorie des "25-34 ans" est la plus représentative de cet échantillon avec 400 votes. Malheureusement dans le processus de la recherche des critiques cinématographiques, nous ne disposons pas toujours de la possibilité de collecter des informations aussi intéressantes sur les auteurs. Nous pensons cependant que l'utilisation de ces informations pourrait améliorer encore la performance de la notation de l'opinion, car il est probable que les personnes de même âge et issues de la même agglomération par exemple utilisent une langue similaire pour décrire leurs sentiments. L'analyse et l'utilisation de ces informations sur les utilisateurs dans le processus de catégorisation semble être prometteur et ouvre un chemin intéressant pour la recherche dans ce domaine. Le seul inconvénient pour l'instant est le manque des ces informations, mais avec l'évolution permanente du Web nous espèrons qu'elles vont bientôt être disponibles.

5.6 Conclusion

Nous avons décrit dans ce chapitre l'architecture globale du système développé. Ce système est un support pour les moteurs prédicatifs et son rôle est de collecter les critiques cinématographiques à partir d'Internet et d'attribuer une note par rapport aux sentiments inclus dans les critiques. Le système développé est autonome et permet d'effectuer la recherche et la notation des critiques cinématographiques d'une manière automatique. Nous avons brièvement présenté les trois modules de notre système : le module de la recherche de critiques cinématographiques, le module de notation des critiques et le module de publication des résultats.

Le module le plus intéressant de notre système, au point de vu recherche, est le module de la notation de l'opinion. Nous n'avons pas précisé dans ce chapitre les méthodes de classification utilisées.

Nous présentons en détail, dans le chapitre suivant, le module de notation de l'opinion qui est basé sur les différentes approches de catégorisation de textes. Nous présentons les classificateurs développés pour effectuer l'attribution de la note à la critique, ces classificateurs étant basés sur les différentes approches de catégorisation et de

compréhension du texte comme l'approche statistique des *Chapitre 2* et *Chapitre 3* et l'approche linguistique du *Chapitre 4*.

Chapitre 6

Module de notation de l'opinion

6.1 Architecture générale du module de notation de l'opinion

Dans le chapitre précédent, nous avons présenté l'architecture générale du système développé sans entrer dans les détails du module de notation des sentiments d'une critique cinématographique, le module le plus intéressant du point de vue de la recherche. Dans ce chapitre, nous allons décrire en détail les techniques de classification. Nous proposons deux nouvelles approches :

- le classificateur de comportement des groupes,
- le classificateur linguistique.

Ensuite, nous comparons les résultats avec le classificateur statistique basé sur la classification "naïf Bayes" et la classification SVM [Figure 6.1]. Pour le marquage de l'opinion

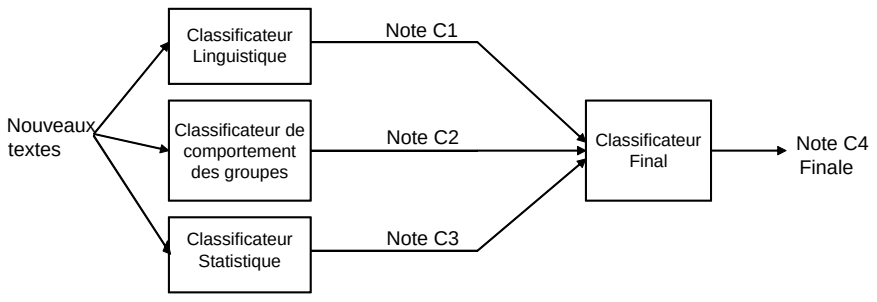


Figure 6.1: Notation de l'opinion - Trois différentes approches pour la notation de l'opinion et classificateur final pour combiner les résultats

nous utilisons trois approches différentes qui sont les suivantes :

- le classificateur de comportement des groupes : c'est une recherche statistique sur les données linguistiques pour déterminer le comportement des critiques cinématographiques qui ont la même note attribuée. Nous avons sélectionné plusieurs éléments que nous considérons comme les caractéristiques qui déterminent une des 5 catégories composées des critiques avec la même note associée. Ces caractéristiques décrivent le comportement des groupes de critiques. Les caractéristiques étudiées sont par exemple : les mots caractéristiques, la longueur des phrases, la taille de l'opinion, la présence de la négation, les expressions caractéristiques ou la ponctuation spéciale. Pour déterminer la note de la nouvelle critique, nous avons calculé la distance entre les caractéristiques de la nouvelle critique et les caractéristiques des groupes
- le classificateur statistique : c'est une recherche basée sur le théorème de Bayes ou SVM
- le classificateur linguistique : pour chaque phrase de la critique, nous attribuons une règle de grammaire qui exprime l'intensité de l'opinion. A la fin, nous calculons la note moyenne des phrases de la critique traitée.

L'attribution de la note finale à la critique est effectuée grâce à un dernier classificateur basé sur un réseau de neurones.

6.2 Le classificateur de comportement des groupes

6.2.1 L'approche générale

Dans cette section, nous présentons le classificateur utilisé pour la notation de l'opinion. L'approche générale est basée sur la vérification que les critiques ayant la même note associée ont des caractéristiques communes. Ensuite, nous déterminons un comportement des critiques qui ont la même note, nous déterminons donc le comportement général de chacun des groupes de critiques (5 groupes correspondant à 5 différentes notes de l'opinion).

Nous avons un très grand nombre de critiques cinématographiques déjà notées, mais pour effectuer l'étude des groupes nous utilisons une base de 1000 critiques (200 critiques par groupe). Nous avons rassemblé toutes les critiques selon leur note. Nous obtenons

alors 5 groupes différents de critiques du film. Ensuite, nous avons essayé de déterminer les caractéristiques typiques de chaque groupe. Nous avons défini tous les paramètres qui pourraient caractériser le comportement d'un groupe tels que :

- les mots caractéristiques,
- les expressions caractéristiques,
- la longueur de phrase,
- la taille de l'opinion
- la fréquence de répétition de plusieurs mots,
- la négation
- le nombre de signes de ponctuation (!, ;), ?)
- et ainsi de suite ...

Le choix des critères que nous avons gardés pour analyse de comportement du groupe a été fait de manière empirique. Tout d'abord en analysant les corpus de critiques, nous avons défini des critères qui nous semblaient intéressants et qui pouvaient déterminer le comportement du groupe. Ensuite, nous avons testé ces critères sur une base d'apprentissage contenant mille critiques. Si les résultats montraient des différences entre les groupes, nous avons considéré ces critères comme critères valides pour nos travaux de recherche.

Dans cette approche, nous présentons l'étude statistique sur les données linguistiques. La base d'apprentissage a été utilisée pour l'analyse des critiques avec la même note afin de trouver les caractéristiques qui déterminent le comportement de chaque groupe. Chacune des approches utilisées dans notre recherche est basée sur différentes caractéristiques pour ne pas les répéter dans le processus de la classification. Néanmoins, nous avons emprunté les classes sémantiques de l'approche linguistique pour la création de la liste des mots caractéristiques. L'utilisation de ces données est différente dans ces deux approches. Apres avoir sélectionné des critères qui caractérisent les groupes de notes, nous avons analysé le corpus pour obtenir des résultats statistiques. Les résultats montrent de grandes différences entre les caractéristiques de ces groupes. La création du comportement global de chaque groupe permet de déterminer à quel groupe appartient une nouvelle critique cinématographique. Pour les nouvelles critiques, nous avons calculé la distance entre ses caractéristiques et les caractéristiques des groupes.

6.2.2 Architecture du processus

Il s'agit de réaliser un logiciel réalisant une classification par groupe d'utilisateurs à partir d'une base de données. La base de données a été présentée dans le *Chapitre 5*. La classification se fait en fonction de critères caractérisant les groupes de notes ; elle sera basée sur le vocabulaire et la syntaxe de la langue anglaise. Nous avons récupéré les critiques à partir des sites d'Amazon et d'IMDB. Amazon utilise la notation de 1 à 5, par contre IMDB utilise une note allons de 1 à 10. Dans un premier temps, il fallait unifier l'ensemble des critiques en 5 catégories : les critiques ayant des notes de 1-2, 3-4, 5-6, 7-8 et 9-10. Nous avons créé la base d'apprentissage de cette base de critiques. Ensuite, pour chaque catégorie, un traitement est effectué sur le vocabulaire utilisé et la syntaxe des phrases afin d'en extraire des statistiques exploitables pour la notation de l'opinion (par exemple : occurrences de certains mots appréciatifs et dépréciatifs, type de ponctuation utilisé le plus souvent...) ce qui permettra de réaliser une étude du comportement du groupe.

La construction de ce classificateur a été découpée en 7 modules qui remplissent chacun une fonction du système [Figure 6.2].
Les fonctions de chaque module sont les suivantes :

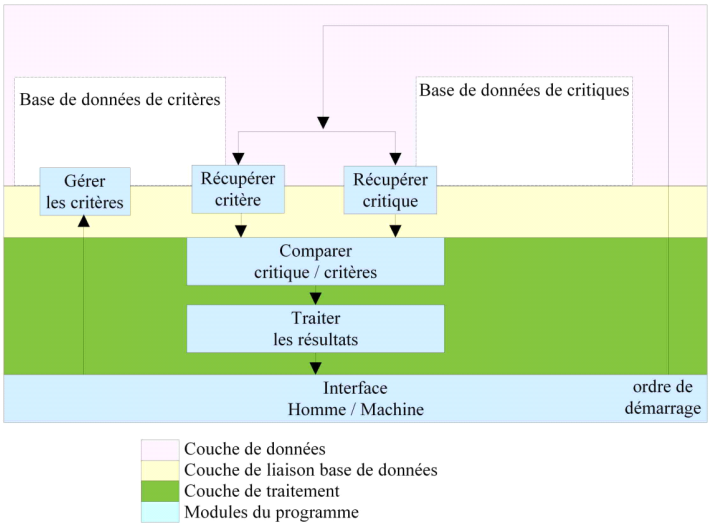


FIGURE 6.2: Architecture du classificateur de comportement des groupes -

– **Récupérer les critiques**

Le but de cette fonction est de récupérer les données présentes dans la base de données des critiques de film et de les mettre en forme pour qu'elles puissent être traitées par la couche suivante (couche de traitement des données). Les catégories de notes vont être également modifiées. Les notes sur dix seront ramenées à des notes sur 5. Les critiques ayant une note inférieure ou égale à zéro ne seront pas traitées. Le texte de la critique ne devra pas être vide. Les principales contraintes de ce module sont :

– la répartition des critiques en 5 catégories,
– la suppression des critiques sans note ou avec une note égale à 0,
– la suppression des critiques vides

Les données seront donc renvoyées sous la forme : {critique : chaîne de caractères}. Cette chaîne contient le titre de la critique, son contenu, et la note attribuée au film concerné. Le module récupère comme paramètre le chemin vers la base de données pour pouvoir lire les critiques séquentiellement : {chemin relatif vers la base : chaîne de caractères ; pointeur dans la base : nombre de caractères à partir du début de la base}.

– **Récupérer les critères**

Le but de cette fonction est de récupérer les données présentes dans la base de données des critères d'appréciation et de les mettre en forme pour qu'ils puissent être traités par la couche suivante (couche de traitement des données). Ces critères sont de nature différente (longueur de la critique, présence de certains mots ou certaines expressions, ponctuation,...) et doivent être choisis au fur et à mesure du développement. Nous aurons néanmoins une valeur retournée du type : {critères : tableau de chaînes de caractères}. Les critères ont été sélectionnés manuellement et ont été testés pour verifier si les critères varient selon l'analyse statistique sur les groupes de critiques annotées.

– **Comparer les critiques aux critères**

Ce module récupère les données mises en forme par les deux modules précédents afin de déterminer la relation entre les critères indiqués dans la base de données et la note attribuée par la critique. Une critique ne présentant aucun des critères proposés est signalée à l'utilisateur. Ce module renvoie un tableau de statistiques généré par les modules de la base des critères sur chacune des critiques examinées. Ce module fait appel aux deux précédents et récupère les valeurs qu'ils retournent en tant que paramètre. Le module récupère donc tous les fichiers critères et les exécute tous pour chaque critique de la base.

– **Traiter les résultats**

Ce module permet d'interpréter, de mettre en forme et de stocker les valeurs retournées par le module précédent. Les résultats sont donnés sous une forme statistique à laquelle peut accéder l'IHM. Nous devons aboutir à une liste de critères permettant de faire ressortir les critères ou les combinaisons de critères relatifs à une note. Toutes les statistiques ont été présentées pour avoir le comportement de chaque groupe. Les résultats ont été normalisés par la taille de la critique pour pouvoir les comparer afin d'avoir l'image globale d'utilisation des critères et pour pouvoir calculer la distance entre les caractéristiques d'une critique que l'on souhaite traiter et les caractéristiques du comportement du groupe.

– **IHM**

Son but est de réaliser l'interface entre l'utilisateur et le système. Elle permet donc de choisir les différents critères de l'analyse, d'afficher les erreurs éventuelles et de présenter les résultats de l'analyse à l'utilisateur. Ce module permet de mettre en forme graphiquement les données générées par l'analyse.

6.2.3 Les critères

Nous allons présenter les exemples des différents critères qui ont été étudiés pour déterminer le comportement des groupes.

– **Taille d'une critique**

Nous voulons ici voir si la longueur d'une critique dépend de son caractère positif ou négatif. Ce critère permet de déterminer le cumul des tailles et la taille moyenne des critiques appartenant à chaque catégorie. Le module renvoie donc un résultat de la forme :

Cumul des tailles pour la note 1	Nombre de critiques pour la note 1	Cumul des tailles pour la note 2	Nombre de critiques pour la note 2	...
...	...	...	...	...

Tableau 6.1: Critères : taille de la critique

– **Majuscules**

Souvent, dans des critiques, lorsqu'une personne veut insister sur une chose, elle utilise des majuscules. Sur le même principe, le module renvoie le cumul du nombre de majuscules présentes dans chaque catégorie de notes ainsi que le nombre de critiques inspectées. Le tableau renvoyé est donc de forme :

Cumul des nombres de majuscules pour la note 1	Nombre de critique pour la note 1	Cumul des nombres de majuscules pour la note 2	Nombre de critique pour la note 2	...
...	...	...	...	...

Tableau 6.2: Critère : nombre de lettres majuscules

– **Superlatifs**

Les superlatifs en anglais sont formés de *l'adjectif+est*. Nous avons donc décidé d'estimer le nombre de superlatifs dans une critique en comptant les mots se terminant par *"-est"*. Ce module renvoie également le cumul du nombre de mots trouvés et le nombre de critiques parcourues pour chaque catégorie.

Cumul des nombres de mots finissant en "est" pour la note 1	Nombre de critiques pour la note 1	Cumul des nombres de mots finissant en "est" pour la note 2	Nombre de critiques pour la note 2	...
...	...	...	...	...

Tableau 6.3: Critères : nombre des adjectifs "-est"

Comme l'exemple des critères, nous avons également étudié les mots caractéristiques. Pour cela, nous nous sommes inspirés des classes sémantiques et du dictionnaire *General Inquirer* qui liste les mots ayant des caractères positifs et négatifs, ayant des niveaux d'intensité forte ou faible. Ce dictionnaire sera traité dans la *Section 6.4*.

6.2.4 Résultats

Pendant la recherche des critères, nous avons remarqué que les résultats de l'analyse effectuée sur la base d'apprentissage montrent bien de grandes différences entre les groupes extrêmes - c'est à dire entre les groupes avec une et cinq étoiles. La tâche difficile était de trouver les critères pour différencier les groupes avec 2 et 3 étoiles. Un exemple des résultats statistiques obtenus par ces approches pour retrouver le comportement des groupes est montré dans le Tableau 6.4.

critere		groupe 1	groupe 2	groupe 3	groupe 4	groupe 5
longeur phrase		168	143	112	43	112
taille critique		174	125	76	84	103
majuscules		320	396	392	348	297
I		316	221	228	156	171
!		182	79	53	58	154
?		67	78	98	112	53
:)		42	21	26	17	36
adj-est		44	62	34	26	75
great		17	8	2	1	2
best		34	16	3	6	1
poor		1	3	21	33	37
boring		2	3	4	13	21
anything		4	20	7	4	4
negation		27	19	33	39	42
phrases(1-20)		62	33	23	45	68
phrases(20-50)		114	211	188	108	220

Tableau 6.4: Résultats d'une étude statistique avec un ensemble de critères donnés sur le corpus de base d'apprentissage

L'exemple montre les résultats obtenus après avoir traité la base d'apprentissage de 200 critiques par note avec seulement quelques critères. Les valeurs numériques dans le tableau sont ou bien des moyennes (par exemple la longueur des phrases) ou bien les valeurs normalisées par les longueurs des critiques (par exemple la fréquence du signe de ponctuation "!"). Les résultats dans le tableau ne montrent pas l'efficacité de la méthode, ils montrent seulement l'étude statistique sur l'ensemble de critiques

de base d'apprentissage avec quelques critères. Les tests sont décrits dans le chapitre suivant [*Section 7.1.5*] où nous précisons le rappel et la précision et nous décrivons les avantages et les inconvénients de toutes les approches.

6.3 Le classificateur statistique

6.3.1 L'approche générale

Dans cette section, nous présentons une approche généralement utilisée dans l'analyse des sentiments. Nous utilisons cette méthode pour comparer les résultats de nos approches avec la même base d'apprentissage. La manière de procéder à une classification est de trouver une caractéristique de chaque classe et d'associer une fonction d'appartenance. Parmi les méthodes connues, nous pouvons citer les classificateurs de Bayes et la méthode de SVM. Nous avons obtenu de meilleurs résultats pour le classificateur "naïf Bayes", nous allons donc nous baser sur ce classificateur. Dans nos travaux de recherche, nous avons utilisé ce classificateur tout d'abord pour déterminer la subjectivité et objectivité des phrases, puis pour attribuer une note aux phrases subjectives de la critique. Le processus général nécessite la préparation des bases d'apprentissage pour deux classificateurs : classificateur de filtrage des phrases subjectives / objectives et classificateur pour l'attribution d'une note. Les étapes intermédiaires sont les suivantes :

- prétraitement et lemmatisation,
- vectorisation et calcul des index complets,
- constitution de bases d'apprentissage pour chaque classificateur,
- réduction de l'index dédié à un classificateur,
- ajout de synonymes,
- classification des textes

6.3.2 Représentation vectorielle

Pour l'attribution d'une note aux sentiments de la critique par l'approche statistique, nous utilisons donc deux classificateurs : un premier pour filtrer la phrase objective et subjective et un deuxième pour l'attribution d'une note à la critique. La notation n'est effectuée que sur les phrases classées phrases subjectives. Ces classificateurs reposent sur une représentation vectorielle du texte de la base d'apprentissage. Cette représentation vectorielle nécessite dans un premier temps un prétraitement linguistique pour

le découpage de la phrase, pour la lemmatisation et pour la suppression de tous les mots ne participant pas au sens du document. Ce prétraitement a été effectué pour le classificateur linguistique, nous allons donc le réutiliser pour la classification statistique. Ensuite, nous construisons un premier index de la base d'apprentissage qui représente les occurrences de mots. Avant d'utiliser le classificateur, nous réduisons la dimension de l'index.

Nous effectuons le prétraitement grâce à l'application *Unitex*. Nous disposons déjà de ressources linguistiques préparées pour cette tâche comme la grammaire de découpage de phrases ou les dictionnaires. Ensuite, nous éliminons des termes vides comme les articles indéfinis et définis et les prépositions. Nous pouvons effectuer cette tâche car ces éléments de grammaire ont une faible influence sur le sens des textes, comme sur l'opinion décrite dans les critiques, contrairement aux adverbes par exemple apportant une forte contribution au jugement de valeur. Puis sur un corpus d'apprentissage, nous calculons la dimension de l'espace vectoriel de représentation des textes pour effectuer l'énumération de tous les lemmes - l'index complet. Chaque document est alors représenté par un vecteur qui contient le nombre d'occurrences de chaque lemme présent dans le document.

Tous les documents de la base d'apprentissage sont représentés par un vecteur dont les dimensions correspondent à l'index complet et les composantes sont les fréquences d'occurrences des unités de l'index dans le document. Donc à ce stade de procédure, les textes sont vus comme des ensembles de phrases. Maintenant, chaque phrase est étiquetée par rapport à la construction des classificateurs (le classificateur de subjectivité et le classificateur de notation). Les étiquettes correspondent aux phrases subjectives (PS) ou objectives (PO) et à la note qu'on estime attribuée (N de 1 a 5). Une phrase j du document i est notée de la façon suivante :

$$\vec{V_{D_iP_j}} = (f_{D_iP_j1}, ...f_{D_iP_jk}, ...f_{D_iP_j|D|}, PS/PO, N) \tag{6.1}$$

où $f_{D_iP_jk}$ représente le nombre d'occurrences du lemme k dans la phrase j du document i. L'étape de l'étiquetage était basée sur les notes des critiques de la base d'apprentissage et les phrases subjectives ont été étiquetées manuellement. De cette manière, nous avons construit l'ensemble d'apprentissage nécessaire à la détermination des classificateurs de

subjectivité et de la notation de sentiments.

La dernière étape de la représentation vectorielle du corpus de documents est la réduction de l'index complet dédié au classificateur. La réduction de l'index complet consiste à éliminer de l'espace vectoriel de la base d'apprentissage des vecteurs qui ont de nombreuses composantes toujours nulles. Cette tâche permet d'éliminer le bruit dans le calcul des classificateurs [Cover & Thomas (1991), Dziczkowski & Wegrzyn-Wolska (2007b)]. Dans la littérature, nous pouvons trouver plusieurs méthodes de réduction de l'index complet [*Chapitre 1*]. Nous avons utilisé la méthode de l'information mutuelle associée à chaque dimension de l'espace vectoriel.

L'information mutuelle est définie de la manière suivante :

Soient X la variable aléatoire associée à l'ensemble des classes et Y la variable aléatoire représentant l'absence ou la présence du mot y dans un document, où Y prend les valeurs $y \in 0, 1$ (l'absence ou la présence du mot y). *P(x)* indique le ratio du nombre des documents de la classe $x \in X$ apparaissant sur le nombre total de documents, *P(y)* indique le ratio du nombre de documents contenant une ou plusieurs occurrences du mot x sur le nombre total de documents, *P(x,y)* indique le ratio du nombre de documents de la classe $x \in X$ et concernant le mot y sur le nombre total de documents. L'entropie de la variable X - *H(X)* est égale à :

$$H(X) = -\Sigma_{x \in X} P(x) log(P(x)). \tag{6.2}$$

L'entropie de la variable X conditionnée par la présence ou l'absence du mot x - *H(X/Y* est égale à :

$$H(X|Y) = -\Sigma_{y \in 0,1} P(y) \Sigma_{x \in X} P(x|y) log(P(x|y)). \tag{6.3}$$

L'information mutuelle moyenne est égale à la différence entre ces deux entropies :

$$I(X,Y) = H(X) - H(X|Y) = \Sigma_{x \in X} \Sigma_{y \in 0,1} P(x,y) log \frac{P(x,y)}{P(x)P(y)}. \tag{6.4}$$

Nous sélectionnons les mots dont l'information mutuelle est supérieure à un seuil donné. Cette méthode permet de calculer pour un classificateur un index réduit. Nous

avons utilisé ces calculs pour les ensembles d'apprentissage pour deux classificateurs présents. Ensuite, nous avons calculé les vecteurs d'occurrences pour ces ensembles d'apprentissage. A la fin de ce processus, nous obtenons la représentation vectorielle, les données sont donc préparées pour la phase de classification.

6.3.3 L'insertion des synonymes

La tâche de l'introduction des synonymes nous permet de compléter et de donner plus de précision à la représentation vectorielle sans demander beaucoup de travail manuel. Avant d'effectuer le calcul de chaque classificateur, nous avons ajouté des synonymes aux vecteurs d'occurrences. Nous effectuons ce processus car nous disposons d'une analyse linguistique profonde grâce à la création du classificateur linguistique [*Section 6.4*]. Pendant l'analyse linguistique, nous avons ajouté des classes sémantiques dans les dictionnaires ce qui nous facilite la tâche de cumulation des synonymes pour les utiliser dans la classification statistique. L'insertion des synonymes consiste à construire une table de correspondance entre les lemmes de l'index réduit du classificateur et leurs synonymes respectifs. Nous ajoutons tous les synonymes du lemme donné dans le vecteur d'occurrence. Si le mot m_i a *k'* synonymes, nous ajoutons *k'* termes dans le vecteur. Le vecteur d'occurrences des phrases est donc le suivant :

$$\vec{V_{D_iP_j}} = (f_{D_iP_j1} + \Sigma_1^{k'_1} f_{D_iP_j s_{1k'}}, ... f_{D_iP_j|D|} + \Sigma_1^{k'_{|D|}} f_{D_iP_j s_{|D|k'}}, PS/PO, N). \qquad (6.5)$$

ou $\Sigma_1^{k'_1} f_{D_iP_j s_{1k'}}$ représente tous les synonymes (de 1 jusqu'a k'_1) de lemme *1* dans la phrase j du document i [Plantie (2006)].

De cette façon, nous avons constitué la base d'apprentissage pour nos classificateurs statistiques. Cette base contient des vecteurs d'occurrences de l'index réduit et elle est utilisée ensuite dans l'étape de la classification.

6.3.4 L'étape de la classification

Dans le chapitre précédent [*Section 2.5*], nous avons montré plusieurs techniques pour la classification des sentiments. Dans nos travaux, nous avons utilisé deux classifications : la classification basée sur modèle de "naïf Bayes" et la classification utilisant SVM. Les deux méthodes ont été testées et les meilleurs résultats (F-score) sont obtenus pour les classificateurs "naïf Bayes". C'est donc le classificateur de Bayes qui a été

utilisé pour le système.
Dans le processus de la classification statistique, nous avons tout d'abord classifié les phrases subjectives et nous avons ensuite attribué une note.

6.3.4.1 Le classificateur de subjectivité

Les phrases intéressantes pour effectuer la notation de l'opinion sont les phrases subjectives car ce sont les seules qui contiennent l'avis de l'auteur. Pour cette raison, nous avons effectué tout d'abord le filtrage des phrases subjectives. Le schéma qui représente ces tâches est montré sur la [Figure 6.3].

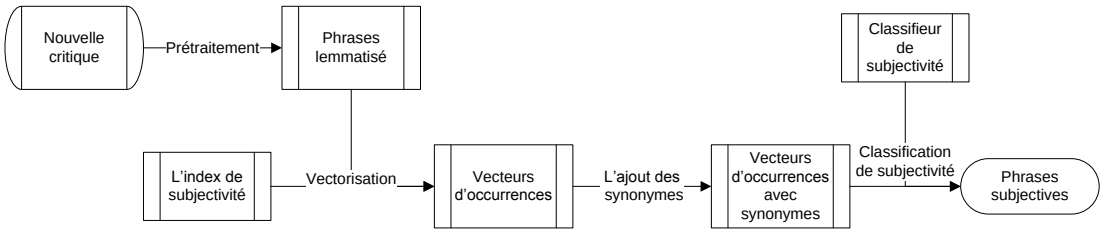


Figure 6.3: Classification de subjectivité - les étapes de la classification

Le processus présenté permet de filtrer uniquement les phrases subjectives, les seules phrases à exprimer une opinion. Les différentes étapes sont les suivantes :

- Le prétraitement consiste à effectuer le découpage en phrase, la lemmatisation et la suppression des mots insignifiants dans notre recherche.
- La vectorisation consiste à mettre toutes les phrases sous forme de vecteur d'occurrences et à réduire l'index complet.
- L'ajout des synonymes consiste à ajouter des termes (synonymes) dans les vecteurs d'occurrences grâce à l'analyse linguistique
- La classification de subjectivité consiste à regrouper les phrases en phrases subjectives et objectives. La classification est basée sur le théorème de Bayes. Pour le reste de la classification (notation), nous gardons seulement les phrases subjectives.

Pour la classification de subjectivité, nous avons utilisé le classificateur de "naïf Bayes".

6.3.4.2 Le classificateur de notation des opinions

Apres avoir effectué la classification de subjectivité, nous ne retenons que les phrases subjectives. Nous effectuons une classification pour pouvoir attribuer une note à ces phrases de chaque critique analysée. Le schéma qui représente ces tâches est présenté sur la [Figure 6.4]. Le processus présenté permet d'attribuer une note aux phrases

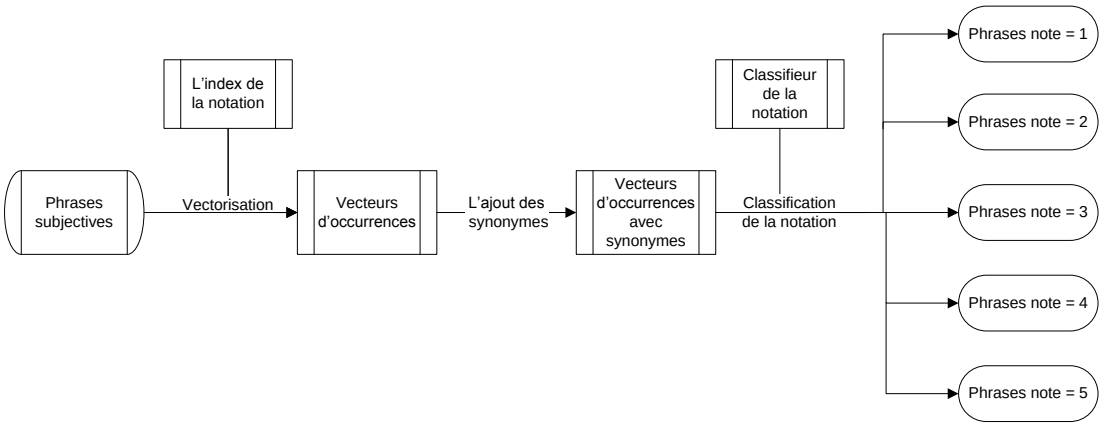


Figure 6.4: Classification de la notation - les étapes de la classification

classées comme phrases subjectives. La notation varie entre les valeurs 1 à 5. Les étapes sont les suivantes :

- La vectorisation et la réduction de l'index complet dédié à la classification de la notation.
- L'ajout de synonymes.
- La classification de la notation qui consiste à regrouper les phrases par rapport à l'intensité des sentiments. Les notes sont entre 1 et 5.

A ce stade du processus nous obtenons toutes les phrases subjectives avec une note associée. La note globale d'une critique de la classification statistique est la moyenne arithmétique de toutes les phrases de cette critique. Les tests effectués et la comparaison de cette approche avec d'autres sont décrites dans le chapitre suivant [*Chapitre 7*].

6.4 Le classificateur linguistique

6.4.1 L'approche générale

Nous effectuons la notation des critique sur une échelle de 1 à 5. Pour l'approche linguistique nous avons créé une règle de grammaire pour chacun de ces groupes. Cette grammaire est basée sur une analyse des critiques de la base d'apprentissage, qui contient environ 2000 phrases pour chaque note (la même base de données que pour les autres classifications). Pour cette partie, nous avons utilisé un traitement linguistique qui exige des lexiques et des grammaires spécialisés. Le développement de ces ressources est une tâche longue et fastidieuse, qui nécessite généralement une expertise dans le domaine et les connaissances en traitement de l'information linguistique telles que les techniques de filtrage, de catégorisation de documents et d'extraction de l'information.

Cette partie du système a été développée avec l'application *Unitex*. Nous utilisons un analyseur linguistique *Unitex* pour effectuer une prétraitement et une lemmatisation des mots, pour ajouter des synonymes, pour détecter la négation, pour ajouter des classes sémantiques aux mots et enfin, pour la partie la plus importante pour nos travaux de recherche - la construction de grammaires locales complexes.
Pour séparer des mots en différents niveaux d'intensité de l'opinion, nous avons introduit les classes sémantiques qui sont associées aux mots et montrent la polarité et l'intensité. Afin d'associer les classes sémantiques aux mots, nous avons utilisé un dictionnaire de subjectivité - *General Inquirer Dictionary*.

Le *General Inquirer Dictionary* est un dictionnaire qui associe des codes décrivant la subjectivité des mots. Il combine les catégories du contenu d'analyse du dictionnaire "Harvard IV-4", les catégories du contenu d'analyse du dictionnaire "Lasswell", et les cinq catégories des travaux des Semin et Fiedler sur la cognition sociale, soit un total de 182 catégories. Chaque catégorie est appliquée à la liste de mots et donne la caractéristique des mots si elle existe pour une catégorie donnée. Contrairement à certaines applications d'intelligence artificielle, le *General Inquirer Dictionary* attribue simplement des étiquettes selon les catégories et ne permet pas de donner un sens au texte. La compréhension à partir de ces étiquettes reste la travail des experts et non de l'ordinateur. Une analyse manuelle est nécessaire pour approuver l'utilisation de ces

données. Ce dictionnaire contient 1915 mots dits positifs, et 2291 mots dits négatifs. Les catégories intéressantes sont la subjectivité : positif, négatif et l'intensité : fort (*ang : strong*), faible (*ang : weak*). Un exemple des entrées de ce dictionnaire est présenté sur la [Figure 6.5].

FATAL	H4Lvd		Negativ							
FATALISTIC	H4		Negativ							
FATE	H4Lvd							Strong		
FATHER	H4Lvd				Affil			Strong		
FATHOM#1	H4Lvd							Strong		
FATHOM#2	H4Lvd							Strong		
FATIGUE	H4Lvd		Negativ			Ngtv				Weak
FAULT	H4Lvd		Negativ			Ngtv				Weak
FAVOR#1	H4Lvd	Positiv		Pstv	Affil					
FAVOR#2	H4Lvd	Positiv		Pstv						
FAVOR#3	H4Lvd	Positiv		Pstv						

Figure 6.5: General Inquirer Dictionary - Entrées limitées à 10 sur 182 catégories, nous pouvons voir les étiquettes positives, négatives, fortes et faibles

L'objectif principal du classificateur linguistique est l'attribution de la note par rapport aux sentiments décrits dans la critique. La notation est réalisée phrase par phrase. Afin de créer des règles de grammaire pour chaque note (dans notre cas la note est entre 1 et 5), l'étude des critiques de la base d'apprentissage a été effectuée. De cette manière, 5 grammaires ont été créées - une pour chaque note. Chaque grammaire contient un grand nombre de règles - des grammaires locales. Pour chaque grammaire, plus de 30 grammaires locales ont été créées. Pour attribuer une note à une nouvelle critique, l'analyse est réalisée phrase par phrase, afin de trouver une règle (de notre base des règles) correspondant à la phrase examinée. À la fin de ce traitement, nous obtenons les phrases de la nouvelle critique examinée avec les règles de grammaire correspondantes. La note finale de cette classification est la moyenne des notes correspondant aux grammaires générales.

La construction de grammaires locales a été réalisée manuellement par voie de l'analyse des phrases des critiques avec les mêmes notes associées. La grammaire locale ne peut pas être trop générale car cela rend les résultats de la recherche trop ambigus. Par contre, si la grammaire est trop spécifique et complexe, l'utilisation de cette grammaire est incertaine parce que le silence augmente de manière significative. Les grammaires ont été créées pour détecter la polarité et l'intensité d'opinion dans une phrase grâce à la forme des grammaires locales qui constituent une grammaire générale pour chaque

groupe de notation. Les travaux de recherche sont basés uniquement sur la forme des grammaires locales, d'autres caractéristiques purement statistiques comme les mots ou les expressions caractéristiques, la longueur de phrase, la fréquence des mots, la répétition des mots, le nombre de signes de ponctuation, etc. ne sont pas pris en compte. Bien sûr, les mots caractéristiques sont dans les dictionnaires avec les classes sémantiques et dans les grammaires locales, mais cette approche est un traitement linguistique (grammaire nécessaire) et non statistique (comme les deux autres classificateurs).

La création locale de la grammaire est une tâche fastidieuse. Les grammaires utilisées dans notre système ont été créées de manière empirique. Nous avons procédé de la manière suivante : tout d'abord, nous avons construit des grammaires générales ensuite nous avons ajouté un niveau de complexité à l'analyse linguistique et nous avons effectué des tests. Après les tests nous avons répété le processus (ajout d'un niveau de complexité). Pour chaque niveau, nous avons effectué des tests et nous avons calculé le *F-score*. Le résultat final des formes de règles de grammaires a été choisi pour obtenir les meilleurs résultats de *F-score*. Malheureusement, nous ne pouvons pas être certains que notre choix soit le plus cohérent. Nous avons pris en considération le fait que chaque classificateur présenté dans notre système devrait avoir ses propres critères et caractéristiques. Il est important de noter que le classificateur linguistique fournit les meilleurs résultats. En particulier, nous pouvons voir que le paramètre de précision est meilleur que celui obtenu en utilisant d'autres approches.

6.4.2 Présentation de l'application

Comme nous l'avons déjà décrit dans le [*Chapitre 4*], cette partie du système était réalisée avec l'analyseur de texte *Unitex*. Unitex permet de traiter en temps réel des textes de plusieurs méga-octets pour l'indexation de motifs morphosyntaxiques, la recherche d'expressions figées ou semi-figées, la production de concordances et l'étude statistique des résultats. La [Figure 6.6] donne un aperçu des ressources développées. Nous pouvons voir le corpus d'une critique (fenêtre à gauche en haut) déjà prétraitée, donc après avoir effectué le découpage en phrases et la normalisation des formes ambiguës. Egalement la liste de tous les tokens (fenêtre à gauche en bas) avec les fréquences d'apparition, ainsi que les unités linguistique (fenêtre à droite) traitée par les dictionnaires de mots simples et de mots composés. La dernière colonne de cette fenêtre

représente des unités linguistiques qui n'ont pas été retrouvées dans les dictionnaires (faute de frappe de l'auteur de la critique).
Sur la [Figure 6.7] nous pouvons voir une grammaire locale et un fichier de résultats

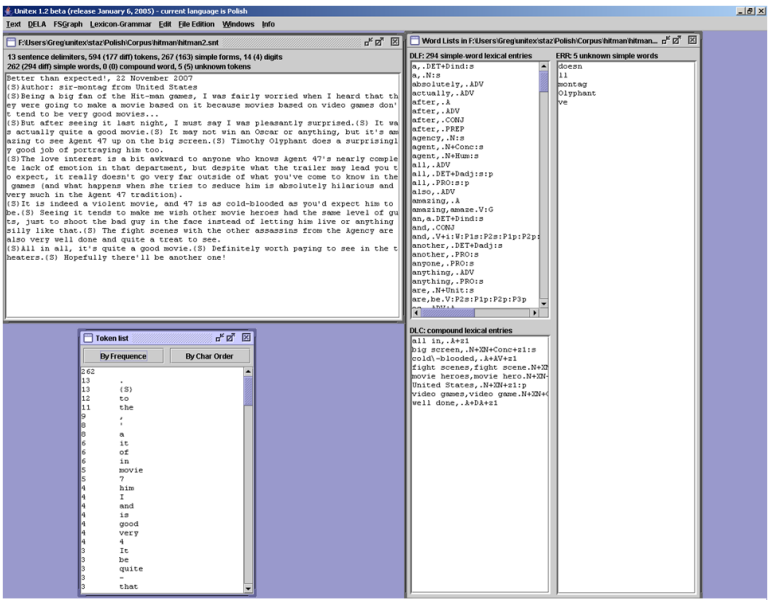


Figure 6.6: L'application Unitex - Représentation des ressources linguistiques, du prétraitement et des dictionnaires

constitué des phrases retrouvées dans le corpus de critiques après traitement.

6.4.3 La forme des grammaires locales

La tâche la plus importante dans l'approche linguistique présentée est la forme et l'utilisation de grammaires locales. C'est grâce à elles que l'attribution de la note aux phrases de la critique est effectuée. Les valeurs de la précision et du rappel dépendent de la forme des grammaires locales. Si les grammaires locales sont trop générales, la valeur de la précision est basse, pourtant la valeur du rappel est élevée. Si nous utilisons les grammaires locales très détaillées, nous obtenons la valeur du rappel très basse car il est impossible de prévoir les règles de grammaire pour toutes les phrases possibles décrivant des sentiments. Bien sûr, dans ce cas la valeur de pertinence s'approche de 100%.

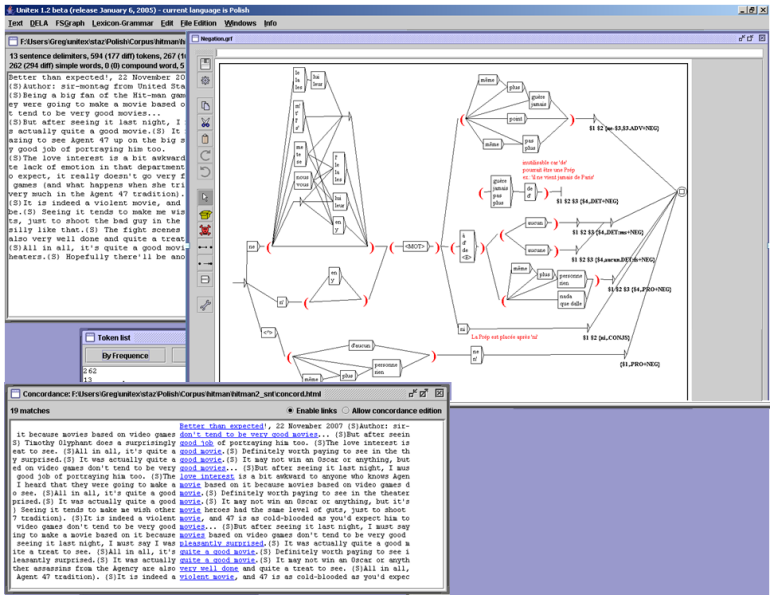


FIGURE 6.7: L'application Unitex suite - Représentation des ressources linguistiques, de la grammaire locale et des résultats de recherche

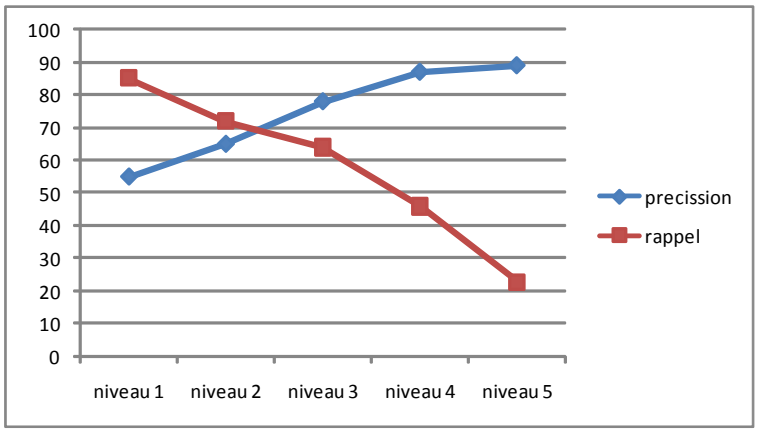


FIGURE 6.8: Niveau de complexité des grammaires et les mesure de pertinence - pour cinq niveau de complexité nous montrons le changement des valeurs de la précision et du rappel

Dans le diagramme [Figure 6.8], nous montrons les changements des paramètres de la précision et du rappel (π, ρ) par rapport au niveau de profondeur de l'analyse linguistique. Les résultats montrent la pertinence d'utilisation des grammaires créées sur la base d'apprentissage sur une nouvelle critique par rapport à la complexité de la forme d'une grammaire locale. Les résultats ne montrent pas l'efficacité de la notation de la nouvelle critique, car les valeurs de F-score n'étaient pas encore présentées. L'exemple montre comment les valeurs de précision et de rappel changent par rapport à la forme de la grammaire locale. Les tests ont été faits manuellement sur un ensemble de 20 critiques et un échantillonnage de grammaires de chaque niveau de compléxité.

En se basant sur ces résultats, nous avons proposé le processus de notation de l'opinion, donc la notation des phrases de chaque critique de la manière suivante. Après le prétraitement, nous avons toutes les phrases séparées que nous analysons phrase par phrase. L'étude débute par les grammaires locales avec le niveau de profondeur le plus haut. Si une grammaire correspond, la phrase est notée avec la note de la grammaire et la prochaine phrase est étudiée. Si aucune grammaire de niveau supérieur n'a été choisie, nous prenons l'ensemble des grammaires de niveau inférieur et le processus se répète jusqu'à parvenir au niveau le plus bas, donc les grammaires les plus générales. Si aucune grammaire ne correspond pas à la phrase, l'étude de la phrase suivante est effectuée avec le même processus [Figure 6.9].
Les mesures de pertinence de l'approche linguistique sont présentées dans le chapitre

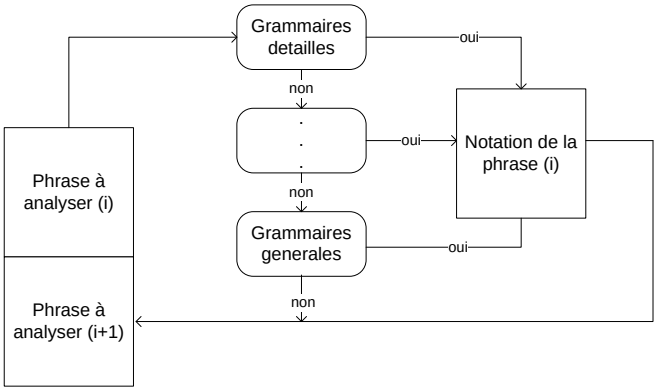


Figure 6.9: Processus de notation des phrases - la notation commence par les grammaires détaillées (analyse linguistique profonde) ; si pour la phrase testée la grammaire n'existe pas, les grammaires plus générales sont appliquées (analyse linguistique inférieure)

suivant [*Chapitre 7*].

6.5 Le classificateur final

Jusqu'à présent, nous avons présenté trois différentes méthodes pour attribuer une note à la critique. Ainsi, nous obtenons trois différentes évaluations (une pour chaque classificateur). La notation est effectuée chaque fois de manière différente, les notes ne sont donc pas toujours les mêmes. Comme nous obtenons trois notes différentes un nouveau problème consiste à effectuer l'évaluation finale pour attribuer une seule note à la critique. Nous avons besoin d'une classification finale pour avoir la note finale qui va être transmise au *Recommender System*. Nous avons remarqué que, si nous calculons la moyenne finale des résultats obtenus par les trois classificateurs, les résultats sont moins performants que ceux obtenus par le classificateur linguistique (selon le calcul du F-score - [*Chapitre 7*]).

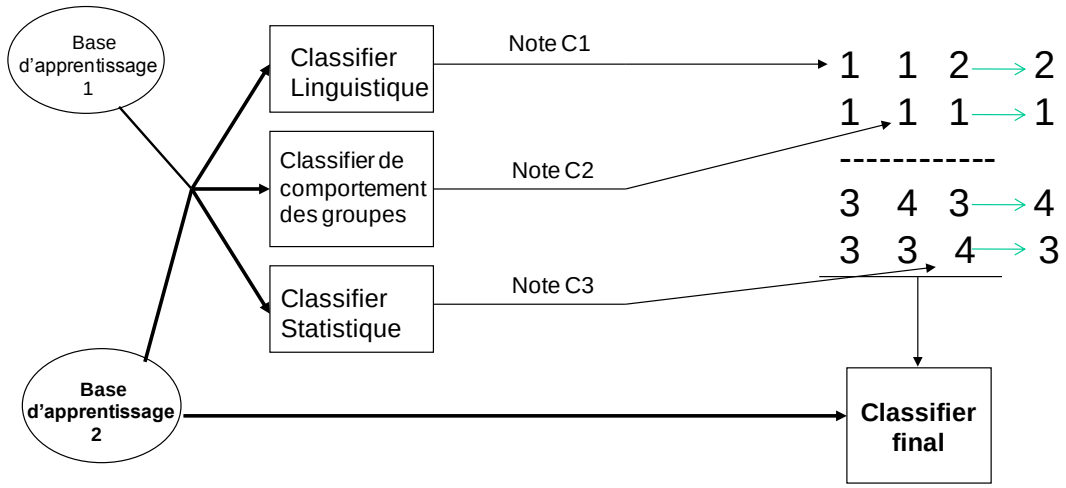


Figure 6.10: Classification finale - le comportement des notes montre la presence d'un classificateur déterminant dans certaines situations

Nous avons aussi remarqué que souvent un classificateur dans des situations spécifiques donne de meilleurs résultats, alors que dans d'autres situations, ce serait un autre classificateur. Nous donnons un exemple [Figure 6.10], souvent lorsque le premier classificateur donne une note égale à 2 et les deux derniers la note égale à 1, le résultat correct est égal à 2. Par conséquent, le premier classificateur est déterminant dans cette

situation.

Si, toutefois, les deux premiers classificateurs donnent des évaluations égales à 1, et le dernier la note de 2, dans ce cas, l'évaluation correcte est égale à 1. Donc, dans ce cas, nous constatons que nous ne devrions pas calculer la note finale comme la moyenne des notes, car un classificateur peut être plus influent. Dans le deuxième exemple [Figure 6.10], la situation est similaire, seulement dans cette situation c'est le deuxième classificateur qui est déterminant avec une note égale à 4 alors que les autres donnent la note de 3. Nous pouvons citer de nombreux exemples avec des comportements similaires.

Pour cette raison nous utilisons un classificateur final. Pour cette classification nous utilisons un réseau de neurones. Le choix de ce classificateur est justifié par la présence d'une très grande base de critiques déjà annotées qui servira pour la base d'apprentissage. En plus, il est facile d'implémenter ces données pour qu'elles servent de base d'apprentissage. La classification prend en compte seulement les probabilités de la note de chaque classificateur. Aucune autre caractéristique n'est prise en compte. Ce choix est justifié car nous pensons que nous avons utilisé toutes les caractéristiques possibles dans le processus de notation (en utilisant les trois classifications présentées) et nous ne voulons pas répéter ces caractéristiques dans les classifications. De plus, l'utilisation d'une caractéristique d'un classificateur de notation de l'opinion dans la classification finale pourrait influencer le choix de ce classificateur.

Pour les entrées du classificateur final, nous utilisons les notes des précédents classificateurs - les notes de chaque classificateur représentées par la probabilité d'appartenance à l'une des cinq classes de notes. Par exemple, le classificateur linguistique attribue la note de la façon suivante : la probabilité que la note est

- égale à 5 est $p_5 = 0{,}6$,
- égale à 4 - $p_4 = 0{,}2$,
- égale à 3 - $p_3 = 0{,}1$,
- égale à 2 - $p_2 = 0{,}1$
- égale à 1 - $p_1 = 0$

Nous avons utilisé un réseau de neurones pour déterminer la corrélation des notes obtenues par les 3 classificateurs. Nous utilisons le réseau de neurones de perceptron

multicouche (PMC) avec l'algorithme de retro propagation de gradient [Figure 6.11]. Nous utilisons :

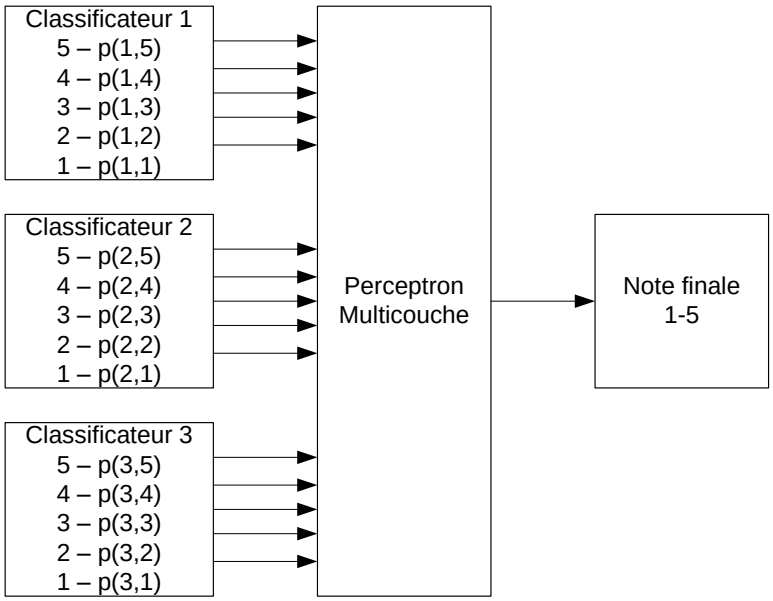


Figure 6.11: Perceptron multicouche - 15 entrées (5 de chaque classificateur), une sortie (la note finale), 3 sous couches

- 15 entrées :
  - C1(5 - p15, 4 - p14, 3 - p13, 2 - p12, 1 - p11),
  - C2(5 - p25, 4 - p24, 3 - p23, 2 - p22, 1 - p21),
  - C3(5 - p35, 4 - p34, 3 - p33, 2 - p32, 1 - p31)
- 1 sortie : la note finale (1-5)
- 3 sous-couches
- base d'apprentissage : 200 critiques pour chaque note (1000 critiques en total)

De cette façon, nous avons amélioré les résultats qui sont meilleurs que les résultats du classificateur le plus pertinent - le classificateur linguistique.

6.6 Conclusion

Dans ce chapitre nous avons présenté l'architecture du module notation de l'opinion. Nous avons présenté trois différentes méthodes de notation des sentiments. Ces méthodes sont basées sur :

- le classificateur de comportement des groupes,
- le classificateur statistique (de "naif Bayes"),
- le classificateur linguistique.

Chaque classification aborde différemment le problème de la notation. La classification de "naif Bayes" est basée sur les techniques statistiques de catégorisation de texte [*Chapitre 2*]. La classification de comportement des groupes, basée sur l'analyse statistique effectuée sur les données linguistiques, est donc une analyse entre les méthodes statistique et linguistique. La classification linguistique est basée sur l'analyse linguistique [*Chapitre 4*].

Nous avons proposé deux nouvelles méthodes : le classificateur linguistique et le classificateur de comportement des groupes. Ensuite, nous comparons les résultats avec la classification généralement utilisée dans la domaine de l'Analyse des Sentiments [Pang *et al.* (2002)], qui consiste à l'application de la classification de subjectivité et ensuite la classification de l'intensité en utilisant les classificateurs de Bayes ou de SVM.
Le processus de notation de l'opinion est effectué de telle manière, que nous avons essayé de ne pas répéter les mêmes caractéristiques dans ces trois traitements. Nous avons effectué les travaux de recherche sur la même base d'apprentissage et la même base de tests, pour pouvoir comparer ces classificateurs.

Nous utilisons trois classificateurs séparément, et obtenons donc trois notes pour une seule critique. Ces notes peuvent être différentes. Nous ajoutons la classification finale au système pour pouvoir associer une seule note à la critique cinématographique. La classification finale est effectuée grâce à l'utilisation des réseaux de neurones.

Dans ce chapitre, nous avons présenté les chaînes de traitement pour chaque classificateur. Dans le chapitre suivant, nous allons présenter les tests effectués et nous discuterons des avantages et inconvénients de chaque classification.

Chapitre 7

L'évaluation et les tests

7.1 Les tests des classifications de notation des sentiments

7.1.1 Le choix de validation des performances

Dans le chapitre précédent nous avons présenté les techniques de classification de texte pour effectuer la notation de l'opinion des critiques cinématographiques. Dans ce chapitre nous présentons nos résultats et nous précisons les avantages et les inconvenants des méthodes décrites dans le chapitre précédent. Pour mesurer les performances des classificateurs nous calculons les paramètres du rappel et de la précision, en déduisant la valeur de F-score.

Dans notre activité de recherche nous avons utilisé le classificateur linguistique pour lequel nous avons créé les grammaires en se basant sur les critiques de la base d'apprentissage (identique pour toutes les méthodes). Pour cette raison nous avons choisi la méthode de **validation par test**. Les autres méthodes de validation [*Section 2.3*] sont basées sur l'estimation de l'erreur et utilisent les données de la base d'apprentissage. La création des grammaires pour le classificateur linguistique est basée sur la base d'apprentissage, donc pour calculer la performance nous avons besoin d'une nouvelle base, la base de test. Nous precisons aussi que nous avons un très important nombre de critiques annotées dans notre base de données ce qui justifie l'utilisation de la méthode de validation par test. Nous comparons les résultats de toutes les approches de classification developpées sur le même ensemble de validation.

Nous avons utilisé la même base de test et la même base d'apprentissage pour tous les classificateurs des sentiments. Nous supposons que l'utilisation des mêmes bases d'apprentissage et de tests nous permet d'effectuer la comparaison des résultats des trois classificateurs, même si l'apprentissage était effectué d'une manière complètement différente.

Dans notre recherche de la notation des sentiments, une des méthode utilisée est la méthode de classification de comportement des groupes. Ce classificateur attribue uniquement la note directement à la critique entière. Les autres classificateurs attribuent la note à chaque phrase de la critique. Pour pouvoir comparer les trois méthodes utilisées, la performance de tous les classificateurs est calculée par rapport à la bonne attribution de la note à la **critique entière** et non à chaque **phrase**.

La mesure de performance d'attribution de la note à la critique entière dans le cas de deux classificateurs (statistique et linguistique) peut sembler moins précise que la mesure de performance par rapport à l'attribution de la note à chaque phrase. En effet, nous effectuons la classification de chaque phrase et non pas la moyenne des notes de toutes les phrases de chaque critique cinématographique. Donc pour ces deux classificateurs nous avons aussi effectué la mesure de la performance par rapport à la note attribuée à chaque phrase. Les résultats que nous avons obtenus n'étaient pas trop éloignés de ceux que nous avons obtenus en regardant la critique entière, pourtant les sens de la précision et du rappel sont différents dans les deux mesures. Cette validation ne peut évidemment pas être effectuée avec le classificateur de comportement des groupes. Pour cette raison nous estimons que pour pouvoir comparer les résultats de toutes les classifications nous devons tenir compte de la note attribuée à la critique entière.

La mesure de la performance d'attribution de la note par rapport à **chaque phrase** (le classificateur linguistique et statistique) demande le calcul de la précision et du rappel. Pour ces calculs nous avons besoin d'avoir :

- l'ensemble de tous les documents pertinents trouvés,
- l'ensemble de tous les documents trouvés,
- l'ensemble de tous les documents pertinents présents dans la base.

Pour l'ensemble de tous les documents pertinents trouvés, nous prenons toutes les phrases subjectives qui ont eu une note associée par le classificateur égale à la note (existante dans la base de données : note d'utilisateur) de la critique de ces phrases.
Pour l'ensemble de tous les documents trouvés, nous prenons toutes les phrases auxquelles nous avons attribué une note.
Pour l'ensemble de tous les documents pertinents présents dans la base, nous prenons toutes les phrases subjectives de la critique.

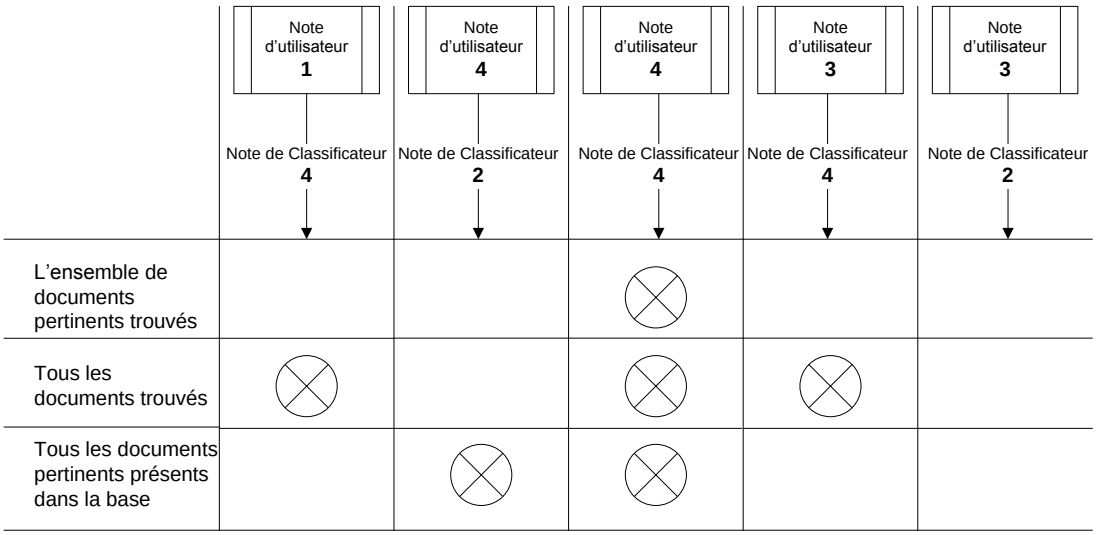


Figure 7.1: La mesure de performance d'attribution de la note par rapport à la critique entière - L'exemple montre le calcul de la précision et du rappel pour une note égale à 4

Dans le cas du calcul de la mesure de la performance d'attribution de la note par rapport à la **critique entière** pour l'ensemble de tous les documents pertinents trouvés, nous prenons toutes les critiques pour lesquelles la notation est correcte.
Pour l'ensemble de tous les documents trouvés, nous prenons toutes les critiques qui ont une note associée par la classification égale à la note pour laquelle nous avons effectué la mesure de performance.
Pour l'ensemble de tous les documents pertinents présents dans la base, nous prenons toutes les critiques qui ont une note associée par l'utilisateur égale à la note pour laquelle nous avons effectué la mesure de performance.

L'exemple de la mesure de performance par rapport à l'attribution de la note à chaque phrase est montré sur la [Figure 7.1].

Dans l'exemple présenté nous avons l'ensemble de documents pertinents trouvés dont la note est égale à 1, l'ensemble de tous les documents trouvés égale à 3 et l'ensemble de tous les documents pertinents présents dans la base égale à 2. Dans l'exemple la précision $\pi = \frac{1}{3} = 33.3\%$ et le rappel $\rho = \frac{1}{2} = 50\%$.

La base de test est constituée de 300 critiques - 60 par note. La base de test que nous avons utilisée pour calculer la performance contient :

- 828 phrases pour la note égale à 5,
- 588 phrases pour la note égale à 4,
- 657 phrases pour la note égale à 3,
- 431 phrases pour la note égale à 2,
- 1130 phrases pour la note égale à 1.

7.1.2 Le classificateur linguistique

Le classificateur linguistique utilise la base d'apprentissage pour la création des règles des grammaires locales pour chaque classe de notes. Pour effectuer la notation nous prenons une nouvelle critique de la base de test. L'attribution de la note est effectuée phrase par phrase. A la fin de processus nous obtenons un nombre des phrases avec les notes associées.

Notre base de tests contient 706 phrases objectives et 2898 phrases subjectives (744 phrases avec une note égale à 5, 533 avec une note égale à 4, 588 avec une note égale à 3, 381 avec une note égale à 2 et 893 phrases avec une note égale à 1).

	5(646)	4(458)	3(557)	2(426)	1(809)	PNC
5*(744)	539	24	44	33	29	75
4*(533)	43	377	25	27	21	40
3*(588)	15	18	399	46	22	88
2*(381)	13	12	23	238	38	57
1*(893)	12	7	39	62	681	92
PO	24	20	27	39	18	-
Précision	83.4%	82.4%	71.6%	55.9%	84.2%	-
Rappel	72.4%	70.8%	67.8%	62.5%	76.3%	-
F-score	76.5%	76.1%	69.6%	59%	80.1%	-

Tableau 7.1: Mesure de performance pour le classificateur linguistique par rapport aux phrases - en haut : la classification des phrases de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

Dans le Tableau 7.1 nous montrons les résultats pour les tests du classificateur linguistique effectués sur la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée pour chaque phrase.
La partie haute du tableau montre la classification des phrases pour chaque groupe de note. Les colonnes répresentent les notes attribuées par notre classificateur. Les lignes représentent les critiques notées par les auteurs (Exemple : 5*(744) - correspond à 744 phrases avec une note égale à 5 selon la base de test). Les colonnes répresentent les notes attribuées par notre classificateur, les valeurs dans le tableau donnent, en détail, la répartition des notes de notre classificateur par rapport aux notes des auteurs. Dans le tableau, PO désigne les phrases objectives, PNC désigne les phrases non classées.

Dans la première colonne par exemple, 5(646) correspond à 646 phrases avec une note égale à 5 selon la note de notre classificateur, où 539 phrases correspondent à des phrases classifiées correctement, 43 correspondent à des phrases classifiées avec une note égale à 5 au lieu de 4, et ainsi de suite.
Le classificateur a attribué aussi les notes pour les phrases objectives (24 phrases pour le groupe 5, 20 pour le groupe 4, 27 pour le groupe 3, 39 pour le groupe 2 et 18 phrases pour la groupe 1). Plusieurs phrases n'ont pas été notées (75 phrases pour la note de 5, 40 pour la note de 4, 88 pour la note de 3, 57 pour la note de 2 et 92 pour la note de 1). La partie basse du tableau montre les valeurs de la précision, du rappel et du f-score pour le classificateur linguistique.

Pour calculer la note de la critique entière nous calculons la moyenne des notes de toutes les phrases notées. Nous pondérons les grammaires en fonction du niveau de l'analyse linguistique de la critique présentée dans la [*Section 6.4*]. La création des grammaires locales était effectuée en ajoutant un niveau de complexité par rapport à l'analyse linguistique. Les grammaires de niveau supérieur sont plus précises, mais le rappel est très faible. La recherche est effectuée de façon à ce qu'une phrase de la critique corresponde à une grammaire d'un niveau supérieur. Les autres grammaires de même note ne sont plus appliquées pour cette phrase. Pour cette raison nous avons la certitude que les résultats de la notation obtenus avec une telle grammaire sont plus précis.

Les grammaires ainsi que leurs pondérations ont été créées manuellement. Nous avons partagé les critiques en 4 groupes en fonction de leur niveau d'analyse linguistique. Nous avons ajouté les pondérations pour chaque groupe. Des grammaires les plus précises jusqu'aux grammaires générales les poids sont respectivement de 2.0 ; 1.6 ; 1.3 ; 1. Les poids ont été choisis pour que la valeur du F-score soit la plus performante, de manière empirique.

Dans le Tableau 7.2 nous montrons les résultats du classificateur linguistique appliqué à la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée pour la critique entière.

	5(60)	4(61)	3(58)	2(55)	1(66)
5*(60)	51	4	3	1	1
4*(60)	6	47	4	2	1
3*(60)	1	6	43	7	3
2*(60)	1	3	4	40	12
1*(60)	1	1	4	5	49
Précision	85%	77%	74.1%	72.7%	74.2%
Rappel	85%	78.3%	71.7%	66.7%	81.7%
F-score	85%	77.6%	72.9%	69.6%	77.8%

Tableau 7.2: Mesure de performance pour le classificateur linguistique par rapport à la critique entière - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

Les résultats obtenus pour la mesure de performance par rapport à critique entière sont meilleurs que dans le cas de la mesure de performance par phrases. La raison en est que dans ce cas nous prenons la moyenne de toutes les phrases notées de la critique, les erreurs de la notation peuvent donc dans plusieurs cas être insignifiantes.

Le principal avantage de ce classificateur linguistique est qu'il donne de meilleurs résultats que les trois autres classificateurs. Il a cependant de nombreux inconvénients. Le plus important de ces inconvénients est que la réutilisation de ce classificateur dans un autre domaine demande la création de nouvelles règles de grammaire. La création de ces règles est effectuée manuellement et demande donc un temps important d'analyse et de test. Un autre problème important est qu'il est difficile de justifier mathématiquement que la forme des règles développées est la plus fiable. Autrement dit nous ne pouvons pas prouver que la forme, le nombre, et la complexité linguistique de nos grammaire sont les plus performants. Ces paramètres dans notre recherche ont été choisis empiriquement par des nombreux tests effectués à chaque étape du travail.

7.1.3 Le classificateur statistique

La base d'apprentissage est composée de 1000 critiques qui correspondent à 200 critiques par groupe de notation. La base est composée de 9289 phrases : 2264 phrases pour la note égale à 5, 1957 phrases pour la note égale à 4, 1308 pour la note égale à 3, 1925 pour la note égale à 2 et 1835 pour la note égale à 1.
Pour la représentation vectorielle nous avons calculé l'index complet qui est égal à 18422 mots lemmatisés. Pour la détection de l'opinion nous utilisons deux classificateurs de Bayes - le premier pour la détection de phrases subjectives et le deuxième pour la détection de l'intensité de l'opinion. Pour chaque classificateur nous effectuons la réduction de l'index complet et nous obtenons un ensemble de 705 mots pour le classificateur de subjectivité et un ensemble de 743 mots pour le classificateur d'intensité.

	Précision	Rappel	F-score
Classe - 5	67.5%	71.4%	69.4%
Classe - 4	71.8%	67.2%	69.4%
Classe - 3	64.2%	63.3%	63.7%
Classe - 2	63.4%	62.4%	62.9%
Classe - 1	69.3%	72.9%	71.1%

Tableau 7.3: Mesure de performance pour le classificateur statistique par rapport aux phrases

Nous utilisons la même base de test pour tous les classificateurs de notation de l'opinion. Le classificateur de subjectivité classe correctement 82,4% des phrases. Dans le Tableau 7.3 nous montrons les résultats du classificateur de l'intensité de l'opinion par la méthode de validation par test. La classification est effectuée phrase par phrase.

Comme nous l'avons déjà précisé, nous voulions comparer entre eux les résultats obtenus par chaque classificateur de la notation de l'opinion. Pour cette raison nous sommes obligés de calculer la performance par rapport à la notation de la critique entière. Nous avons procédé de la même manière que pour le classificateur linguistique sauf que dans ce cas nous n'avons attribué aucune pondération aux phrases. Nous avons attribué sa note à la critique en calculant la note moyenne de toutes les phrases. Dans

	5(63)	4(56)	3(58)	2(57)	1(66)
5*(60)	43	10	2	4	1
4*(60)	12	41	4	3	0
3*(60)	3	3	39	8	7
2*(60)	3	2	6	37	12
1*(60)	2	0	7	5	46
Précision	68.3%	73.2%	67.2%	64.9%	69.7%
Rappel	71.7%	68.3%	65%	61.7%	76.7%
F-score	70%	70.7%	66.1%	63.3%	73%

Tableau 7.4: Mesure de performance pour le classificateur statistique par rapport à la critique entière - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

le Tableau 7.4 nous montrons les résultats pour le classificateur statistique obtenus sur la base de test de 300 critiques cinématographiques par la méthode de validation par test.

7.1.4 Classification des sentiments par phrases

Nous avons présenté les résultats de la mesure de performance pour deux classificateurs, l'un linguistique et l'autre statistique. Ces sont les classificateurs qui traitent la critique phrase par phrase. Nous montrons la comparaison de ces deux approches en montrant la valeur de la précision [Figure 7.2], du rappel [Figure 7.3] et du F-score [Figure 7.4] pour chaque groupe de critique.

Nous pouvons constater que le classificateur linguistique donne de meilleurs résultats que le classificateur statistique. Nous avons donc réussi à appliquer l'analyse linguistique pour la mesure de l'intensité des sentiments.

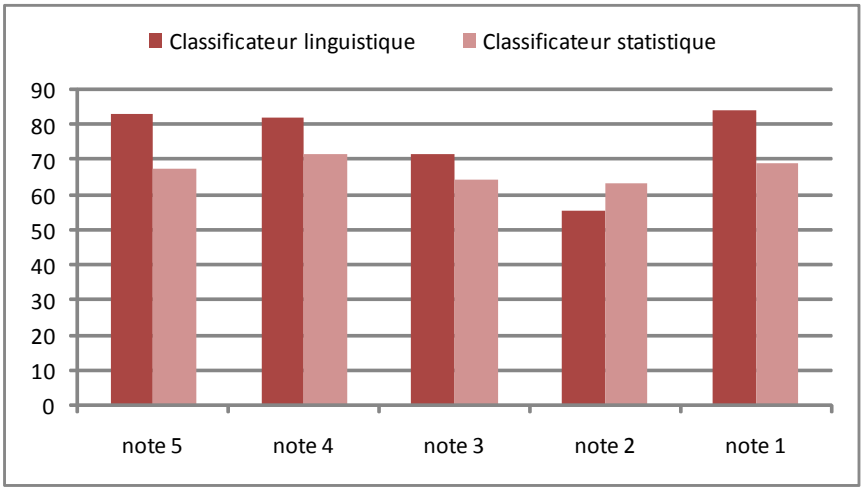


FIGURE 7.2: Précision pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant la précision (classification par phrase)

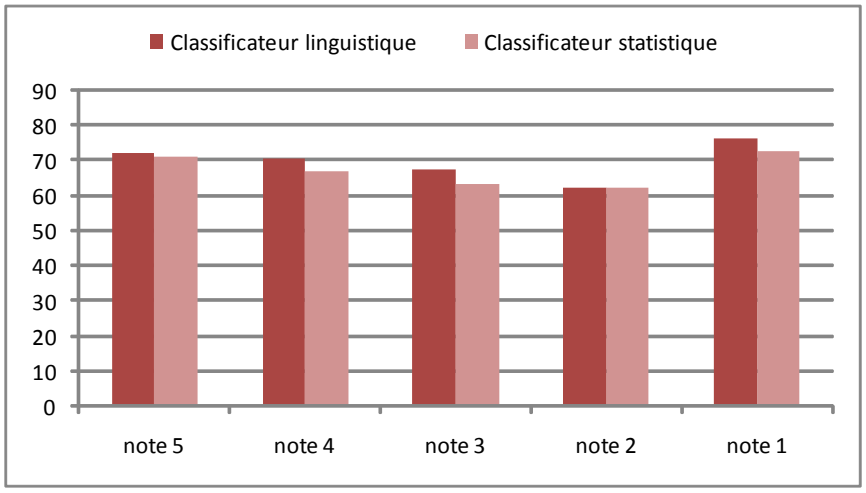


FIGURE 7.3: Rappel pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant le rappel (classification par phrase)

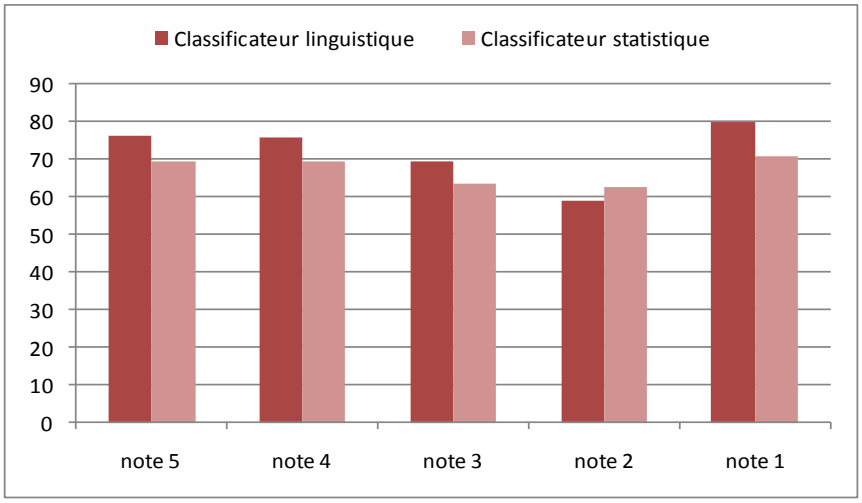


Figure 7.4: F-score pour la classification par phrases - Comparaison des résultats des classificateurs (linguistique et statistique) en mesurant la F-score (classification par phrase)

7.1.5 Le classificateur de comportement des groupes

Pour le classificateur de comportement des groupes nous avons utilisé la base d'apprentissage pour déterminer le comportement de chaque groupe composé par les critiques qui ont la même note associée. Pour effectuer le processus de la notation nous prenons une nouvelle critique de la base de test. La détermination du comportement globale de chaque groupe permet de déterminer à quel groupe appartient une nouvelle critique cinématographique. Pour les nouveaux critiques nous calculons la distance euclidienne entre ses caractéristiques et les caractéristiques des groupes. La note attribuée à la critique est celle pour laquelle la distance est la plus courte. Dans le Tableau 7.5 nous montrons les résultats du classificateur de comportement des groupes appliqué à la base de test de 300 critiques cinématographiques par la méthode de validation par le test. La mesure de performance est effectuée par la critique entière.

Comme nous pouvons le constater les résultats obtenus par cette classification sont moins fiables que ceux obtenus par la classification linguistique. Pourtant il faut préciser que les résultats sont légèrement meilleurs que ceux obtenus par la classification de Bayes. Un grand avantage de cette classification (contrairement au classificateur linguistique) est la facilité de sa réutilisation dans un nouveau domaine, celle-ci ne demandant pas beaucoup de travaille manuel. Il suffit d'appliquer une nouvelle base d'apprentis-

	5(62)	4(61)	3(56)	2(58)	1(63)
5*(60)	45	6	3	2	4
4*(60)	11	43	2	3	1
3*(60)	3	7	41	6	3
2*(60)	2	3	6	39	13
1*(60)	1	2	3	8	42
Précision	72.6%	70.5%	73.2%	67.2%	66.7%
Rappel	75%	71.6%	68.3%	65%	70%
F-score	73.8%	71%	70.8%	66.1%	68.3%

Tableau 7.5: Mesures de performance pour le classificateur de comportement des groupes - en haut : la classification des critiques de chaque groupe de notation (lignes - notes des auteurs, colonnes - notes de classification), en bas : les mesures de performance

sage pour pouvoir calculer les nouvelles caractéristiques des groupes. Le travail manuel nécessaire consiste uniquement à rechercher les mots et les expressions caractéristiques de ce nouveau domaine (qui devraient être peu éloignés de ceux présentés) et de fournir de nouveaux critères pour la recherche. Un des principaux inconvénients de cette classification est qu'elle nécessite une très grande base d'apprentissage pour pouvoir rechercher les caractéristiques de comportement des groupes. Cela n'est pas génant dans notre cas, car nous avons une très grande base de critiques cinématographiques déjà notées. L'utilisation de cette méthode dans un domaine où l'on ne dispose pas de ces ressources est remise en question car l'annotation manuelle des données demanderait beaucoup trop de temps.

7.1.6 Classification des sentiments par la critique entière

Nous avons présenté les résultats de la mesure de performance pour les trois classificateurs : linguistique, statistique et de comportement des groupes. Nous avons présenté la comparaison de toutes les approches appliquées à la notation des sentiments. Nous avons mis en valeur la comparaison de la précision [Figure 7.5], du rappel [Figure 7.6] et du F-scores [Figure 7.7] pour chaque groupe de critique.

Nous pouvons constater encore une fois que le classificateur linguistique donne de meilleurs résultats. Les résultats obtenus grâce au classificateur de comportement des groupes sont légèrement meilleurs que ceux obtenus par le classificateur statistique.

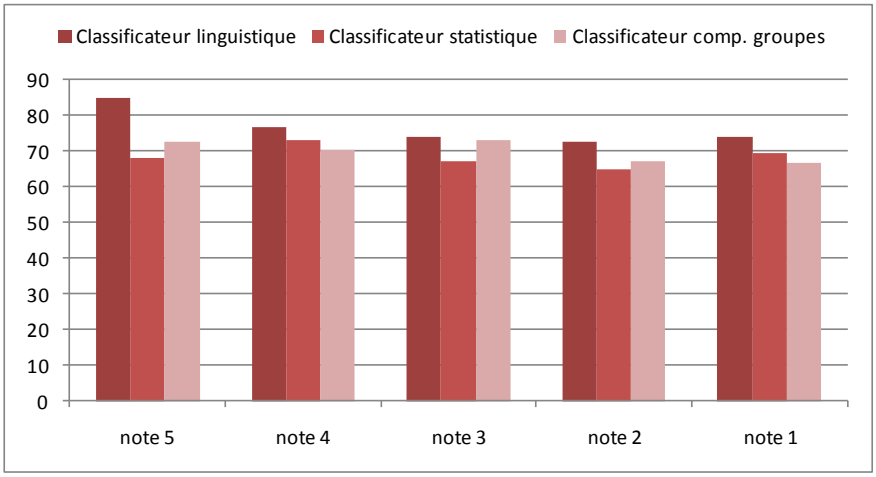


Figure 7.5: Précision pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant la précision

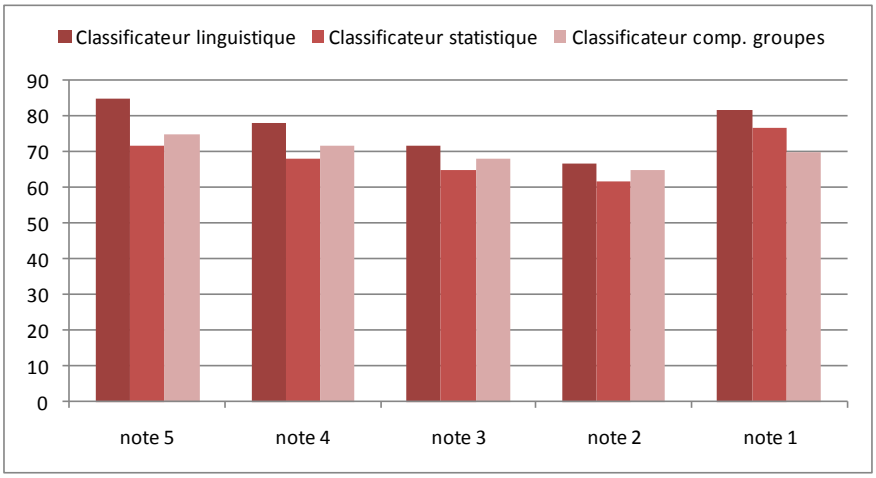


Figure 7.6: Rappel pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant le rappel

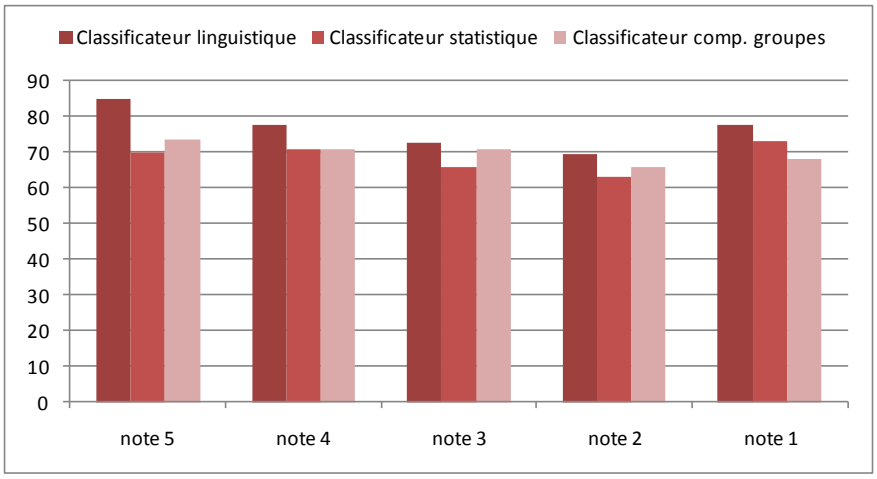


Figure 7.7: F-score pour la classification par la critique entière - Comparaison des résultats de tous les classificateurs de la notation de l'opinion en mesurant le F-score

7.2 Les tests de classification finale

Comme nous l'avons remarqué nous avons un classificateur déterministe dans plusieurs situations. Nous avons donc amélioré nos résultats obtenus en utilisant les réseaux de neurones. Pour cette étape de classification nous avons fondé notre approche uniquement sur les résultats des 3 classificateurs décrits précédemment. Les résultats finaux obtenus par le calcul de la moyenne basée uniquement sur les notes entières de chaque classificateur (1 à 5) sont moins bonnes (la précision et le rappel) que les résultats obtenus par le meilleur classificateur - classificateur linguistique. Pourtant nous avons amélioré nos résultats par l'utilisation de réseaux de neurones en prenant en considération chaque probabilité de chaque note de chaque classificateur. Ces résultats ont été améliorés d'un ordre de 4% par rapport au résultat du meilleur classificateur - le classificateur linguistique [Figure 7.8].

Le F-score calculé par rapport à la note finale est de 83,1% pour 5*, 81,2% pour 4*, 74,5% pour 3*, 72,2% pour 2* et 81,4% pour 1*. Pour l'apprentissage de ce classificateur nous avons utilisé une nouvelle base d'apprentissage et la même base de tests que pour les trois classificateurs présentés précédemment.

L'utilisation des réseaux de neurones est justifiée par la présence d'une très grande base d'apprentissage - la base des critiques déjà notées. Un avantage de ce classificateur

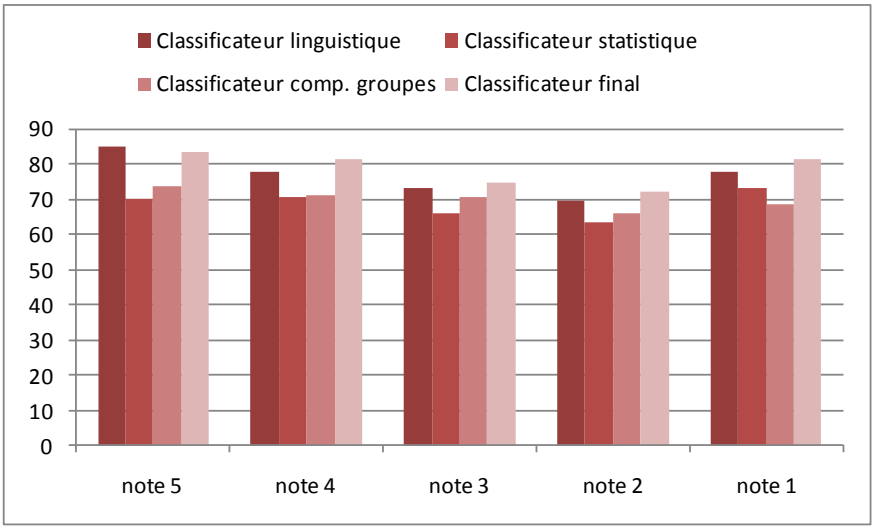


Figure 7.8: Les résultats du classificateur final - Comparaison des résultats de toutes les classifications

est qu'il ne demande pas de prétraitement spécial de données. Le classificateur final n'est pas utilisé pour la notation des sentiments. Cependant il est utilisé pour combiner les résultats obtenus des trois classificateurs présentés dans cette thèse.

7.3 Conclusion

Nous avons remarqué que nous obtenons de meilleurs résultats avec le classificateur linguistique (surtout la précision). Les moins bons résultats sont ceux du classificateur statistique de "naïf Bayes" (le rappel est correct mais la précision est faible). Cela démontre la nécessité d'une analyse linguistique profonde. Nous avons observé que les meilleurs résultats ont été obtenus dans chaque approche pour des opinions extrêmes. Il est plus facile de noter automatiquement et de juger les critiques cinématographiques ayant des notes de 1 ou 5. Cela semble évident, car les émotions extrêmes sont plus fortes et généralement la personne les exprime de manière plus intense. De plus le texte des opinions extrêmes est plus long ce qui favorise l'attribution d'une note correcte. Partant du principe qu'il est nécessaire de disposer de grammaires plus complexes, nous avons montré que le classificateur linguistique donne de meilleurs résultats que le classificateur statistique ou le classificateur de comportement de groupe.

Un point important à la vue des résultats obtenus est la réussite de l'implémentation de l'approche linguistique, ce qui démontre l'importance de l'utilisation de l'analyse linguistique dans le domaine de l'*Opinon Mining*.

Chapitre 8

Conclusion générale et perspectives

8.1 Synthèse

Le sujet de recherche de cette thèse est la notation de l'opinion. Nous avons développé un système autonome d'exploration des opinions exprimées dans les critiques cinématographiques. Les objectifs du système présenté sont :

- la collecte automatique des critiques cinématographique,
- l'attribution automatique d'une note aux critiques par rapport aux émotions décrits,
- la création et publication des profils des utilisateurs.

Le but de notre travail est la réalisation d'un système qui prépare la base de données des profils pour le moteur prédicatif (RS). Il s'agit d'attribuer une note à des critiques d'utilisateurs en appliquant les connaissances du domaine d'*Analyse des Sentiments*.

La partie la plus intéressante au niveau de la recherche est la notation automatique des sentiments. Pour cette partie nous avons présenté trois méthodes différentes pour effectuer la classification des sentiments. Nous avons nommé les méthodes présentées de la façon suivante :

- la classification linguistique,
- la classification de comportement des groupes,
- la classification statistique.

8. CONCLUSION GÉNÉRALE ET PERSPECTIVES

Les deux premières classifications ont été proposées par nos soins. Ensuite, nous avons comparé nos approches avec l'approche généralement utilisée dans ce domaine [[Pang *et al.* (2002)]] - la classification statistique qui est basée sur les classificateurs de "naïf Bayes".

Nous avons justifié le choix d'architecture proposé et utilisé dans le système développé. C'est une architecture parallèle, ce qui revient à un traitement individuel de chaque critique cinématographique par chaque classificateur présenté. Nous avons obtenu de cette façon à la sortie de ce module trois notes pour une seule critique. Afin d'avoir une seule note attribuée de la façon le plus performante, nous avons utilisé le classificateur final basé sur les réseaux de neurones.

Après les tests effectués, nous pouvons constater que nous avons réussi à implanter une première méthode innovante basée sur un classificateur linguistique. Les résultats obtenus après cette classification donnent une plus grande satisfaction. Nous pouvons donc conclure que l'analyse linguistique plus profonde est une voie importante de recherche dans le domaine de l'*Analyse de Sentiments*.

Nos travaux de recherche concernant la classification linguistique ont été effectués sur l'application *Unitex* qui permet l'intégration des grammaires, des tables de lexique-grammaire et des dictionnaires. Notre objectif était de préparer et d'implémenter des ressources linguistiques et de créer des grammaires locales complexes afin de pouvoir associer ces grammaires à des phrases des critiques cinématographiques. Chaque grammaire associe la note obtenue aux phrases des critiques cinématographiques. C'est de cette manière qu'est effectuée la notation de la critique.

Malgré le fait que le classificateur linguistique permet d'obtenir les meilleurs résultats, son utilisation ne peut pas être universelle. Son application à un nouveau domaine nécessite la création d'une nouvelle base des ressources linguistiques et il est nécessaire d'effectuer l'analyse linguistique profonde de nouveau. Ces traitements sont inévitables car le langage est très dépendant du domaine. L'analyse linguistique est effectuée manuellement et demande beaucoup de temps pour l'analyse et les tests.

La deuxième méthode proposée dans cette thèse est la classification des comportements des groupes. Cette approche est basée sur une étude statistique de données linguistiques. Nous disposons d'un très grand nombre des critiques déjà annotées par les utilisateurs, nous pouvons alors les utiliser pour retrouver le comportement caractérisant le groupe de notation. Les résultats de cette méthode montrent une assez bonne performance. Un des grands avantages de cette approche par rapport à l'approche linguistique est la facilité de son implémentation sur un autre domaine.

Nous avons présenté dans ce manuscrit les techniques existantes dans le domaine de la catégorisation du texte [*Chapitre 2*]. Ces techniques sont généralement basées sur l'analyse statistique. Nous utilisons ces techniques dans le troisième classificateur de la notation de l'opinion, basé sur le classificateur "naïf Bayes" pour comparer les résultats obtenus.
Nous avons ensuite exposé l'utilisation des techniques de catégorisation de texte dans le domaine de l'Opinion Mining, domaine qui a pour but la détection et la notation des sentiments [*Chapitre 3*].
Notre objectif principal était l'intégration du traitement linguistique dans ce domaine. Nous avons donc présenté les techniques utilisées dans le Traitement Automatique des Langues Naturelles [*Chapitre 4*].
Nous avons ensuite décrit le système mis en oeuvre pour la collecte et notation des sentiments exprimés dans les critiques [*Chapitre 5*].
Nous avons consécutivement présenté les chaines de traitement de tous les classificateurs du module de la notation [*Chapitre 6*]. Et enfin, nous avons présenté les résultats en concluant par la nécessité de l'analyse linguistique [*Chapitre 7*].

La principale utilisation du notre système est la création des profils pour le moteur de prédiction. Nous avons mis en évidence [*Section 3.2*] le besoin de connaitre les sentiments des autres personnes. Parmi les nombreuses possibilités d'utilisation des nos approches nous pouvons citer l'amélioration des multiples sites de vente en ligne, un meilleur ciblage des études de marché pour des entreprises souhaitant développer de nouveaux produits ou services, la veille technologique, ou même l'amélioration des résultats des moteurs de recherche.

La libre collecte et l'utilisation des informations sur les goûts et opinions des internautes, leurs réutilisations ou stockages soulèvent des questions d'ordre éthiques (où commence et s'arrête la vie privée ?).

8.2 Perspectives

Le système présenté dans cette thèse effectue automatiquement la collecte et la notation de l'opinion. Pourtant, il est utile de préciser que les approches que nous proposons donnent de nombreuses perspectives de recherche.

La première perspective est de trouver des moyens de déterminer la forme finale des règles de grammaire. Dans nos travaux, les formes des grammaires locales sont trouvées empiriquement. Il serait très intéressant de pouvoir décrire mathématiquement le niveau de la complexité de l'analyse linguistique ainsi que le nombre des règles nécessaires pour que la recherche soit la plus performante.

Lors de l'aboutissement de ce travail, une deuxième perspective apparaît. Il s'agit de l'automatisation partielle ou totale de l'analyse linguistique.

Nous avons mentionné plusieurs fois dans cette thèse que le langage dépend du domaine concerné. Nous avons aussi précisé que nous pouvons retirer de plus en plus d'informations sur les utilisateurs qui postent les critiques. Nous espérons donc que nous pourrons améliorer encore les résultats obtenus en tenant compte de ces informations sur les auteurs des critiques. Ces informations peuvent être très intéressantes dans la phase de classification. Nous pouvons générer l'hypothèse que la connaissance des informations sur les auteurs, comme par exemple l'âge, le sexe, le centre d'intérêt, le code postal, le statut socioprofessionnel, l'avis sur d'autre produits etc., améliore la compréhension des sentiments. Car les personnes d'un même âge, venant d'un même milieu, aimant les mêmes choses peuvent exprimer leurs goûts de façon similaire. Nous avons besoin de plus d'informations sur les utilisateurs pour effectuer les travaux de recherche qui prennent en compte ces caractéristiques. Le système présenté dans ce manuscrit

pourrait être utilisé pour extraire ces informations en collaboration avec un moteur prédicatif.

Glossaire

Agent de recherche (agent intelligent, robot, spider, wanderer, Web worm) Assistant électronique personnalisé, qui peut être paramétré. C'est un logiciel, qui accomplit un certain nombre de tâches répétitives à partir des règles de fonctionnement qui définissent son architecture. Il peut surveiller un thème à partir de différents filtres et de différentes sources, communiquer avec d'autres agents, observer un environnement que le veilleur lui commande d'explorer, etc. Il existe plusieurs niveaux de recherche :

- recherche générale : un premier survol
- recherche avancée : une observation plus fine
- recherche sectorielle : spécialisée dans un secteur
- recherche d'alertes : à la recherche de nouveautés.

Ambiguïté Caractère de ce qui présente plusieurs sens possibles. Les langues naturelles sont par nature ambiguës. Cette ambiguïté enrichit la langue, pensons aux calembours, aux phrases suscitant des quiproquos.

Ambiguïté lexicale se situe au niveau du mot ; elle est de nature catégorielle, ou sémantique (homonymie, polysémie, ..). Un logiciel de traduction, un correcteur seront performants si l'ambiguïté est résolue. Le mot marche est une ambiguïté catégorielle : la marche d'escalier ; il marche (verbe marcher).

Ambiguïté pragmatique est levée par des connaissances extra-linguistiques. Le touriste était dans l'avion et il n'a pas décollé : il = le touriste ou l'avion ?

Ambiguïté syntaxique se situe au niveau de la structure des énoncés. <Nous écoutons les bruits de la fenêtre = nous écoutons les bruits que fait la fenêtre> ou <les bruits provenant par la fenêtre> ou <nous sommes à la fenêtre et nous écoutons les bruits>. Pour l'esprit humain, cette phrase replacée dans

son contexte ne pose pas de difficulté de compréhension. Pour un programme de traduction assistée par ordinateur, d'analyse de contenu, etc., il faudra d'abord lever cette ambiguïté.

Ambiguïté sémantique se situe au niveau du sens des mots et occasionne plusieurs représentations logiques d'un énoncé. <Toutes les filles de la famille X aiment un homme> = <Elles aiment un homme différent> ou <un même homme> ?

Analyse syntaxique Analyse qui consite à assigner des étiquettes de nature syntaxique aux mots ou aux phrases. En général, en extraction d'information, les mécanismes mis en jeu se fondent uniquement sur des informations de surface

Analyse sémantique Analyse qui consiste à assigner des étiquettes de nature sémantique aux mots ou aux phrases. Les mécanismes peuvent se fonder sur des informations de surface ou sur des mécanismes plus complexes comme l'analyse de la coréférence, l'inférence ...

Analyseur Terme générique qui désigne, en Analyse de texte <un programme ou un ensemble de programmes informatiques fournissant des renseignements analytiques sur des mots donnés ou sur un ou plusieurs textes. Les informations fournies par un analyseur peuvent être - d'ordre numérique : des indices numériques sur la répartition d'un mot donné dans un texte, par exemple - d'ordre symbolique - des représentations graphiques mettant en évidence la structure syntaxique ou le sens des énoncés>. C'est la finalité de l'analyse d'un texte qui justifie la mise en oeuvre d'un type d'analyseur ou la combinaison de plusieurs analyseurs.

Analyseur lexicométrique ou lexico-statistique l'analyseur effectue des calculs sur les mots pris hors de leur contexte. L'approche est de type quantitatif. L'analyse lexicométrique permettra, par exemple, de vérifier la richesse objective d'un texte en en comparant le vocabulaire à des listes de référence.

Analyseur linguistique programme ou ensemble de programmes informatiques visant à produire des représentations (sous forme symbolique ou graphique) caractéristiques des phénomènes linguistiques (morphologie, syntaxe, sémantique et pragmatique) dans un texte, et cela dans le but d'en mettre en évidence le ou les sens. Nous trouvons les analyseurs linguistiques aussi bien dans

le domaine du traitement des données linguistiques écrites que dans celui du traitement de la parole.

Analyseur statistique ce programme ou ensemble de programmes informatiques décompose le texte en une suite de signaux numériques. Le texte est alors considéré comme un ensemble de phénomènes dont les occurrences peuvent faire l'objet d'une analyse statistique et mathématique. Des informations lexicales, stylistiques (informations de structure) sont ainsi disponibles. L'analyseur statistique peut avoir une - approche quantitative : indices sur les données brutes (nombre d'occurrences, etc...) - approche quantitative et qualitative : indices sur des données catégorisées.

Apprentissage Technique par laquelle un processus tire des connaissances de son environnement, généralement pour améliorer ses traitemenents en fonction des données. Appliqué à l'extraction d'information, l'apprentissage permet l'acquisition semi-automatique de connaissances à partir de textes (ressources linguistiques), facilitant la mise au point des systèmes.

Apprentissage interactif Type d'apprentissage au cours duquel l'utilisateur intervient régulièrement pour valider ou guider l'analyse

Apprentissage non supervisé Type d'apprentissage où le système n'a aucune connaissance préalable susceptible de le guider dans la tâche d'apprentissage

Apprentissage supervisé Type d'apprentissage où le système a un ensemble de connaissances préalables susceptibles de le guider dans la tâche d'apprentisage (par exemple, dans un système d'apprentissage de classes sémantiques, nous pouvons fournir au système une liste de mots intéressants sur lesquels le système va focaliser l'analyse

Automate Graph représentant un ensemble de séquences qui peuvent être reconnues dans les textes. Pour UNITEX, les automates à nombre fini d'etats sont un cas particulier de transducteur qui ne produit aucune information en sortie

Bruit désigne toute réponse non pertinente à une recherche documentaire (AFNOR) Le taux de bruit est le % exprimant le rapport entre le nombre de documents non pertinents extraits et le nombre total de documents extraits. Ce terme

désigne également, en Analyse de Texte par ordinateur, les résulats non désirés ou les fausses réponses parmi les résultats fournis par un programme informatique d'analyse de texte.

Catégorisation Cette procédure consiste à associer à un mot, à un groupe de mots, ou à tout objet relevant d'un texte (signes typographiques, segments de texte, caractères spéciaux), des informations d'ordre linguistique : les catégories grammaticales (nom, verbe, ...), les traits sémantiques (humain, animal, ...), les traits narratifs (argument, contre-argument, ...) et/ou d'ordre sociologique : le domaine d'emploi, la réalité sociale sous-jacente, ...

Co-occurrence Une co-occurrence est un groupe de mots apparaissant fréquemment ensemble.

Corpus Un corpus est, dans notre acceptation un ensemble de productions linguistiques (langue écrite ou langue parlée) qui partagent les mêmes conditions de production, et qui seraient donc comparables entre elles.

EA L'ensemble d'apprentissage

EA L'ensemble de validation

ET L'ensemble de test

Etiquettage Opération qui consiste à assigner une partie du discours à un mot dans un corpus

Fréquence d'un mot le nombre de fois qu'apparaît un mot donné dans un texte. La fréquence peut être exprimée en nombre absolu ou en pourcentage.

Grammaire Désigne un ensemble de règles représentant des expressions linguistique. Dans notre système les grammaires sont modélisées, en utilisant la technologie à nombre fini d'états, sous forme d'automates ou de transducteurs.

Grammaire locale Désigne une grammaire se limitant généralement à l'analyse de constituants continus. Des grammaires locales peuvent être imbriqus pour avoir un degré de localité étendu.

IMDB L'Internet Movie Database (IMDb ou Base de données cinématographiques de l'Internet) est une base de données en ligne sur le cinéma mondial, restituant

les informations concernant les films, les acteurs, réalisateurs, scénaristes et toutes personnes et entreprises intervenant dans l'élaboration d'un film, d'un téléfilm, d'une série TV ou d'un jeu vidéo. L'accès aux informations publiques est gratuit. Un service payant IMDbPro donne accès aux informations intéressant les professionnels.

Index Liste des éléments, sous forme de mots-clés, contenus dans un domaine donné. Il peut devenir une liste de références permettant de localiser ces éléments dans un serveur.

Indexation Procédure destinée à décrire et à caractériser le contenu d'un texte ou d'un document (ou d'une partie d'un texte ou d'un document) à l'aide de mots (dits mots-clés ou descripteurs) représentant les concepts ou informations dans ce document, et ce, à des fins de classification et pour en faciliter la recherche. De fait, un processus d'indexation comporte, d'une part, la reconnaissance et l'extraction des concepts informatifs (ou mots porteurs d'informations) et, d'autre part, la traduction de ces concepts dans un langage (dit langage documentaire) approprié à la classification et à la recherche de documents.

Information Élément de sens perceptible. Ensemble cohérent qui constitue pour l'utilisateur une unité de connaissances. L'information entre dans la formation et l'acquisition de la connaissance, celle-ci étant la compréhension que l'on a d'une situation ou d'un événement.

Informations non structurées Les données non structurées sont des informations qui ne sont ni classées, ni identifiées, comme par exemple les documents sur votre ordinateur ou sur le WEB. Les documents à gérer sont composés de données textuelles, numériques, sonores, graphiques, et des images (fixes ou animées). Ce type d'information représente environ les neuf dixièmes de l'information utilisée dans les organisations.

Informations structurées Les données structurées sont stockées dans des BD hiérarchiques, relationnelles, orientées objet, bibliographiques ou autres. Les informations sont ainsi identifiables et leurs chemins d'accès sont déterminés à l'avance par le concepteur de la BD. Les documents structurés sont des documents électroniques représentés selon un format structuré, c'est-à-dire un

format qui utilise des balises pour décrire la structure logique des documents. Souvent, la structure logique d'un document sera sa division en parties, chapitres, sections, etc., de même que certaines autres unités telles des notes de bas de page ou des références bibliographiques. La norme XML permet de définir des documents structurés.

Informatique documentaire Désigne l'utilisation des moyens informatiques à des fins d'assistance, ou de résolution de problèmes, dans les domaines de la documentation et du traitement de l'information. Ainsi, que ce soit au niveau des analyses des textes ou des documents (par exemple, l'analyse du contenu des textes à des fins de repérage de mots porteurs d'informations) ou au niveau de l'activité documentaire (par exemple, la catégorisation du contenu des textes), on trouve de nombreux outils technologiques permettant la recherche, l'extraction, l'archivage et la gestion d'informations, l'analyse et le traitement de la documentation, etc.

Ingénierie linguistique (language engineering) est l'application de la connaissance des langues à l'élaboration de systèmes informatiques capables de reconnaître, de comprendre, d'interpréter et de produire du langage humain sous toutes ses formes.

IR (ang. Information Retrieval) la recherche d'information est la science qui consiste à rechercher l'information dans des documents - les documents eux-mêmes ou les métadonnées qui décrivent les documents -, dans des bases de données - qu'elles soient relationnelles ou mises en réseau par des liens hypertexte comme dans le World Wide Web, l'internet, et les intranets, pour le texte, le son, les images, les données. Au sens large, la recherche d'information inclut deux aspects l'indexation des corpus, et l'interrogation du fonds documentaire ainsi constitué.

Lemmatisation consiste à donner à un mot (accordé, conjugué) une forme canonique (forme de base = le lemme) pour, entre autres, qu'il puisse entrer dans un dictionnaire.

Lemme désigne la forme de référence d'un mot, c'est-à-dire la forme du mot sans les marques (dites marques de flexion) qui l'actualisent dans le discours.

Lexique Recueil de mots et d'informations qui leur sont associées, comme leur forme grammaticale, leur structure sonore ou leur signification en contexte

Linguistique la science du langage, qu'elle étudie à travers la diversité des langues naturelles parlées sur la Terre. Le statut scientifique de la linguistique implique un certain nombre de contraintes sur la méthode. En général, on procède par la proposition de modèles qu'on essaie de tester contre des données pour les infirmer. A la lumière des faiblesses découvertes, on modifie le modèle pour le tester encore, et ainsi de suite.

ML (ang. Machine Learning) L'Apprentissage Automatique est un des champs d'étude de l'intelligence artificielle. L'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. La reconnaissance de caractères est une tâche complexe car deux caractères similaires ne sont jamais exactement égaux. On peut concevoir un système d'apprentissage automatique qui apprend à reconnaître des caractères en observant des exemples, c'est à dire des caractères connus.

Mot-clé un mot ou groupe de mots choisi librement dans le titre ou dans le corps d'un texte (ou d'un document), permettant d'en caractériser le contenu et d'en faciliter la classification et la recherche. Le mot-clé peut aussi éventuellement faire partie d'une liste fermée de mots (dits descripteurs) utilisée à des fins de description et de recherche de documents (on parle de vocabulaire contrôlé).

Moteur de recherche Outil pour des recherches précises grâce à des mots clés spécifiques. Il est constitué de deux éléments : un robot (spider) qui visite les sites et un système d'indexation qui, à partir de filtres, analyse leurs contenus. Certains outils utilisent des opérations booléennes comme AND, OR, NOT, etc. Il y a quatre types de moteurs de recherche : par mots-clés, par thèmes, par cartes, par méta-sites.

Occurrence un élément linguistique ou un mot, toutes les fois qu'il apparaît dans un texte. Ainsi, l'apparition du mot informatique dans un texte constitue une

occurrence du mot informatique ; de même, ordinateur et ordinateurs constituent deux occurrences du mot ordinateur. Les mots ou éléments linguistiques qui figurent en même temps aux côtés de l'occurrence dans le texte sont les cooccurrences, ou sont dits ses cooccurents.

Segmentation Opération qui consiste à découper un texte en phrases, mots ou groupes de mots afin de pouvoir ensuite les analyser

Silence l'absence totale de résultat, suite à une analyse supposée faite par un programme informatique d'analyse de texte.

Syntagme est formé d'une ou de plusieurs chaînes de caractères séparées par des blancs. Un syntagme peut contenir des mots vides mais il doit comporter au moins un mot significatif.

TC (ang. Text Categorization) Catégorisation de Texte est une spécialisation de la fouille de données et fait partie du domaine de l'intelligence artificielle. C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.

Terme de l'information (documentation) Un mot ou groupe de mots en ce qu'il est succeptible de décrire ou de caractériser sans ambiguïté le contenu d'un texte ou d'un document, et cela, à des fins de classification et de repérage du texte ou du document. Autrement dit, il s'agit d'un mot dont le sens n'est pas nécessairement univoque, donc qui peut avoir des équivalents (ou synonymes), mais qui, avant tout, catégorise le mieux le contenu du texte ou du document.

Terme linguistique (terminologie) Un mot ou groupe de mots qui ne s'applique qu'à un et un seul objet ou concept, et ce, dans un domaine donné. Autrement dit, il s'agit d'un mot qui désigne de façon univoque un objet ou un concept dans un domaine, qui, généralement est un domaine de spécialité. De fait, dans l'idéal terminologique, un terme n'a pas d'équivalent (ou synonyme) dans le domaine désigné. Un terme formé d'un seul mot (par exemple, avion) est dit terme simple ou uniterme, alors que celui constitué de plusieurs mots

est appelé terme complexe ou multi-terme (par exemple, avion à réacteur), bien que le mot terme soit couramment employé dans les deux cas. On parle également d'unité terminologique.

Texte une séquence ou suite de caractères appartenant à une langue naturelle (on parle de lexique), respectant les règles de fonctionnement de la langue (dites grammaire) et de structuration du document dans cette langue. En d'autres termes, le texte est une collection de mots organisés non seulement sur le plan linguistique, mais aussi selon une norme de présentation donnée, pour être objet d'interprétation de nature intelligente chez l'humain.

TF (ang. Text Filttering) Filtrage de texte est l'activité de classification d'un flux de documents expédiés de manière asynchrone par un producteur d'information à destination d'un consomateur d'information.

Thésaurus – ensemble de mots ou de termes (dits descripteurs) constituant un vocabulaire défini (vocabulaire contrôlé de termes), ayant entre eux des relations d'ordre sémantique (par exemple, une relation hiérarchique orientée du générique vers le spécifique) ou pragmatique, et qui s'appliquent à un ou plusieurs domaines de la connaissance.

- Les relations entre les termes représentent un corpus sémantique d'un domaine et tiennent compte de l'évolution du domaine concerné. Le thésaurus est donc un outil en construction permanente. Comme le signalait Hudon, le thésaurus se doit d'être un instrument de travail éminemment flexible et adaptable.
- ... vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relations générique-spécifique).

Thésaurus électronique L'usage de l'informatique favorise la manipulation d'un thésaurus : le terme recherché est entouré des termes reliés qui sont tous des liens hypertextuels qui nous ramènent à leurs réseaux. Selon la norme ISO [ISO :2788-1986], le contenu d'un thésaurus peut être représenté en trois modes principaux :

- présentation alphabétique ;

- systématique (organisation en domaines ou disciplines et organisation par facettes, ou la combinaison des deux) ;
- présentation graphique (schéma fléché ou disposition graphique). Les deux derniers types sont accompagnés d'un index alphabétique.

Traduction automatique Traduction automatique par ordinateur de textes d'une langue naturelle à une autre.

Traitement automatique des langues naturelles (TALN) Pour que les ordinateurs puissent analyser, générer, traduire, interroger, traiter, manipuler des textes, de nombreuses connaissances sur le langage naturel sont requises : la prononciation, l'orthographe, la signification, l'emploi des mots ; la combinaison des mots pour donner un sens à une phrase, etc.

Traitement de l'information En informatique : traitement électronique des données à l'aide d'un langage de programmation. En linguistique : traitement sémantique du contenu.

Unité textuelle tout élément ou signe constitutif d'un texte. Par exemple, tous les mots du texte suivant sont des unités textuelles : Jean a mangé la soupe. En somme, il s'agit de tout caractère ou chaîne caractères reconnus et traitables par le système d'analyse de texte en regard du texte et seulement par rapport au texte ; contrairement aux unités lexicales ou aux lexèmes qui sont plutôt des éléments du lexique. Autrement dit, est désignée comme unité textuelle les éléments qui sont soumis à des opérations de description ou de catégorisation, mais seulement à l'égard de leur contexte.

Vocabulaire contrôlé (ou mots-clé ou descripteur) désigne une liste fermée de mots reconnus comme étant le vocabulaire technique d'un domaine de spécialité donné. Il sert à : la description du contenu d'un texte l'indexation la recherche d'information. L'utilisation du vocabulaire contrôlé sert ainsi à limiter les problèmes relatifs aux traitements et à la recherche d'informations pouvant résulter d'une catégorisation ou d'une classification des documents en langage naturel ou vocabulaire libre. On parle aussi de lexique contrôlé ou de liste de termes normalisés.

WSD (ang. Word Sense Disambiguation) C'est un processus d'identification du sense d'un mot dans la phrase. Voire Ambiguïté

References

Alshawi, H. 1992. *The core language Engine.* MIT Press - ACL-MIT Press Series in Natural language Processing. 48

Amati, G., & Crestani, F. 1999. Probabilistic learning for selective dissemination of information. *Inform. Process. Man.*, **35**, 633–654. 11

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. 2000. An experimental comparison of naive Bayesian and keywordbased anti-spam filtering with personal e-mail messages. *23rd ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-00**, 160–167. 10

Apt'e, C., Damerau, F. J., & Weiss, S. M. 1994. Automated learning of decision rules for text categorization. *ACM Trans. on Inform. Syst.*, **12**, 233–251. 19

Attardi, G., Di Marco, S., & Salvi, D. 1998. Categorization by context. *J. Univers. Comput. Sci*, **4**, 719–736. 11

Balvet, A. 2001. Grammaires locales et lexique-grammaire pour le filtrage dŠinformation Vers une (re)utilisabilite des ressources linquistique pour la recherche dŠinformation. *In : Conference TIA, Nancy.* 53

Beineke, P., Hastie, T., & Vaithyanathan, S. 2004. Exploring Sentiment Summarization. *In : Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications.* 39

Belkin, N. J., & Croft, W. B. 1992. Information filtering and information retrieval : two sides of the same coin ? *Commun. ACM 35*, **12**, 29–38. 10

REFERENCES

Besancon, R., & Rajman, M. 2000. Le modele DSIR : une approchea base de semantique distributionnelle pour la recherche documentaire. *Revue TAL*, **41**, 1–27. 22

Borko, H., & Bernick, M. 1963. Automatic document classification. *J. Assoc. Comput. Mach*, **10**, 151–161. 9

Cavnar, W. B., & Trenkle, J. M. 1994. N-grambased text categorization. *3rd Annual Symposium on Document Analysis and Information Retrieval*, **SDAIR-94**, 161–175. 11, 17

Cleverdon, C. 1984. Optimizing convenient online access to bibliographic databases. *Inform. Serv. Use*, **4**, 37–47. 9

Cohen, W. W., & Singer, Y. 1999. Contextsensitive learning methods for text categorization. *ACM Trans. Inform. Syst.*, **17**, 141–173. 24, 27

comScore/the Kelsey group. 2007 (November). *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. Press Release. `http://www.comscore.com/press/release.asp?press=1928`. 36

Cover, T.M., & Thomas, J.A. 1991. *Elements of Information Theory*. John Wiley. 85

Dagan, I., Karov, Y., & Roth, D. 1997. Mistakedriven learning in text categorization. *2nd Conference on Empirical Methods in Natural Language Processing*, **EMNLP-97**, 55–63. 29

Das, S., & Chen, M. 2001. Yahoo! for Amazon : Extracting Market Sentiment from Stock Message Boards. *In : Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*. 35, 43

Dave, K., Lawrence, S., & M., Pennock D. 2003. Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews. *Pages 519–528 of : Proceedings of WWW*. 35, 41, 44

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. 1990. Indexing by latent semantic indexing. *J. Amer. Soc.Inform.Sci*, **41**, 391–407. 21

Drucker, H., Vapnik, V., & Wu, D. 1999. Automatic text categorization and its applications to text retrieval. *IEEE Trans. Neural Netw.*, **10**, 1048–1054. 10, 25

Dumais, S. T., Platt, J., Heckerman, D., & Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. *7th ACM International Conference on Information and Knowledge Management*, **CIKM-98**, 148–155. 19, 41

Dziczkowski, G. 2005. *Analyse et développement d'un outil de filtrage de données.* Master thesis, l'Université de Marne-la-Valée. 50, 58

Dziczkowski, G., & Wegrzyn-Wolska, K. 2007a. Graph based system purpose - built for automatic retrieval and extraction of the electronics data. *In : Internet and Multimedia Systems and Applications. ACTA Press.* 58, 67

Dziczkowski, G., & Wegrzyn-Wolska, K. 2007b. Rcss - rating critics support system purpose built for movies recommendation. *In : Advances in Intelligent Web Mastering. Springer.* 57, 68, 85

Dziczkowski, G., & Wegrzyn-Wolska, K. 2008a. An autonomous system designed for automatic detection and rating of film. Extraction and linguistic analysis of sentiments. *In : Accepted to publication in : Proceedings of IEEE/WIE/ACM International Conference of Web Intelligence, Sydney.* 57, 61

Dziczkowski, G., & Wegrzyn-Wolska, K. 2008b. Tool of the intelligence economic : Recognition function of reviews critics. *In : ICSOFT 2008 Proceedings. INSTICC Press.* 57, 61

Escudero, G., M'arquez, L., & Rigau, G. 2000. Boosting applied to word sense disambiguation. *11th European Conference on Machine Learning*, **ECML-00**, 129–141. 11

Forsyth, R. S. 1999. New directions in text categorization. *n Causal Models and Intelligent Data Management*, **1**, 151–185. 11

Freitag, D. 1998. *Machine Learning for Information Extraction in Informal Domains.* Ph.D. thesis, Carnegie Mellon University. 57

Fuhr, N., & Knorz, G. 1984. Retrieval test evaluation of a rule-based automated indexing (AIR/PHYS). *7th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-84**, 391–408. 9

REFERENCES

Fuhr, N., Hartmann, S., Knorz, G., Lustig, G., Schwantner, M., & Tzeras, K. 1991. AIR/XŮa rule-based multistage indexing system for large subject fields. *3rd International Conference ŞRecherche dŠInformation Assistee par OrdinateurŤ*, **RIAO-91**, 606–623. 27

Furnkranz, J. 1999. Exploiting structural information for text classification on the WWW. *3rd Symposium on Intelligent Data Analysis*, **IDA-99**, 487–497. 11

Gale, W. A., Church, K. W., & Yarowsky, D. 1993. A method for disambiguating word senses in a large corpus. *Comput. Human.*, **26**, 415–439. 11

Gardent, C., Guillaume, B., Falk, I., & Perrier, G. 2005. Le lexique-grammaire de M.Gross et le traitement automatique des langues. *In : LORIA & ATILF.* 55

Gray, W. A., & Harley, A. J. 1971. Computerassisted indexing. *Inform. Storage Retrieval*, **7**, 167–174. 9

Gross, M. 1997. The construction of local grammars. *Finite-State Languauga Processing,*, **MIT Press**, 329–354. 53

Hatzivassiloglou, V., & McKeown, K. 1997. Predicting the Semantic Orientation of Adjectives. *Pages 174–181 of : Proceedings of the Joint ACL/EACL Conference.* 42

Hatzivassiloglou, V., & Wiebe, J. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *In : Proceedings of the International Conference on Computational Linguistics (COLING).* 39, 42

Hayes, P. J., Andersen, P. M., Nirenburg, I. B., & Schmandt, L. M. 1990. Tcs : a shell for content-based text categorization. *6th IEEE Conference on Artificial Intelligence Applications*, **CAIA-90**, 320–326. 10, 12

Heaps, H. 1973. A theory of relevance for automatic document classification. *Inform. Control 22*, **3**, 268–278. 9

Hoffman, T. 2008 (February). *Online reputation management is hot — but is it ethical ?* Computerworld. 36

Horrigan, J.A. 2008. *Online shopping.* Pew Internet & American Life Project Report. 36

Hull, D. A. 1994. Improving text retrieval for the routing problem using latent semantic indexing. *17th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-94**, 282–289. 22

Iyer, R. D., Lewis, D. D., Schapire, R. E., Singer, Y., & Singhal, A. 2000. Boosting for document routing. *ACM International Conference on Information and Knowledge Management*, **CIKM-00**, 70–77. 11

Joachims, T. 1998. Text categorization with support vector machines : learning with many relevant features. *10th European Conference on Machine Learning*, **ECML-98**, 137–142. 25, 27

Joachims, T. 1999. Transductive inference for text classification using support vector machines. *16th International Conference on Machine Learning*, **ICML-99**, 200–209. 25, 33

Joachims, T., & Sebastiani, F. 2002. Guest editorsŠ introduction to the special issue on automated text categorization. *J. Intell. Inform. Syst.*, **18**, 103–105. 8

Kamp, H. 1981. Evenemts, representations discursives et reference temporelle. *In : Langages, nb 64*. 48

Kessler, B., Nunberg, G., & Schutze, H. 1997. Automatic detection of text genre. *35th Annual Meeting of the Association for Computational Linguistics*, **ACL-97**, 32–38. 11

Kim, Y.H., Hahn, S.Y., & Zhang, B.T. 2000. Text filtering by boosting naive Bayes classifiers. *23rd ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-00**, 168–175. 11

Knight, K. 1999. Mining online tex. *Commun. ACM 42*, **1**, 58–61. 8

Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. *21st ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-98**, 90–95. 11

Larkey, L. S. 1999. A patent search and classification system. *Conference on Digital Libraries*, **DL-99**, 179–187. 10

Lewis, D. D. 1992a. An evaluation of phrasal and clustered representations on a text categorization task. *15th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-92**, 37–50. 19

Lewis, D. D. 1992b. *Representation and Learning in Information Retrieval.* Ph.D. thesis, Department of Computer Science, University of Massachusetts. 17

Lewis, D. D. 1995. Evaluating and optmizing autonomous text classification systems. *18th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-95**, 246–254. 23

Lewis, D. D., & Gale, W. A. 1994. A sequential algorithm for training text classifiers. *17th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-94**, 3–12. 25

Lewis, D. D., & Haues, P. J. 1994. Guest editorial for the special issue on text categorization. *In : ACM Trans. Inform. Syst.* 8

Li, Y. H., & Jain, A. K. 1998. Classification of text documents. *Comput. J. 41*, **8**, 537–546. 27

Liddy, E. D., Paik, W., & Yu, E. S. 1994. Text categorization for multiple users based on semantic features from a machine-readable dictionary. *ACM Trans. Inform. Syst.*, **12**, 278–295. 11

Liu, B. 2006. *Web data mining ; Exploring hyperlinks, contents, and usage data.* Springer. Chap. 11 : Opinion Mining. 35

Maron, M. 1961. Automatic indexing : an experimental inquiry. *J. Assoc. Comput. Mach. 8*, **3**, 404–417. 9

Mitchell, T.M. 1996. *Machine Learning.* McGraw Hill. 12, 14, 27

Mullen, T., & Collier, N. 2004 (July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *Pages 412–418 of : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Poster paper. 42

Myers, K., Kearns, M., Singh, S., & Walker, M. A. 2000. A boosting approach to topic spotting on subdialogues. *17th International Conference on Machine Learning*, **ICML-00**, 655–662. 11

Na, J.C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. 2004. Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. *Pages 49–54 of : Conference of the International Society for Knowledge Organization (ISKO)*. 43

Ng, H. T., Goh, W. B., & Low, K. L. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. *20th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-97**, 67–73. 29

Oh, H.J., Myaeng, S. H., & Lee, M.H. 2000. A practical hypertext categorization method using links and incrementally available class information. *23rd ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-00**, 264–271. 11

Pang, B., & Lee, L. 2004. A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Pages 271–278 of : Proceedings of the Association for Computational Linguistics (ACL)*. 43

Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up ? Sentiment Classification using Machine Learning Techniques. *Pages 79–86 of : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 36, 38, 41, 42, 43, 98, 116

Paumier, S. 2000. *Recherche dŠexpressions dans de grands corpus : le systeme AGLAE.* Master thesis, . Universite de Marne-la-Valee. 50

Paumier, S. 2003. *De La reconnaissance de formes linquistique a lŠanalyse syntaxique.* Ph.D. thesis, Marne-la-Valee. 50

Paumier, S. 2004. *Unitex 1.2 Manuel dŠutilisation. PAUMIER S.* Universite Marne la Valee. 50, 52

Pazienza, M. T. 1997. Information Extraction. *In : Lecture Notes in Computer Science Vol. 1299*. 8, 47

Plantie, M. 2006. *Extraction automatique de connaissances pour la decision multicritere.* Ph.D. thesis, Ecole Nationale Superieure des Mines Saint-Etienne. 23, 86

Porter, W. A. 1980. Synthesis of polynomic systems. *j-SIAM-J-MATH-ANA*, **11**, 308–315. 18

Quinlan, J. R. 1993. *Programs for Machine Learning.* Morgan Kaufmann. 28

Rijsbergen, C. J. V. 1979. *Information Retrieval.* Butterworths. 34

Riloff, E. 1993. Automatically constructing a Dictionary for Information Extraction Tasks. *Proceedings of 11th National Conference on Artificial Intelligence*, **AAAI'93**, 811–816. 56

Riloff, E., & Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. *In : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 45

Riloff, E., Patwardhan, S., & Wiebe, J. 2006. Feature Subsumption for Opinion Analysis. *In : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 42

Robertson, S. E., & Harding, P. 1984. Probabilistic automatic indexing by learning from human indexers. *J. Document. 40*, **4**, 264–270. 9

Sabach, G. 2001. Sens et traitements automatique des lanque. *Ingenierie des langues*, **PIERREL**, 77–129. 47

Sable, C. L., & Hatzivassiloglou, V. 2000. Textbased approaches for non-topical image categorization. *Internat. J. Dig. Libr.*, **3**, 261–275. 11, 33

Salton, G., & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inform. Process. Man.*, **24**, 513–523. 19, 20

Salton, G., Wong, A., & Yang, C. 1975. A vector space model for automatic indexing. *Commun. ACM*, **18**, 613–620. 20, 21

Salton, G., McGill M. 1983. *Introduction to modern information retrieval.* McGaw Hill Publications. 20, 48

Saracevic, T. 1975. Relevance : a review of and a framework for the thinking on the notion in information science. *J. Amer. Soc. Inform. Sci.*, **6**, 321–343. 9

Schapire, R. E., & Singer, Y. 2000. BoosTexter : a boosting-based system for text categorization. *Mach. Learn.*, **39**, 135–168. 11

Schmid, H. 1994. Part-of-Speech Tagging with Neural Network. *15th conference on Computational linguistics*, **1**, 172–176. 18

Schutze, H. 1998. Automatic word sense discrimination. *Computat. Ling.*, **24**, 97–124. 22

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys,*, **Vol. 34**, 1–47. 8, 9, 10

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal*, **27**, Bell System Technical Journal. 17

Silberztein, M. 1993. *Dictionnaires electronique et analyse automatique de texte, le sesteme INTEX.* Masson. 50, 51

Soderland, S., Fisher, D., Aseltine, J., & Lehnert, W. 1995. Crystal : Inducing a Conceptual Dictionary. *In : Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* 56

Sowa, J. 1984. *Conceptual Structures. Information processing in Mind and Machine.* Addison Wesley Publishing CO. 48

Taira, H., & Haruno, M. 1999. Feature selection in SVM text categorization. *16th Conference of the American Association for Artificial Intelligence*, **AAAI-99**, 480–486. 25

Tarveen, L., & Littman, M. 2001. Beyond recommender systems : helping people help each other. *In : HCI in the millennium. Addison-Wesley.* 59

Tauritz, D. R., Kok, J. N., & Spronkhuizen-Kuyper, I. G. 2000. Adaptive information filtering using evolutionary computation. *Inform. Sci.*, **122**, 121–140. 11

Tong, R. M. 2001. An Operational System for Detecting and Tracking Opinions in Online Discussion. *In : Proceedings of the Workshop on Operational Text Classification (OTC).* 35

Turney, P. 2002. Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Pages 417–424 of : Proceedings of the Association for Computational Linguistics (ACL).* 35, 42

Tzeras, K., & Hartmann, S. 1993. Automatic indexing based on Bayesian inference networks. *16th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-93**, 22–34. 9

Voorhess, E.M. 1999. Natural language processing and information retrieval. *Information extraction, toward scalable, adaptable systems*, **Lecture Notes in Computer Science**, 32–48. 47

Wang, Y., Hodges, J., & Tang, B. 2005. Classification of Web Documents using Naive Bayes Method. *IEEE*, **1**, 560–564. 25

Weigend, A. S., Wiener, E. D., & Pedersen, J. O. 1999. Exploiting hierarchy in text catagorization. *Inform. Retr.*, **1**, 193–216. 22

Whitelaw, C., N., Garg, & Argamon, S. 2005. Using appraisal groups for sentiment analysis. *Pages 625–631 of : Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM).* ACM. 42

Wiebe, J, & Mihalcea, R. 2006. Word Sense And Subjectivity. *In : Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL).* 40

Wiebe, J.M., Wilson, T., & Bell, M. 2001. Identifying Collocations for Recognizing Opinions. *In : Proceedings of the ACL/EACL Workshop on Collocation : Computational Extraction, Analysis, and Exploitation.* 39

Wiebe, J.M., Wilson, T., Bruce, R., Bell, M., & Martin, M. 2004. Learning Subjective Language. *Computational Linguistics*, **30**(3), 277–308. 39

Wiener, E. D., Pedersen, J. O., & Weigend, A. S. 1995. A neural network approach to topic spotting. *4th Annua Symposium on Document Analysis and Information Retrieval*, **SDAIR-95**, 317–332. 24

Wilson, t., Wiebe, j., & Hwa, r. 2004. Just How Mad Are You ? Finding Strong and Weak Opinion Clauses. *Pages 761–769 of : Proceedings of AAAI.* Extended version in *Computational Intelligence* 22(2, Special Issue on Sentiment Analysis) :73–99, 2006. 40

Wilson, T., Wiebe, J., & Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Pages 347–354 of : Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).* 39

Woods, W.A. 1973. Progress in natural language understanding : An application to lunar geology. *The American Federation of Information Processing Saocieties Conference Proceedings*, **AFIPSŠ1973**, 441–450. 48

Yang, Y. 1995. Noise reduction in a statistical approach to text categorization. *18th ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-95**, 256–263. 22

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Inform. Retr.*, **1**, 69–90. 24, 33

Yang, Y., & Liu, X. 1999. A re-examination of text categorization methods. *22nd ACM International Conference on Research and Development in Information Retrieval*, **SIGIR-99**, 42–49. 25

Yang, Y., Slattery, S., & Ghani, R. 2002. A study of approaches to hypertext categorization. *J. Intell. Inform. Syst.*, **18**, 219–241. 11

Yu, H., & Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions : Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 40

Yu, K. L., & Lam, W. 1998. A new on-line learning algorithm for adaptive text filtering. *7th ACMInternationalConference on Information and Knowledge Management*, **CIKM-98**, 156–160. 11

Zabin, J., & Jefferies, A. 2008 (January). *Social Media Monitoring and Analysis : Generating Consumer Insights from Online Conversation.* Aberdeen Group Benchmark Report. 37

9

Annexe

9.1 Expressions régulières, automates et transducteurs dans Unitex

Le système Unitex permet de construire des expressions régulières, des automates et des transducteurs à nombre fini d'états pour le traitement automatique des langues. Nous présenterons ici certains aspects formels des langages à nombre finis d'états.

9.1.1 Rappels sur les langages formels

Soit un ensemble fini non vide A, que l'on appelle alphabet. Toute suite finie d'éléments de A est un mot. Le mot vide, noté ϵ, est une suite qui ne comporte aucun élément. Etant donnés deux mots x et y, on peut leur associer un troisième mot qui est obtenu par la concaténation de x et y. On note xy la concaténation de x et y. L'ensemble de tous les mots composés d'éléments de l'alphabet A est le monoïde libre sur A, noté A*.

9.1.1.1 Langage

Tout sous-ensemble d'un monoïde libre A* constitue un langage formel défini sur A. On connaît les opérations suivantes sur les langages :

- Intersection : à deux langages L1 et L2, on associe un troisième langage formé par l'intersection ensembliste de L1 et L2. Ce langage est noté L1∩L2.
- Union : à deux langages L1 et L2, on associe un troisième langage formé par l'union ensembliste de L1 et L2. Ce langage est noté L1∪L2 ou L1+L2.

- Complément : à un langage L1 et L2, on associe un autre langage noté L ou A* - L, composé de tous les mots de A* qui ne sont pas dans L.
- Produit : à deux langage L1 et L2, on associe un troisième langage noté L1L2 tel que pour tout mot x de L1, pour tout mot y de L2, xy appartient à L1L2.
- Opération étoile : à un langage L, on associe un langage noté L* tel que, pour tout nombre entier n, si x un mot de L, la concaténation x avec lui-même (n-1) fois forme un mot de L*.

9.1.1.2 Expressions régulières

Une expression régulière sur un alphabet A est définie de la manière suivante :

1. le mot vide ϵ est une expression régulière
2. si a est un élément de A, alors a est une expression régulière
3. si R est une expression régulière, alors R* est une expression régulière
4. si R et S sont des expressions régulières, alors RS et R+S sont des expressions régulières (concaténation et union)

9.1.1.3 Automates

Un automate à nombre fini s'états est défini par un quadruplet

$$< K, V, Q, \delta >$$

- K est un ensemble fini d'états E0, E1...En où E0 désigne toujours l'état initial
- V désigne un vocabulaire ou alphabet des entrées a0, a1...ak
- Q est un sous-ensemble de K dont les éléments sont appelés états terminaux
- δ est une application, dite fonction de transition qui, à tout couple composé d'un élément q de K et d'un élément a de V, associe un sous-ensemble de K. On écrit r $\in \delta$(q,a) dans le cas où r appartient à ce sous-ensemble. (E0, a0) –> Ek (qui se lit : si l'automate est dans l'état E0 et reçoit l'entrée a0, alors il passe à l'état Ek)

Un automate est dit déterministe, si, à un couple composé d'un élément q de K et d'un élément a de V, la fonction δ associe au plus un élément r de K.
Théorème de Kleene (1956)
Pour tout automate à nombre fini d'états, il existe une expression régulière qui représente

le langage reconnu.
Pour toute expression régulière, il existe un automate à nombre fini d'états qui représente le langage représente.

9.1.1.4 Transducteurs

Un transducteur à nombre fini d'états est défini par un sextuplet

$$< K, Ve, Vs, Q, F, \delta >$$

- K est un ensemble fini d'états
- Ve désigne le vocabulaire ou alphabet d'entrée (destiné à être reconnu)
- Vs désigne le vocabulaire ou alphabet de sortie (ou de réécriture)
- Q est un élément appelé état initial
- F est un sous-ensemble de K dont les éléments sont appelés états terminaux
- δ est une application, dite fonction de transition qui, à tout couple composé d'un élément q de K et d'un élément a de Ve, associe un sous-ensemble de K. On écrit r $\in \delta$(q,a) dans le cas où r appartient à ce sous-ensemble.

Un transducteur permet donc, via le vocabulaire Vs, d'associer une séquence de sortie à une séquence reconnue. On utilise les transducteurs pour annoter le texte (ajouter des informations linguistiques, baliser les séquences pertinentes, etc.).

9.1.2 Unitex et la technologie à nombre fini d'états

9.1.2.1 Alphabet et symboles utilises

1. **Alphabet** Unitex permet à l'utilisateur de définir son propre alphabet pour une langue donnée. Voici un exemple d'alphabet défini dans Unitex, avec la correspondance minuscule majuscule.
   Aa, Bb, Cc, Dd, Ee, Éé, Ëë, Ff, Gg, Hh, Ii, Jj, Kk, Ll, Mm, Nn, Oo, Óó, Öö, Pp, Qq, Rr, Ss, Tt, Uu, Üü, Vv, Ww, Xx, Yy, Zz,
2. **Codes grammaticaux usuels**
   Le Tableau 9.1 montre des exemples de codes grammaticaux usuels.
3. **Codes sémantiques**
   Le Tableau 9.2 montre des exemples de codes sémantiques usuels.

Code	Signification	Exemple
A	adjectif	fabuleux
ADV	adverbe	réellement, à la longue
CONJC	conjonction de coordination	mais
CONJS	conjonction de subordination	puisque, à moins que
DET	déterminant	ses, trente-six
INTJ	interjection	adieu, mille millions de mille sabords
N	nom	prairie, vie sociale
PREP	préposition	sens, à la lumière de
PRO	pronom	tu, elle-même
V	verbe	continuer, copier-coller

Tableau 9.1: Codes grammaticaux usuels

Code	Signification	Exemple
z1	langage courant	blague
z2	langage specialisé	sepulcre
z3	langage tres specialisé	houer
Abst	abstrait	bon goût
Anl	animal	cheval de race
AnlColl	animal collectif	troupeau
Conc	concret	abbaye
ConcColl	concret collectif	décombres
Hum	humain	diplomate
HumColl	humain colectif	vieille garde
t	verb transitif	foudroyer
i	verb intransitif	fraterniser
en	particule pré-verbale (PPV) obligatoir	en imposer
se	verbe probominal	se marier
ne	verbe à negation obligatoire	ne pas cesser de

Tableau 9.2: Codes sémantiques

4. **Codes flexionnelles usuels**

   Le Tableau 9.3 montre des exemples de codes flexionnelles usuels.

Code	Signification
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1,2,3	1, 2, 3 personne
P	présent de l'indicatif
I	impératif de l'indicatif
S	présent du subjonctif
T	impératif du subjonctif
Y	présent de l'impératif
C	présent de conditionnel
J	passé simple
W	infinitif
G	participe présent
K	participe passé
F	futur

Tableau 9.3: Codes flexionnelles usuels

5. **Symboles spéciaux**

   Les exemples des symboles spéciaux sont les suivants :

   – <E> : mot vide, ou epsilon. Reconnaît la séquence vide ;
   – <^> : reconnaît un retour à la ligne ;
   – <$> : séparateur de phrase ;
   – <L> : reconnaît n'importe quelle lettre ;
   – <PNC> : reconnaît les symboles de ponctuation ;
   – <TOKEN> : reconnaît n'importe quelle unité lexicale ;
   – <MOT> : reconnaît n'importe quelle unité lexicale formée de lettres ;
   – <MIN> : reconnaît n'importe unité lexicale formée de lettres minuscules ;
   – <MAJ> : reconnaît n'importe unité lexicale formée de lettres majuscules ;

- <PRE> : reconnaît n'importe unité lexicale formée de lettres et commençant par une majuscule ;
- <DIC> : reconnaît n'importe quel mot composé figurant dans les dictionnaires du texte ;
- <NB> : reconnaît n'importe quelle suite de chiffres contigus ;
- # : interdit la présence de l'espace.

6. **Lemme**

   Avec Unitex il est possible de faire référence à toutes les formes fléchies d'un mot en décrivant le lemme de ce mot entre chevrons. Par exemple dans une grammaire local nous avons une lemme d'un nom en langue polonais <przypadek> et un adjective en langue polonais <pytac>. Nous allons avoir les références à toutes les formes fléchies de ce mot qui se trouvent dans le dictionnaire Table [9.4].

Formes fléchies	Lemme	Codes	Formes fléchies	Lemme	Codes
przypadek	przypadek	N+Gi+Ns+Ca	pytalabym	pytac	V+Ai+Vp+Mc+Ns+P1+Gf
przypadek	przypadek	N+Gi+Ns+Cn	pytalabys	pytac	V+Ai+Vp+Mc+Ns+P2+Gf
przypadkach	przypadek	N+Gi+Np+Cl	pytalaby	pytac	V+Ai+Vp+Mc+Ns+P3+Gf
przypadkami	przypadek	N+Gi+Np+Ci	pytalam	pytac	V+Ai+Vp+Md+Ta+Ns+P1+Gf
przypadka	przypadek	N+Gi+Ns+Cg	pytalas	pytac	V+Ai+Vp+Md+Ta+Ns+P2+Gf
przypadkiem	przypadek	N+Gi+Ns+Ci	pytala	pytac	V+Ai+Vp+Md+Ta+Ns+P3+Gf
przypadki	przypadek	N+Gi+Np+Ca	pytalbym	pytac	V+Ai+Vp+Mc+Ns+P1+Gpai
przypadki	przypadek	N+Gi+Np+Cn	pytalbys	pytac	V+Ai+Vp+Mc+Ns+P2+Gpai
przypadki	przypadek	N+Gi+Np+Cv	pytalbym	pytac	V+Ai+Vp+Mc+Ns+P3+Gpai
przypadkom	przypadek	N+Gi+Np+Cd	pytalem	pytac	V+Ai+Vp+Md+Ta+Ns+P1+Gpai
przypadkowi	przypadek	N+Gi+Ns+Cd	pytales	pytac	V+Ai+Vp+Md+Ta+Ns+P2+Gpai
przypadkôw	przypadek	N+Gi+Np+Cg	pytaloby	pytac	V+Ai+Vp+Mc+Ns+P3+Gn
przypadku	przypadek	N+Gi+Ns+Cg	pytalo	pytac	V+Ai+Vp+Md+Ta+Ns+P3+Gn
przypadku	przypadek	N+Gi+Ns+Cl	pytac	pytac	V+Ai+Vp+Md+Ta+Ns+P3+Gpai
przypadku	przypadek	N+Gi+Ns+Cv	pytalabyscie	pytac	V+Ai+Vp+Mc+Np+P2+Gaifn
			pytalabysmy	pytac	V+Ai+Vp+Mc+Np+P1+Gaifn
			pytalaby	pytac	V+Ai+Vp+Mc+Np+P3+Gaifn
			pytalyscie	pytac	V+Ai+Vp+Md+Ta+Np+P2+Gaifn
			pytalysmy	pytac	V+Ai+Vp+Md+Ta+Np+P1+Gaifn
			pytaly	pytac	V+Ai+Vp+Md+Ta+Np+P3+Gaifn
			pytac	pytac	V+Ai+Vb
			pytacie	pytac	V+Ai+Vp+Md+Tr+Np+P2
			pytajcie	pytac	V+Ai+Vp+Mi+Np+P2
			pytajmy	pytac	V+Ai+Vp+Mi+Np+P1
			pytaja	pytac	V+Ai+Vp+Md+Tr+Np+P3
			pytaj	pytac	V+Ai+Vp+Mi+Ns+P2
			pytalibyscie	pytac	V+Ai+Vp+Mc+Np+P2+Gp
			pytalibysmy	pytac	V+Ai+Vp+Mc+Np+P1+Gp
			pytaliscie	pytac	V+Ai+Vp+Md+Ta+Np+P2+Gp
			pytalismy	pytac	V+Ai+Vp+Md+Ta+Np+P1+Gp
			pytali	pytac	V+Ai+Vp+Md+Ta+Np+P3+Gp
			pytam	pytac	V+Ai+Vp+Md+Tr+Ns+P1
			pytamy	pytac	V+Ai+Vp+Md+Tr+Np+P1
			pyta	pytac	V+Ai+Vp+Md+Tr+Ns+P3
			pytasz	pytac	V+Ai+Vp+Md+Tr+Ns+P2
			pytano	pytac	V+Ai+Vi+Ta

Tableau 9.4: Exemple des références aux formes fléchies

9.1.2.2 Automates, transducteurs et expressions régulières

Le système Unitex permet de traiter des expressions régulières, des automates et des transducteurs. Nous donnons ci-dessous des exemples de chacun de ces formalismes dans Unitex.

1. **Automate**
   Unitex permet de représenter un ensemble d'expressions linguistique sous la forme d'un automate. Dans la représentation proposée, les graphes contiennent les éléments du vocabulaire dans des "boites" (correspondant aux états de l'automate, alors que traditionnellement ces éléments figurent sur les transitions), mais cela ne change en rien les propriétés des objets manipulés [Figure 9.1]

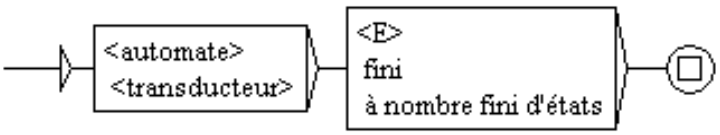


Figure 9.1: Automate 1 - a nombre fini d'états

   Cet automate est strictement équivalent à l'expression régulière donnée précédemment. Unitex permet également de modéliser des automates récursifs (RTN), ou un état correspond en fait à un sous-ensemble appelé dynamiquement. L'appel à un sous-graphe apparaît en grise. Le graphe suivant [Figure 9.2] est équivalent au précédent s'il existe des automates appelé *automate* et *fini* équivalents.

Figure 9.2: Automate 2 - récursif

2. **Transducteur**
   Suivant la définition un transducteur Unitex est un automate auquel est associé un vocabulaire de sortie. Le transducteur suivant permet de reconnaître un certain

nombre de formes et leur associe systématiquement la sortie *automate fini* [Figure 9.3].

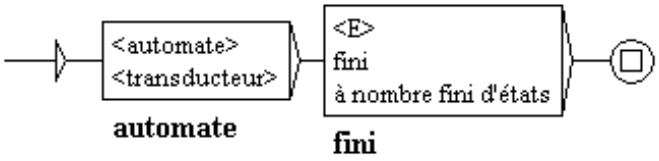


Figure 9.3: Transducteur -

3. **Expression régulière**
   (<automate> + <transducteur>) (<E> + à nombre fini d'états + fini)
   Cette expression permet de reconnaître les séquences suivantes :
   automate ; automates, transducteur ; transducteurs ; automate à nombre fini d'états ; automates à nombre fini d'états, transducteur à nombre fini d'états ; transducteurs à nombre fini d'états ; automate fini ; transducteurs fini ; etc.
   L'étiquette <E> autorise une chaîne vide comme second élément, ce qui permet de reconnaître Automate ou Transducteur de manière isolée.

9.1.2.3 Opération sur les graphes

Les automates peuvent subir l'opération étoile, ainsi que l'union, l'intersection et le calcul du complémentaire. Par exemple : l'union consiste en fait à essayer d'appliquer un automate A à partir des états de l'automate B et l'automate A doit être complètement parcouru à partir d'un état quelconque de l'automate B.

1. **Passage des dictionnaires sur le texte**
   Les dictionnaires Unitex sont compilés sous forme de transducteurs à nombre fini d'états. Le passage d'un dictionnaire sur le texte revient à faire l'union entre le transducteur du texte et le transducteur du dictionnaire. On obtient un graphe où chaque ambiguïté est représentée par une nouvelle boîte dans le graphe résultat.
2. **Passage d'une grammaire sur le texte** Une grammaire étant elle-même un transducteur ou un automate, le passage d'une grammaire sur un texte revient

à calculer l'intersection de la grammaire avec le texte. A l'inverse de l'union de l'union de graphe, l'opération d'intersection telle qu'elle est défini dans Unitex a tendance à enlever des chemins et à regrouper certains éléments.

Abstract

Sentiment analysis - an autonomous system exploring opinions expressed in cinema reviews
Directeur de thèse : Robert Mahl, ENSMP,
Co-encadrement : Katarzyna Wegrzyn-Wolska, ESIGETEL,

This thesis describes the study and development of a system designed for the evaluation of sentiments within cinema reviews. Such a system offers :
– an automatic search of reviews on the Internet,
– the valuation and the attribution of marks to the opinions given by cinema critics,
– the publication of the results.
In order to improve the application results of predictive algorithms the objective of this system is to supply a support system for the prediction engines analysing users profiles. Firstly the system seeks and fetches likely reviews by cinema reviewers on the internet, particularly those who are prolific. Then the system will evaluate and attribute a mark to the opinion expressed in the cinema reviews and automatically associate a numerical mark to each review ; this is the objective of the system. The final stage is to regroup the reviews (as well as the marks) with the user who wrote them so as to create complete profiles and to propose these profiles the prediction engines.

For the development of this system research for this thesis was based principally on the marking of sentiments, this work is in the realm of *Opinion Mining* and *Sentiment Analysis*. Our system uses three different methods for the classification of opinions. We present here two new methods ; one founded on pure linguistic knowledge and the other on a combination of statistic

and linguistic analysis. Subsequently the results are compared using the statistical method based on Bayes' classifier frequently used in this domain.

The ensuing results are then combined in order to make the final evaluation as precise as possible. For this task we used a fourth classifier based on the neuron network.

Between one and five points are attributed to reviews. This mark requires a deeper linguistic analysis than the binary notation- positive/negative which may be objective or subjective and which is habitually used.

This thesis gives a general account of all the system modules which we have created and a detailed analysis of the one dedicated to opinion marking. We wish to show the advantages of deep linguistic analysis which is less commonly used than statistical analysis in the domain of sentiment analysis.

Key Words : Opinion Mining, Sentiment Analysis, Text Classification. Text Categorization. Natural Language Processing, Information Retrieval, Prediction Engine.

Résumé

Cette thèse décrit l'étude et le développement d'un système conçu pour l'évaluation des sentiments des critiques cinématographiques. Un tel système permet:
– la recherche automatique des critiques sur Internet,
– l'évaluation et la notation des opinions des critiques cinématographiques,
– la publication des résultats.
Afin d'améliorer les résultats d'application des algorithmes prédicatifs, l'objectif de ce système est de fournir un système de support pour les moteurs de prédiction analysant les profils des utilisateurs. Premièrement, le système recherche et récupère les probables critiques cinématographiques de l'Internet, en particulier celles exprimées par les commentateurs prolifiques. Par la suite, le système procède à une évaluation et à une notation de l'opinion exprimée dans ces critiques cinématographiques pour automatiquement associer une note numérique à chaque critique ; tel est l'objectif du système. La dernière étape est de regrouper les critiques (ainsi que les notes) avec l'utilisateur qui les a écrites afin de créer des profils complets, et de mettre à disposition ces profils pour les moteurs de prédictions.
Pour le développement de ce système, les travaux de recherche de cette thèse portaient essentiellement sur la notation des sentiments ; ces travaux s'insérant dans les domaines de (ang : Opinion Mining) et d'Analyse des Sentiments. Notre système utilise trois méthodes différentes pour le classement des opinions. Nous présentons deux nouvelles méthodes ; une fondée sur les connaissances linguistiques et une fondée sur la limite de traitement statistique et linguistique. Les résultats obtenus sont ensuite comparés avec la méthode statistique basée sur le classificateur de Bayes, largement utilisée dans le domaine.
Il est nécessaire ensuite de combiner les résultats obtenus, afin de rendre l'évaluation finale aussi précise que possible. Pour cette tâche nous avons utilisé un quatrième classificateur basé sur les réseaux de neurones.
Notre notation des sentiments à savoir la notation des critiques est effectuée sur une échelle de 1 à 5. Cette notation demande une analyse linguistique plus profonde qu'une notation seulement binaire : positive ou négative, éventuellement subjective ou objective, habituellement utilisée.
Cette thèse présente de manière globale tous les modules du système conçu et de manière plus détaillée la partie de notation de l'opinion. En particulier, nous mettrons en évidence les avantages de l'analyse linguistique profonde moins utilisée dans le domaine de l'analyse des sentiments que l'analyse statistique.

Mots clefs : Opinion Mining, Analyse des Sentiments, Classification du Texte, Catégorisation du Texte, Traitement Automatique de la Langue Naturelle (TALN), Information Retrieval, Moteur de Prédiction.