Domaines et fouille d'opinion : une étude des marqueurs multi-polaires au niveau du texte

Morgane Marchand

Université Paris-Sud

Ecole Doctorale 427
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)
Laboratoire Vision et Ingénierie des Contenus (LVIC)

Discipline : Informatique

Thèse de doctorat

Soutenance le 4 mars 2015 par

Morgane Marchand

Domaines et fouille d'opinion
Une étude des marqueurs multi-polaires au niveau du texte

Directrice de thèse :	Mme Anne Vilnat	Professeure (Université Paris Sud)
Co-directeur de thèse :	M. Romaric Besançon	Ingénieur chercheur (CEA, Saclay)
Encadrant CEA :	M. Olivier Mesnard	Ingénieur chercheur (CEA, Saclay)

Composition du jury :

Rapporteurs :	Mme Béatrice Daille	Professeure (Université de Nantes)
	M. Thierry Poibeau	Directeur de recherche (LaTTiCe, Montrouge)
Examinateurs :	Mme Sophie Rosset	Directrice de recherche (LIMSI, Orsay)
	M. Mathieu Valette	Professeur (InaLCO, Paris)

DOMAINES ET FOUILLE D'OPINION

Une étude des marqueurs multi-polaires au niveau du texte

MORGANE MARCHAND

REMERCIEMENTS

J'aimerai tout d'abord remercier les trois personnes à m'avoir accompagnée tout au long de cette thèse, à savoir Anne Vilnat ma directrice de thèse ainsi que mes encadrants Romaric Besançon et Olivier Mesnard. Ma thèse ne serait certainement pas ce qu'elle est aujourd'hui sans leurs conseils avisés.

Je remercie également Béatrice Daille et Thierry Poibeau d'avoir relu mon manuscrit avec beaucoup de précision, ainsi que Sophie Rosset et Mathieu Valette d'avoir accepté de juger mon travail.

J'ai passé trois très bonnes années au sein de mes deux laboratoires, le LVIC et le LIMSI, et j'en attribue le mérite à tous les collègues que j'ai pu y côtoyer. Je garde notamment de très bons souvenirs de toutes les discussions au coin café ainsi que des soirées restaurants ou laser-game.

L'aide entre doctorants est très importante lors d'une thèse et je dois beaucoup à Mathieu qui m'a grandement aidé à mettre en place mon interface d'annotation et en a été le plus grand contributeur. J'ai également souvent bénéficié de l'aide technique et amicale de Quentin avec qui il m'a été très agréable de partager mon bureau durant ces trois ans et plus.

Enfin, merci de tout cœur à ma famille pour son soutien moral, logistique et gastronomique. Merci en particulier à ma grande sœur pour ses conseils et à ma petite sœur pour son enthousiasme.

TABLE DES MATIÈRES

1 INTRODUCTION 2

I PROBLÉMATIQUE 7

2 ETAT DE L'ART 8
- 2.1 La fouille d'opinion ou l'analyse de sentiments 8
  - 2.1.1 La subjectivité dans le langage 8
  - 2.1.2 L'opinion et le sentiment 9
  - 2.1.3 La fouille d'opinion 9
- 2.2 De l'influence du domaine sur l'opinion 10
  - 2.2.1 Les constructions classiques de ressources et classifieurs et leur performance sur plusieurs domaines 11
  - 2.2.2 Prendre en compte le domaine 12
- 2.3 Transfert d'apprentissage et adaptation au domaine 15
  - 2.3.1 Schéma de fonctionnement d'un transfert 15
  - 2.3.2 Classification des différents types de transfert d'apprentissage 15
  - 2.3.3 L'adaptation au domaine dans le TAL 16
- 2.4 Conclusion 18

3 PRÉSENTATION DE LA PROBLÉMATIQUE 19
- 3.1 Choix de la tâche étudiée 19
- 3.2 Problématique de la représentation commune 20
  - 3.2.1 Espace de représentation commun 20
  - 3.2.2 Comment évaluer la transportabilité d'un domaine à un autre ? 22
- 3.3 Les travaux de Blitzer et al. 23
  - 3.3.1 Utilisation du corpus Multi-Domain Sentiment Dataset 23
  - 3.3.2 Implémentation du *Structural Correspondance Learning* 24
- 3.4 Limitations du Structural Correspondance Learning 26
  - 3.4.1 Créer des clusters pour rester interprétable 27
  - 3.4.2 Utilisations possibles des clusters de mots 27
- 3.5 Conclusion 28

4 CRÉATION DE CLUSTERS DE MOTS 29
- 4.1 Examen préliminaire sur des clusters par projection 29
  - 4.1.1 Méthode générale 29
  - 4.1.2 Sélection des mots pivots 30
  - 4.1.3 Création de la matrice de correspondance 31
  - 4.1.4 Regroupement des mots du vocabulaire par Markov Clustering 32
- 4.2 Utilisation des clusters de mots 32

4.2.1 De nouveaux traits d'apprentissage 32
4.2.2 Transfert d'information de polarité 33
4.3 Comment évaluer la qualité d'un cluster 35
4.3.1 Mesure de mixité 35
4.3.2 Mesure de Pureté 36
4.4 Conclusion 41

II LES MARQUEURS MULTI-POLAIRES 42

5 CONCEPT ET DÉTECTION DES MARQUEURS MULTI-POLAIRES 43
5.1 Définition de la notion de marqueurs multi-polaires 43
5.1.1 Liens entre subjectivité, polarité, adjectifs, lexiques et marqueurs multi-polaires 44
5.1.2 Approches complémentaires 45
5.2 Sélection supervisée des marqueurs multi-polaires 45
5.2.1 Description de la méthode de sélection 46
5.2.2 Exemples de marqueurs multi-polaires 49
5.3 Vers une sélection non supervisée 52
5.3.1 Description d'une méthode de sélection semi-supervisée 52
5.3.2 Sélection des mots pivots 54
5.3.3 Résultats et pistes d'amélioration 55
5.4 Conclusion 57
6 MISE EN PLACE D'UNE EXPÉRIENCE D'ANNOTATION MULTI-ANNOTATEURS 59
6.1 Création d'une interface d'annotation 59
6.2 Validation du guide d'annotation 60
6.2.1 Sélection des phrases 60
6.2.2 Mesures d'accord entre deux annotateurs 62
6.2.3 Généralisation et adaptation des mesures d'accords 63
6.2.4 Analyse de l'accord inter-annotateurs global 66
6.2.5 Analyse des accords inter-annotateurs deux à deux 67
6.2.6 Validation des hypothèses et lancement de l'annotation à grande échelle 69
6.3 Expérience multi-annotateurs 70
6.3.1 Présentation des mots choisis 70
6.3.2 Accords inter-annotateurs globaux multi-annotateurs 70
6.3.3 Score de positivité au niveau des phrases 71
6.4 Analyse des positivités des marqueurs au niveau des phrases 71
6.4.1 Différences dans les écarts de positivité 72
6.4.2 Coefficient d'accord S 73
6.5 Conclusion 74
7 CLASSIFICATION DES MARQUEURS MULTI-POLAIRES 75
7.1 Comportement des marqueurs dans un domaine 75

7.1.1 Contexte : parler d'autre chose 76
7.1.2 Opinion : expression d'un jugement 77
7.1.3 Raison : qualité ou défaut 78
7.1.4 Cible : élément de l'objet 79
7.1.5 Possibilité de préférence 79
7.2 Paires de comportements distinguables entre deux domaines 80
7.2.1 Polaire versus neutre : description prédominante 81
7.2.2 Polaire versus neutre : mélange de trois classes 83
7.2.3 Polaire versus neutre : mélange de deux classes 85
7.2.4 Positif versus négatif : présence de préférences 88
7.2.5 Autres comportements distinguables 89
7.3 Catégories de justification de changement de comportement 90
7.3.1 Description contextuelle 91
7.3.2 Changement de sens 91
7.3.3 Changement d'objet 91
7.3.4 Changement d'utilisation 91
7.3.5 Biais de corpus 93
7.4 Conclusion 94

III UTILISATION PRATIQUE DES MARQUEURS MULTI-POLAIRES 96

8 ADAPTATION D'UN DOMAINE À UN AUTRE 97
8.1 Faciliter l'adaptation d'un domaine à un autre 97
8.2 Intégration des marqueurs multi-polaires au transfert d'apprentissage 98
8.2.1 Méthode d'intégration des marqueurs multi-polaires 98
8.2.2 Détection des marqueurs multi-polaires 99
8.2.3 Particularisation des corpus d'apprentissage et test 99
8.3 Apport des marqueurs multi-polaires à l'adaptation au domaine 100
8.3.1 Résultats sur les corpus entiers 100
8.3.2 Discussion des résultats 102
8.4 Conclusion 103

9 CLASSIFICATION D'OPINION SUR UN CORPUS MULTI-DOMAINE 105
9.1 Utiliser plusieurs domaines 105
9.2 Présentation de la méthode 106
9.3 Corpus utilisés 107
9.4 Détection des marqueurs 108
9.5 Corpus d'entraînement particularisés 109
9.6 Présentation des résultats 111
9.7 Conclusion 111

10 CLASSIFICATION D'OPINION SUR UN CORPUS EN DOMAINE OUVERT 113

10.1 Différence avec le cas du multi-domaines 113
10.2 Génération de domaines 114
10.2.1 Séparation des corpus à l'aide de l'allocation de Dirichlet latente 114
10.2.2 Adéquation entre la séparation manuelle et la séparation par LDA 115
10.2.3 Génération de domaines sur un corpus de tweets 119
10.3 Détection des marqueurs multi-polaires et particularisation 120
10.4 Une classification par fusion 120
10.5 Évaluation des résultats et influence des différents paramètres 122
10.5.1 Un comportement semblable au cas des corpus multi-domaines 122
10.5.2 Recentrage des sous-domaines détectés 124
10.5.3 Test des différents mixages 125
10.5.4 pVal et minDiff : description des marqueurs multi-polaires 126
10.5.5 minOcc : pertinence statistique 127
10.6 Participation à la campagne d'évaluation SemEval 2013 128
10.7 Conclusion 130
11 CONCLUSION 132
A EXEMPLE DE CRITIQUES 138
B CRÉATION DE CLUSTERS 147
C MOTS PIVOTS ET COOCCURRENCES 151
D EXEMPLES DE MARQUEURS MULTI-POLAIRES 154
E PHRASES ANNOTÉES LORS DE LA CAMPAGNE D'ÉVALUATION 164

BIBLIOGRAPHIE 172

1

INTRODUCTION

CONTEXTE ACADÉMIQUE DE LA THÈSE

Cette thèse de l'école doctorale d'informatique de Paris-Sud a été effectuée au Laboratoire de Vision et d'Ingénierie des Contenus du CEA-LIST (LVIC) et au Laboratoire d'Informatique pour la Mécanique et les sciences de l'ingénieur du CNRS (LIMSI). Elle a été encadrée par Anne Vilnat du LIMSI et Olivier Mesnard et Romaric Besançon du LVIC.

LE CONTEXTE DE L'ADAPTATION AU DOMAINE POUR LA FOUILLE D'OPINION

Avec le développement du Web 2.0, de plus en plus de personnes donnent leur avis sur internet à propos d'un grand nombre de sujets. Les plate-formes le permettant sont désormais multiples. Il peut s'agir de blogs personnels, de commentaires laissés à la suite d'un article ou encore de critiques de produits via les nombreux sites de vente en ligne. Il existe donc de gros gisements de données potentiellement exploitables. Du fait de leur quantité, des techniques de traitement automatique sont nécessaires afin de transformer ces données en connaissances utiles. Parmi ces techniques, le traitement automatique de l'opinion a de nombreuses applications allant du résumé automatique de critiques de films à la surveillance de l'e-réputation d'une entreprise. Il s'agit aujourd'hui d'un sujet en pleine expansion.

Dans le cadre de la fouille d'opinion, il existe plusieurs tâches de traitement automatique des langues, qu'il est utile ou non de mettre en œuvre selon les applications visées. Il est possible, par exemple, de chercher à détecter la présence ou non d'une opinion ou d'une appréciation, de vouloir classer des opinions exprimées en fonction de leur polarité sur l'axe positif négatif ou bien de leur intensité. D'autres travaux s'attachent à identifier l'objet sur lequel porte l'opinion ou bien la personne qui exprime cette opinion. Toutes ces tâches peuvent être réalisées à différents niveaux selon les applications envisagées : au niveau global du texte, au niveau très précis d'un aspect particulier ou à des niveaux intermédiaires comme la phrase, le paragraphe ou bien une thématique.

Nous nous intéressons dans cette thèse à la classification de l'opinion au niveau du texte. Pour créer un classifieur automatique d'opi-

nion, il est possible de s'appuyer sur des règles définies manuellement, par exemple à partir de lexiques ou bien d'utiliser des approches à base d'apprentissage automatique, avec des modèles qui doivent alors être entraînés sur des corpus qui doivent être annotés. La création des ressources nécessaires à l'entraînement d'un classifieur peut être coûteuse, en temps et en argent. Il serait donc souhaitable de pouvoir utiliser un classifieur déjà créé pour n'importe quel nouveau texte. Or, nous n'exprimons pas notre opinion de la même façon selon ce dont nous parlons. Il existe par exemple des mots qui ne s'emploient que dans un contexte particulier. Ainsi, savoir qu'un vin bouchonné est une mauvaise caractéristique pour un restaurant n'est d'aucune utilité lorsque l'on veut connaître l'avis d'une personne sur le dernier film paru. A l'inverse, si le classifieur est entraîné à partir de critiques de films, il n'a aucun moyen de savoir qu'un vin bouchonné est mauvais. Une autre possibilité est qu'un même mot ne désigne pas la même chose ou bien n'a pas la même connotation selon le thème de la discussion. Pour reprendre l'exemple des films et des restaurants, parler de navets est signe d'opinion négative pour les uns mais est totalement neutre pour les autres. Aussi, l'un des défis majeurs de la fouille d'opinion est l'adaptation d'un domaine à un autre. Le mot domaine désigne ici le thème général du texte, et plus spécifiquement le type d'objets sur lesquels l'opinion est portée.

ADAPTATION POUR LA CLASSIFICATION DE L'OPINION AU NIVEAU DU TEXTE

Une volonté de méthode multilingue

En abordant la question de l'adaptation au domaine pour la classification de l'opinion, nous avons souhaité nous focaliser sur des méthodes les plus indépendantes possible de la langue utilisée. Cela permet ainsi de les utiliser pour un grand nombre de langues, y compris certaines peu dotées. C'est pourquoi nous avons choisi d'étudier la tâche de classification au niveau du texte représenté en sacs de mots d'unigrammes et bigrammes puisque ce type de représentation se transfère très bien d'une langue à l'autre. Le choix de classifieurs statistiques à base de corpus et non à base de règles permet également de ne pas utiliser au cours de l'apprentissage de connaissance propre à la langue. En contrepartie, il est nécessaire de développer des méthodes utilisant le moins possible d'annotations manuelles des corpus.
C'est pourquoi nous nous sommes finalement intéressée au problème de la classification de l'opinion au niveau du texte représenté en sacs de mots d'unigrammes et bigrammes en utilisant des méthodes d'apprentissage statistique.

Projection et possibilité d'interprétation

Lorsque l'on entraîne un classifieur sur un corpus afin de l'utiliser pour classer des textes issus d'un autre corpus, on fait l'hypothèse implicite que la distribution des traits de représentation des textes est similaire dans les deux corpus. Or lorsqu'un classifieur est entraîné sur un domaine pour être utilisé sur un autre, cette hypothèse ne tient plus. Il existe deux types de différences dans les distributions. Certains traits peuvent être existant dans un domaine et inexistant dans l'autre. D'autres peuvent ne pas avoir la même répartition à l'intérieur des classes d'intérêt d'un domaine à l'autre.

Beaucoup de travaux de recherche s'intéressent au premier problème, c'est à dire celui des mots qui sont connus dans un domaine et inconnus dans un autre. En effet, ne pas prendre ces mots en compte résulte en une perte d'information importante. Aussi essaie-t-on de relier ces mots inconnus à des mots connus afin de propager l'information polaire d'un domaine à un autre. Les nouveaux espaces de représentation ainsi formés sont en général plus difficilement interprétables. Cela ne pose aucun problème si les classifieurs sont développés dans le but de servir d'outil de type boîte noire. Cependant, si l'on veut s'en servir afin de créer de nouvelles ressources telles que des lexiques polaires bi-domaines, il est important de conserver le plus possible les liens avec les mots d'origine.

Le problème des marqueurs multi-polaires

Par ailleurs, le transfert entre les domaines ne peut se faire dans de bonnes conditions que si les mots connus des deux domaines, sur lesquels les liens avec les mots inconnus s'appuient, se comportent exactement de la même façon dans les deux domaines. Ce qui nous ramène au second type de différences de distribution. Au cours de cette thèse, nous nous sommes plus particulièrement intéressée à ces mots que nous appelons marqueurs multi-polaires. Plus précisément, nous définissons un marqueur multi-polaire comme une unité linguistique (mot ou bigramme) donnant un indice de polarité au niveau du texte et pour lequel cette polarité change en fonction du domaine. Ces marqueurs peuvent aisément être repérés à l'aide de corpus annotés au niveau du texte dans plusieurs domaines.

Les études classiques du phénomène des mots changeant de polarité selon le domaine se focalisent en général uniquement sur les adjectifs ou bien sur des mots appartenant au préalable à un dictionnaire de

mots polarisés. Or, tous types de mots peuvent être concernés par le phénomène. Il peut s'agir de mots désignant un jugement, une qualité, une description, une réaction, une partie de l'objet dont on est en train de parler. Certains de ces mots sont intrinsèquement polaires, d'autres ne le deviennent que dans un contexte particulier. Il est donc préférable, d'un point de vue pragmatique, de ne pas se restreindre dans le choix des mots étudiés.

PLAN DE LA THÈSE

Au cours de cette thèse, nous nous intéressons à la problématique de l'adaptation au domaine pour la fouille d'opinion. La première partie dresse un panorama des défis liés à cette thématique (chapitre 2). Nous insistons plus précisément sur les techniques de projection des vocabulaires de plusieurs domaines dans un espace commun afin de pouvoir exploiter l'information apportée par les mots connus en gardant une caractérisation explicite de l'opinion qui reste proche des mots (chapitre 3). Cette proximité permet de nombreuses applications. Dans ce contexte expérimental, nous mettons en évidence l'importance de l'impact des marqueurs multi-polaires, comme le mot "navet" dont nous avons parlé plus haut (chapitre 4).

Dans la seconde partie, nous étudions plus en détail ces marqueurs multi-polaires. Nous proposons tout d'abord une méthode de détection simple à partir de corpus annotés et nous proposons également un processus permettant de les repérer sans utiliser d'annotation dans le domaine cible (chapitre 5). Nous avons ensuite mis en place une expérience de collecte d'annotations manuelles afin de déterminer la polarité des phrases contenant des marqueurs multi-polaires. Nous vérifions ainsi que l'information de polarité du texte est bien présente autour du marqueur (chapitre 6). À partir de ces annotations, nous avons étudié les différents comportements des marqueurs à l'intérieur des différents domaines et établi une catégorisation des causes de changement de comportement entre domaines (chapitre 7).

Enfin, dans une troisième partie, nous revenons sur l'impact de la prise en compte des marqueurs multi-polaires pour les classifieurs automatiques de l'opinion au niveau du texte, en abordant trois situations. Nous nous intéressons dans un premier temps à l'adaptation au domaine proprement dite, c'est à dire au cas où un classifieur est entraîné sur un domaine et utilisé sur un autre (chapitre 8). Dans un deuxième temps, nous étudions le cas où un classifieur est entraîné sur des textes provenant de plusieurs domaines identifiés et utilisé pour classer des nouveaux textes issus de ces domaines. Nous regardons ici l'influence des marqueurs entre les domaines à l'intérieur même du corpus d'apprentissage (chapitre 9). Enfin, nous proposons

d'utiliser une séparation automatique en domaines dans le cas de collections en domaine ouvert, où l'on sait que le classifieur est entraîné sur des textes issus de domaines variés mais où l'on ne dispose d'aucune information de domaine (chapitre 10).

Première partie

PROBLÉMATIQUE

Dans cette partie, nous expliquons comment, à partir de la problématique de l'adaptation au domaine pour la fouille d'opinion, nous en sommes venue à nous intéresser au cas des marqueurs multi-polaires. Nous présentons tout d'abord une revue de littérature à propos de la fouille d'opinion et du transfert d'apprentissage. Ensuite, à partir de l'étude de travaux sur la projection de domaines différents dans un espace de représentation commune, nous proposons la création de clusters de mots afin de permettre le transfert d'information polaire entre les mots de deux domaines différents afin de créer de nouvelles ressources. Enfin, l'étude de ces clusters a mis en évidence la présence de mots ou bigrammes présents dans les deux domaines mais présentant une polarité différente selon les cas. Il convient donc de détecter ces mots ou bigrammes, que nous appelons marqueurs multi-polaires afin de permettre la création de clusters plus performants.

2

ETAT DE L'ART

Savoir ce que les autres pensent est, depuis toujours, une information très importante pour prendre une décision. Nous consultons des critiques de consommateurs avant d'acheter un appareil photo, des sondages avant des élections ou encore dans le domaine professionnel des lettres de recommandation. Depuis le développement d'Internet, de plus en plus de personnes rendent leurs avis disponibles. Nous avons donc facilement accès à un très large corpus d'opinions en tout genre. Le domaine de la fouille d'opinion s'est développé pour exploiter ces informations.
Les applications possibles de la fouille d'opinion sont multiples [Pang and Lee, 2007; Liu, 2012]. Elle peut, par exemple, être utilisée pour agréger des critiques, faire des systèmes de recommandation ou bien des outils de marketing et de business intelligence. Certains moteurs de recherche proposent déjà des applications pour résumer les opinions des consommateurs dans des interfaces dédiées au shopping [Blair-Goldensohn et al., 2008]. L'idéal serait de pouvoir disposer de telles fonctionnalités pour des recherches d'ordre général.

La diversité et la quantité de ces témoignages rendent leur traitement manuel long et coûteux. C'est pourquoi l'exploitation automatique de ces données est un enjeu majeur.

2.1 LA FOUILLE D'OPINION OU L'ANALYSE DE SENTIMENTS

2.1.1 *La subjectivité dans le langage*

La terminologie utilisée en fouille d'opinion est multiple : opinion, sentiment, subjectivité, polarité, etc. Nous nous intéressons ici spécifiquement à l'expression de l'opinion, qui peut se classer sur un axe positif/négatif.
On peut distinguer deux niveaux de subjectivité dans le langage [Benveniste, 1966] :

- le premier niveau n'implique pas l'expression d'une évaluation. Il témoigne simplement du degré de présence de l'énonciateur dans son énoncé. Cette présence peut être implicite ou bien explicite en fonction de la présence ou l'absence de certains marqueurs ;
- le second niveau est celui des évaluations exprimées par l'énonciateur. Elles se caractérisent par la présence d'un prédicat ex-

primant l'évaluation. Ce prédicat peut avoir ou non une valeur axiologique (positif, négatif, neutre...)

C'est ce deuxième niveau qui nous intéresse ici. Il est cependant parfois difficile de distinguer les deux niveaux de subjectivité et cela peut amener à des erreurs de classification.

2.1.2 *L'opinion et le sentiment*

L'étude des opinions et sentiments dans les textes est un axe de recherche qui s'est constitué au début des années 2000. Sa relative nouveauté explique que les termes techniques utilisés pour le décrire ne soient pas toujours normés. Les termes plus couramment utilisés sont ceux de "fouille d'opinion" et "d'analyse des sentiments" parfois réunis sous l'appellation de "analyse de la subjectivité", parfois utilisés de manière interchangeable. Le terme "fouille d'opinion" est apparu pour la première fois dans l'article de Dave et al. [2003] et a été plus utilisé par la communauté de recherche d'information. Le terme "d'analyse des sentiments" a quant à lui fait son apparition plus au cœur de la communauté de traitement automatique des langues [Pang and Lee, 2007].
aujourd'hui cependant, la plupart s'accordent à parler d'opinion lorsque l'on souhaite classer un avis sur une axiologie positif/négatif. L'analyse de sentiments concerne quant à elle l'étude des émotions telles que la peur, la colère ou la joie.

2.1.3 *La fouille d'opinion*

La fouille d'opinion se compose de plusieurs tâches, qu'il est utile ou non de mettre en œuvre selon les applications visées. :

- détection de la présence ou non de l'opinion ;
- classification de l'axiologie de l'opinion (positif, négatif, neutre) ;
- classification de l'intensité de l'opinion ;
- identification de l'objet de l'opinion (ce sur quoi porte l'opinion) ;
- identification de la source de l'opinion (qui exprime l'opinion).

Toutes ces tâches peuvent se pratiquer à différents niveaux selon les applications envisagées : au niveau global du texte, au niveau très précis d'un aspect particulier ou à des niveaux intermédiaires comme la phrase, le paragraphe ou bien une thématique.
Afin de réaliser ces analyses d'opinion plusieurs types de ressources existent : des lexiques valués, des ontologies, des règles ou patrons lexicaux et des corpus annotés ou non et ce aux différents niveaux possibles.
Enfin, que l'on utilise des classifieurs à base de règles ou des classifieurs statistiques, il existe une multitude de représentations possiblement utilisables. Cela va du sac de mots ou de n-grammes à l'utilisa-

tion d'arbres de relations syntaxiques, de rôles lexicaux ou de parties du discours.

2.2 DE L'INFLUENCE DU DOMAINE SUR L'OPINION

Selon le sujet d'un texte, ce ne sont pas les mêmes mots de vocabulaire qui sont employés. On pourrait cependant penser que les expressions d'évaluation sont universelles. En effet, certains mots et certaines structures reviennent avec régularité tels que "j'adore" ou bien "je le déconseille". De plus, les dictionnaires notent que certains mots sont péjoratifs ("avare"), d'autres, au contraire, mélioratifs ("généreux"). Ainsi, selon Pup [1998], il y a des mots à valeur intrinsèquement positive ("généreux, délicieux") et d'autres à valeur intrinsèquement négative ("avare, mauvais"). D'autres mots semblent en revanche neutres : "table" est l'exemple classiquement donné par les linguistes. On parle ici d'orientation *a priori*.

Néanmoins, à côté de mots intrinsèquement positifs ou négatifs, il existe des mots dont l'orientation peut changer selon le contexte dans lequel ils sont employés [Riloff and Wiebe, 2003]. Ce sont ces mots ou expressions que nous étudierons à partir du chapitre 5 de cette thèse. Il peut s'agir de mots polysémiques ou bien d'homonymes ayant des axiologies différentes. C'est le cas du "navet" qui est un légume tout à fait ordinaire en cuisine mais un film à éviter dès lors que l'on parle de cinéma. La désambiguïsation sémantique (savoir quel sens est effectivement utilisé) s'appuie justement sur les mots du contexte. Les méthodes existantes utilisent des corpus, annotés ou non, ainsi que des dictionnaires inventoriant les sens existants [Navigli, 2009].
L'orientation d'un mot non polysémique peut également changer à l'intérieur d'un même domaine, selon l'objet qu'il évalue. Par exemple pour un ordinateur portable, une batterie "large" est un inconvénient mais un écran "large" est un atout. L'orientation des mots peut aussi dépendre des préférences et de l'idéologie de l'auteur et c'est alors bien plus difficile à détecter. Les textes politiques sont notamment très sensibles à cela. Par exemple, le mot "bourgeois" est fondé sur une sémantique neutre mais quand il s'agit de préjugé ou d'opinion, ce qui est "bourgeois" est souvent mal vu.

Un problème proche de l'adaptation au domaine est l'adaptation au niveau de langue. On retrouve un vocabulaire différent selon les niveaux mais aussi des mots communs qui changent de polarité ("C'est terrible !, C'est mortel !"). Ces inversions de sens peuvent être extrêmement fortes comme le mot "bad" qui signifie exactement le contraire de son sens littéral dans le domaine du blues à une certaine époque.

Dans la prochaine partie, nous allons voir que les méthodes clas-

siques pour obtenir des lexiques et des classifieurs d'opinions ne sont pas toujours adaptées pour prendre en compte le changement de vocabulaire induit par le changement de domaine.

2.2.1 *Les constructions classiques de ressources et classifieurs et leur performance sur plusieurs domaines*

Cette partie se focalise sur les problèmes d'adaptation des ressources et des classifieurs classiques. Pour obtenir plus de détails sur les méthodes de construction classiques, le lecteur se référera à Pang and Lee [2007].

2.2.1.1 *Les ressources*

Pour la constitution de ressources, on distingue deux grandes familles d'approches. La première consiste à utiliser des dictionnaires. À partir d'un petit ensemble de mots, appelés mots graines, le lexique est étendu en utilisant les relations de synonymie et d'antonymie [Kim and Hovy, 2005; Esuli and Sebastiani, 2006] ou bien les définitions [Andreevskaia and Bergler, 2006]. La seconde consiste à s'appuyer sur un corpus. Le lexique de mots graines est étendu en s'appuyant sur plusieurs indices comme les conjonctions et/mais [Hatzivassiloglou and McKeown, 1997], la co-occurrence entre mots[Turney and Littman, 2002] ou la proximité des contextes d'évaluation [Turney et al., 2003]. Il existe également des approches mixtes, combinant l'utilisation de corpus et de dictionnaires [Taboada et al., 2011] ou bien de patrons d'extraction [Riloff and Wiebe, 2003].

Les lexiques obtenus en utilisant des dictionnaires ne sont pas spécifiques à un domaine mais leur couverture est souvent faible et ils sont pour la plupart limités au sens *a priori* des mots, c'est-à-dire hors contexte. Les méthodes à base de corpus sont quant à elles applicables à tous les corpus, quel que soit leur domaine. Cependant, le lexique finalement appris dépendra du domaine du corpus utilisé. Enfin, les patrons d'extraction sont longs et coûteux à créer. De plus, les résultats nécessitent souvent un nettoyage manuel avant d'être réellement exploitables.

2.2.1.2 *Les classifieurs*

En ce qui concerne la création de classifieurs pour l'axiologie positive/négative, on distingue également deux approches principales. La première consiste à utiliser principalement des lexiques et des indices linguistiques [Takamura et al., 2005; Ferrari, 2009]. La seconde consiste à utiliser des données d'apprentissage afin de construire un classifieur statistique. Le type de classifieur a moins d'importance que les traits utilisés qui peuvent être des n-grammes [Pang et al.,

2002], des arbres de relations syntaxiques [Kudo and Matsumoto, 2004], tous les mots ou bien certains mots particuliers comme les adjectifs et les adverbes [Benamara et al., 2007].

Les classifieurs développés à partir de ressources générales ont plusieurs défauts. En effet ces ressources sont trop générales et ne captent pas la spécificité des domaines. Par exemple, Denecke [2009] teste les scores du lexique général SentiWordNet dans la tâche de classification des opinions sur six corpus différents. Leurs classifieurs statistiques mono-domaines ont de bien meilleurs résultats que les classifieurs à base de règles utilisant uniquement les mots de SentiWordNet. Un autre problème des ressources générales est que certains mots *a priori* positifs ou négatifs peuvent en réalité être employés dans des contextes neutres voire de polarité opposée [Wilson et al., 2009]. Quant aux classifieurs développés sur un domaine particulier, les utiliser directement sur d'autres domaines donne en général de mauvais résultats. Par exemple, dans Aue and Gamon [2005], les auteurs comparent des classifieurs entraînés sur quatre domaines différents. Leurs résultats montrent que l'utilisation d'un classifieur entraîné sur un domaine source différent du domaine cible fait perdre entre 2 et 38 % d'exactitude (*accuracy*).

2.2.2 *Prendre en compte le domaine*

Afin de surmonter les défauts de performance des méthodes classiques, une possibilité est de s'attacher à développer des ressources générales plus performantes. Le but est d'obtenir une performance acceptable sur tous les domaines ou, au moins, sur un grand nombre de domaines. Une autre possibilité est de développer des méthodes permettant, à moindre coût, d'adapter automatiquement une ressource générale à un domaine particulier.

2.2.2.1 *Améliorer les performances des classifieurs généraux*

Comme nous l'avons vu précédemment, les lexiques d'opinion généraux donnent des scores de polarité *a priori*. Or cet *a priori* change selon le contexte et il faudrait disposer de lexiques capables d'en rendre compte.
Tous les auteurs ne retiennent pas la même définition pour le contexte d'un mot d'opinion : cela peut aller de la cible directe de l'opinion [Jijkoun et al., 2010] à un sac de mots représentant le thème abordé [Li and Zong, 2008]. Un lexique donnant des scores différents selon l'étiquette grammaticale du mot, comme SentiWordNet, peut être considéré comme faiblement contextuel [Dang et al., 2010]. On peut également imaginer des lexiques d'opinion généraux bien plus fortement contextuels. Par exemple, Gindl et al. [2010] créent tout d'abord deux lexiques contextuels et valués sur deux corpus A et B. Ils déterminent

ensuite pour quels termes l'ajout du contexte a été utile, nocif ou neutre pour A et B. Les résultats obtenus grâce au lexique contextuel sont ainsi comparés à ceux obtenus grâce au lexique non-contextuel. Ils ne gardent ensuite que les termes contextuels qui sont soit utiles soit neutres sur les deux domaines à la fois, créant ainsi un lexique contextuel qui donne de bon résultats sur plusieurs domaines.
Wilson et al. [2009] ne créent pas un lexique contextuel, mais utilisent les relations déduites d'arbres de dépendances syntaxiques afin de tempérer les informations apportées par les orientations des mots *a priori*.

Une autre carence des lexiques d'opinion généraux classiques est de manquer souvent d'expressions polylexicales. Les mots simples sont les plus faciles à repérer mais ils ne suffisent pas à capter la richesse de l'expression de l'opinion dans la langue. Certaines expressions polylexicales sont même intégralement composées de mots qui ne sont pas eux même évaluatifs, par exemple "un coup de bol" ou bien "une bouffée d'air frais". C'est pourquoi des lexiques exhaustifs sont très difficiles à constituer.
Les travaux de Vernier et al. [2010] utilisent des marqueurs d'intensité (comme "très") pour pallier ce manque. Ils ont en effet observé que ces marqueurs s'appliquaient le plus souvent à des expressions subjectives. Ils utilisent donc des requêtes Yahoo pour sélectionner les candidats qu'ils séparent ensuite entre objectif et subjectif à l'aide d'un SVM. Ils ont évalué manuellement l'efficacité de ce nouveau lexique par rapport à un lexique de base sur un corpus de blog qui mélangeait des textes de domaines différents. Ils observent un gain de 15,6% en précision par rapport au lexique de base pour la détection de fragments subjectifs.

Enfin, pour construire un classifieur statistique le plus général possible, il est utile d'utiliser des données d'apprentissage venant du plus grand nombre de domaines possible afin de pratiquer du co-apprentissage ou bien de la fusion de traits et de classifieurs [Li and Zong, 2008; Li et al., 2011]. Les classifieurs obtenus avec ce type d'approche donnent de bons résultats sur plusieurs domaines si l'on dispose d'un peu de données annotées pour tous. Néanmoins, il est impossible de garantir des résultats pour des domaines complètement nouveaux.

2.2.2.2 *Passage du général au particulier*

Les lexiques d'opinion généraux peuvent être adaptés à un domaine particulier en utilisant les méthodes d'expansion classiques sur un corpus sélectionné pour être thématique. C'est le cas de Harb et al. [2008] qui extraient automatiquement du Web un corpus thématique en utilisant des requêtes du type "+opinion +cinema +good

-bad -poor -nasty ...". Ils extraient ensuite les adjectifs porteurs d'opinion en mesurant la cooccurrence dans les phrases entre les adjectifs candidats et les mots graines du lexique initial.
La méthode des doubles propagations, décrite dans Qiu et al. [2009, 2011], peut être utilisée pour trouver de nouveaux mots d'opinion associés à leur cible sur un corpus particulier. Elle permet à la fois de découvrir les mots d'opinion et leurs cibles grâce à un processus d'amorçage (*bootstrap*). Les travaux se fondent sur la reconnaissance des relations grammaticales reliant les mots d'opinion et leur cible. Ces relations sont décrites au préalable manuellement. Lors de l'expansion, les relations sont détectées à l'aide d'un analyseur en dépendances. Ainsi, à partir d'un lexique d'opinion général, on augmente d'une part les cibles détectées et d'autre part le lexique de mots d'opinion en utilisant les relations une fois dans un sens et une fois dans l'autre.

Une autre manière d'adapter un lexique général à un domaine particulier est non pas de l'étendre mais de le restreindre. C'est ce que font Jijkoun et al. [2010] dans leurs travaux. Ils réalisent une détection de relations syntaxiques afin d'associer à chaque mot du vocabulaire général un certain nombre de candidats pouvant être la cible de l'opinion. Ils font l'hypothèse que les cibles des opinions sont plus diverses que les autres éléments syntaxiquement liés à un terme d'opinion et ne retiennent donc que les mots cibles ayant un fort score d'entropie.

Sans étendre ou restreindre le vocabulaire, il est possible de juste vouloir adapter au domaine le score de polarité des mots contenus dans le lexique général. C'est par exemple le cas dans les travaux de Choi and Cardie [2009]. A l'aide d'une formulation en problème linéaire en nombres entiers, ils exploitent les relations entre les mots d'une même expression et la polarité des expressions afin d'adapter la polarité *a priori* des mots.

Enfin, d'après l'hypothèse distributionnelle de Harris [1954], des mots qui apparaissent dans des contextes semblables ont des significations semblables. Aussi, si l'on dispose de corpus spécifiques à un domaine, il est possible d'apprendre des représentations vectorielles des mots capturant cette similarité de contexte [Mikolov et al., 2013; Erk, 2012; Van de Cruys et al., 2011]. L'utilisation de ces nouvelles représentations spécifiques à un domaine donne en général de bons résultats.

2.3 TRANSFERT D'APPRENTISSAGE ET ADAPTATION AU DOMAINE

2.3.1 *Schéma de fonctionnement d'un transfert*

Jiang and Zhai [2007a] ont analysé d'un point de vue théorique quelques propriétés générales des problèmes d'adaptation au domaine dans les modèles discriminatifs. Soit $\mathcal{X}$ l'espace des instances pour un problème d'apprentissage donné et $\mathcal{Y}$ l'espace des étiquettes possibles pour une instance. Le but du problème est de trouver une estimation de la fonction prédictive $f : \mathcal{X} \to \mathcal{Y}$. Une représentation est une fonction $R : \mathcal{X} \to \mathcal{Z}$ avec $\mathcal{Z}$ un espace de traits adapté. Les dimensions de $\mathcal{Z}$ sont appelées traits de représentation au *features*. Pour une instance $x \in \mathcal{X}$, on parle de traits ou *features* de x pour désigner les valeurs des dimensions particulières de $R(x)$.

Dans un cas complètement supervisé, chaque paire d'entraînement (x, y), rassemblant une instance et son étiquette associée, est tirée à partir d'une distribution jointe $p(x, y)$ avec $x \in \mathcal{X}$ et $y \in \mathcal{Y}$. L'objectif de l'apprentissage discriminatif, auquel nous nous intéressons plus particulièrement ici, est d'estimer les paramètres θ du modèle utilisé afin de pouvoir reconstituer la distribution conditionnelle $p(y|x)$. L'apprentissage génératif a, quant à lui, pour but de reconstruire la distribution jointe $p(x, y)$. Après le processus d'entraînement, étant donné un objet test x_i, le modèle peut prédire son étiquette de sortie y en évaluant $\mathrm{argmax}_y p(y|x; \theta)$.
En pratique, puisque la distribution réelle $p(x, y)$ est inconnue, on utilise en approximation une estimation $\tilde{p}(x, y)$ issue des données d'entraînement. Dans le cas de l'apprentissage supervisé, l'hypothèse de base est que si l'on dispose d'un suffisamment grand nombre de données d'entraînement, alors $\tilde{p}(x, y)$ devient une bonne approximation de $p(x, y)$.

Cependant, dans le cas de l'adaptation au domaine, les données d'entraînement et les données de test sont tirées de deux distributions sous-jacentes différentes ($p_S(x, y) \neq p_T(x, y)$). Par conséquent, l'hypothèse précédente ne tient plus et le problème devient de savoir comment estimer $p_T(x, y)$ ou $p_T(x|y)$.

2.3.2 *Classification des différents types de transfert d'apprentissage*

La notion d'adaptation au domaine est très étroitement liée à celle de transfert d'apprentissage. Le transfert d'apprentissage est un terme générique qui fait référence à un type de problème d'apprentissage qui implique plusieurs tâches ou domaines. Il n'y a pas de consensus à propos de la définition exacte de ces deux termes et ils sont parfois employés de manière interchangeable. Nous présentons ici la

définition du transfert d'apprentissage présentée dans Pan and Yang [2010].

> Étant donné un domaine source D_S et sa tâche d'apprentissage associée T_S, un domaine cible D_T et sa tâche d'apprentissage associée T_T, le transfert d'apprentissage a pour but d'améliorer l'apprentissage de la fonction prédictive cible f_T sur D_T en utilisant la connaissance apprise de D_S et T_S avec $D_S \neq D_T$ et $T_S \neq T_T$.

Si le but est également d'améliorer les performances de la tâche T_S dans le même temps, on parle alors d'apprentissage multi-tâches.
Selon les propriétés de D_S, D_T, T_S et T_T, il y a plusieurs catégories :

- Transfert inductif : T_S est différent de T_T ;
- Transfert transductif : T_S et T_T sont semblables.

Dans chacune de ces catégories, il y a plusieurs cas de figure possibles relatifs à la disponibilité des données :

- Des données étiquetés sont disponibles dans D_S mais pas dans D_T ;
- Il existe des données étiquetés dans les deux domaines mais en très faible quantité pour D_T ;
- Des données étiquetés sont disponibles dans D_T mais pas dans D_S ;
- Aucune donnée annotée n'est disponible.

Les deux premiers cas sont les plus courants. En effet, pour beaucoup d'applications, la collecte de données annotées pour un nouveau domaine est coûteux alors que l'on dispose d'annotations conséquentes pour certains domaines utilisés de longue date. Dans ce cas, si T_T et T_S sont les même tâches, on parle d'adaptation au domaine. Ces techniques permettent de minimiser le besoin d'annotations manuelles coûteuses pour un nouveau domaine tout en conservant une très grande performance.

2.3.3 *L'adaptation au domaine dans le TAL*

Si l'on regarde dans la littérature existante, il existe trois grands types d'algorithme d'adaptation au domaine. Le premier type utilise la transformation de traits de représentation [Daumé, 2007; Blitzer et al., 2006; Jiang and Zhai, 2007b; Guo et al., 2009; Xue et al., 2008]. Dans ces travaux, l'hypothèse est que $p_T(y|x)$ diffère de $p_S(y|x)$ mais qu'il existe une façon de représenter les données avec des traits qui ont des distributions conditionnelles identiques ou similaires dans les deux domaines source et cible ($p_T(y|R(x)) \simeq p_S(y|x)$).
Il y a deux principaux défis dans de tels algorithmes. Tout d'abord, comment distinguer les traits spécifiques aux domaines et les traits généraux. Et deuxièmement, comment trouver un nouvel espace de représentation $R(\mathcal{X})$ pour coder la correspondance entre les domaines source et cible. Pour répondre à ces questions, des algorithmes ont été

proposés tels que le *Structural Correspondance Learning* [Blitzer et al., 2006] ou le *Topic Modelling* [Guo et al., 2009; Xue et al., 2008].

Le deuxième type d'algorithme exploite les *a priori* des modèles pour réduire la différence entre les deux domaines [Finkel and Manning, 2009; Chelba and Acero, 2006; Chan and Ng, 2006]. Lors d'un apprentissage discriminatif, on utilise souvent en *a priori* une distribution gaussienne à des fins de régularisation. Le vecteur de paramètre θ est considéré comme étant tiré à partir d'une distribution *a priori* $p(\theta)$. Afin d'approximer $p_S(y|x)$ à partir d'un grand nombre de données d'entraînement ainsi qu'un petit échantillon de données du domaine cible, on peut ajuster la distribution *a priori* $p(\theta)$ et produire une distribution $p_S(y|x;\theta)$ raisonnable. Cette possibilité a en particulier été étudiée par Finkel and Manning [2009] et Chelba and Acero [2006].
En apprentissage génératif, le terme *a priori* fait référence à une distribution estimée avant le tirage des étiquettes selon $p(y)$. Une hypothèse est que la distribution conditionnelle $p(x|y)$ est semblable ou similaire dans les deux domaines. La disparité entre les distributions *a posteriori* $p(y|x)$ provient en majorité des différences entre $p_S(y)$ et $p_T(y)$. C'est pour cette raison qu'une bonne estimation de $p_T(y)$ à l'aide des jeux de données peut grandement améliorer les performances. Cette possibilité a été étudiée par Chan and Ng [2006] qui ont utilisé des réseaux bayésiens naïfs.

Le troisième type d'algorithme se focalise quant à lui au niveau des instances sans chercher à modifier les modèles d'apprentissage [Axelrod et al., 2011; Xu et al., 2011; Jiang and Zhai, 2007a]. Ce type d'algorithme est étroitement apparenté à des schémas d'apprentissage semi-supervisé classiques comme l'auto-apprentissage. L'idée générale est la suivante : en pondérant ou en sélectionnant les instances d'entraînement, il est possible de rendre $p_S(y|x;\theta)$ proche de $p_T(y|x;\theta)$ avec θ représentant ici les paramètres du schéma de pondération. Il suffit pour cela de donner un faible poids ou bien de retirer les instances qui ont des probabilités $p_S(y|x)$ et $p_T(y|x)$ différentes. En d'autres mots, on suppose que $p(y|x)$ reste semblable entre les deux domaines mais que $p(x)$ varie grandement.Ainsi, pondérer ou sélectionner les instances d'entraînement [Bickel et al., 2007] ou les traits de représentation [Satpal and Sarawagi, 2007] peut contrebalancer cet impact. Ces méthodes s'appliquent également à l'extraction d'information générale [Gupta and Sarawagi, 2009]. Le principal défi pour ce type d'algorithme est de déterminer le schéma de pondération ou de sélection.

2.4 CONCLUSION

La fouille d'opinion est un domaine d'étude vaste et se développant très rapidement pour lequel les applications pratiques sont nombreuses. Le domaine d'un texte, c'est à dire ici le type d'objet dont il parle, influe sur la manière d'exprimer l'opinion. Aussi, l'adaptation au domaine ainsi que la création de ressources spécifiques sont des défis actuels et majeurs de la fouille d'opinion. Dans le chapitre suivant, nous présentons les choix théoriques et pratiques que nous avons effectués en abordant cette question.

3

PRÉSENTATION DE LA PROBLÉMATIQUE

3.1 CHOIX DE LA TÂCHE ÉTUDIÉE

En abordant le thème de l'adaptation au domaine pour la fouille d'opinion, nous avons gardé à l'esprit une volonté de rester assez indépendant vis-à-vis de la langue traitée. Nous avons donc très tôt décidé d'utiliser le moins possible d'annotations manuelles et notamment, lorsque l'on parlera de transfert d'un domaine à un autre, d'annotations sur le domaine cible. En effet, les annotations humaines peuvent être fastidieuses et coûteuses à obtenir. Lorsqu'émerge un nouveau domaine d'intérêt, il serait souhaitable de ne pas avoir besoin d'annoter un nouveau corpus, sauf en très petite quantité, avant de pouvoir le traiter. L'anglais est un cas particulier pour lequel les ressources annotées sont déjà disponibles en grand nombre. Il a par ailleurs été montré, que lorsque l'on a la possibilité de s'entraîner sur beaucoup de domaines, on obtient un classifieur robuste pour un nouveau domaine [Yoshida et al., 2011; Mansour et al., 2013]. Outre le fait, qu'il est toujours intéressant de particulariser quand même le classifieur robuste pour le nouveau domaine, la disponibilité de ressources annotées en grand nombre et pour de multiples domaines reste faible pour presque toutes les autres langues.
Cela justifie notre décision d'investiguer des méthodes très indépendantes de la langue, ou pouvant s'adapter à toute langue autre que l'anglais pour laquelle beaucoup de ressources existent. Dans cette optique, le choix des méthodes statistiques était donc logique. De plus, nous avons choisi un niveau de supervision assez bas avec une annotation globale au niveau du texte et une représentation en sac de mots d'unigrammes et de bigrammes. Cela permet une transposition assez facile à d'autres langues, les langues agglutinantes nécessitant peut-être un pré-traitement supplémentaire pour séparer les unités lexicales. De plus, même si des annotations plus précises d'un point de vue lexical ou syntaxique sont très utiles à la fouille d'opinion et permettent d'obtenir des résultats plus précis, les unigrammes ou les bigrammes restent des traits extrêmement importants, surtout dans le cas d'une classification au niveau du texte [Zhu et al., 2013b,a].

En résumé, notre préoccupation d'une **dépendance minimale à la langue utilisée** et également d'une **utilisation minimale des annotations manuelles**, nous a amenée à nous intéresser au problème de la fouille d'opinion en adoptant les choix suivants :

– Classification au niveau du texte

– Représentation en sacs de mots
– Utilisation de méthodes statistiques

3.2 PROBLÉMATIQUE DE LA REPRÉSENTATION COMMUNE

De précédents travaux ont réalisé les mêmes choix que nous, à savoir étudier l'adaptation au domaine sur le cas de la classification au niveau du texte avec des représentations en sacs de mots en utilisant des méthodes statistiques. Cela a souvent été abordé sous l'angle de la création d'une représentation commune pour les domaines concernés.

3.2.1 *Espace de représentation commun*

Deux domaines différents n'utilisent pas forcément les mêmes mots et pas à la même fréquence. Les mots ont donc des distributions différentes selon les domaines. Or, lorsque l'on entraîne un classifieur sur un corpus pour le tester sur un autre, on fait l'hypothèse implicite que les deux corpus partagent la même distribution de traits. Ce qui n'est bien évidemment pas le cas lorsque l'on utilise deux corpus issus de domaines différents. Il est alors intéressant de créer un espace commun de représentation dans lequel projeter les deux corpus. Dans cet espace, les différences de distribution des traits seraient par définition amoindries.

Afin de créer cet espace commun, beaucoup de travaux essaient de détecter des pivots, des structures communes entre deux domaines. Plusieurs approches ont été proposées. Daumé [2007] a utilisé un noyau heuristique pour augmenter les traits afin de résoudre un problème spécifique de l'adaptation au domaine dans le traitement automatique des langues. Blitzer et al. [2006] ont proposé l'algorithme *Structural Correspondence Learning* (SCL) pour induire des correspondances entre les traits de différents domaines. SCL se fonde sur la recherche de pivots entre les deux domaines permettant de comparer les histogrammes de répartition des différents termes des domaines. Cette approche est motivée par un algorithme d'apprentissage multitâches, ASO (*Alternating Structural Optimization*), proposé par Ando and Zhang [2005]. Cette méthode a été appliquée à la recherche d'opinion dans Blitzer et al. [2007], travaux que nous décrirons un peu plus en détail à la partie 3.3. Les pivots sont ici des mots fréquents utiles à la détermination de l'opinion dans le domaine source annoté. Des classifieurs pivots sont créés qui permettent de comparer les distributions des autres mots par rapport à ces mots pivots. Ce sont les projections de ces distributions qui deviennent les traits représentatifs des textes.
Dans un article plus récent, Blitzer et al. [2011] s'intéressent plus spé-

cifiquement au cas où les supports des domaines source et cible (l'ensemble des mots qui apparaissent dans chaque domaine) ont peu de mots en commun. Les cooccurrences entre les termes des domaines source et cible ne sont donc pas uniquement apprises par rapport à des mots pivots communs aux deux domaines mais également par rapport à des mots spécifiques à un seul domaine.

Un autre travail à ce sujet est celui de Pan et al. [2010] qui développe la méthode de *Spectral Feature Alignment* (SFA). Ils se servent également comme pivots de mots indépendants du domaine, sélectionnés pour leur fréquence dans le domaine cible et leur information mutuelle par rapport aux étiquettes du corpus source. Ils construisent ensuite un graphe bipartite de corrélation entre les traits pivots et les traits non-pivots. Puis à l'aide d'algorithmes de *clustering* spectral, ils créent des *clusters* entre des traits dépendants des domaines source et cible. Ils obtiennent ainsi un espace de représentation commun aux deux domaines. Les résultats obtenus dans Pan et al. [2010] montrent que la méthode SFA obtient de meilleurs résultats en exactitude que d'autres méthodes, dont SCL.
La plupart des méthodes précédentes ne minimisent pas explicitement la distance en terme de distribution entre les deux domaines. von Bünau et al. [2009] ont proposé la méthode *Stationary Subspace Analysis* (SSA) pour faire correspondre les distributions dans un espace latent. Cependant, SSA se focalise sur la détection d'un sous-espace latent sans se préoccuper de la conservation de propriétés dans ce sous-espace telles que la variance des données. Pan et al. [2008] ont proposé une nouvelle méthode de réduction de dimension pour l'adaptation au domaine appelée *Maximum Mean Discrepancy Embedding* (MMDE). MMDE vise à apprendre un espace latent commun sous-jacent aux domaines dans lequel la variance des données peut être préservée. Cependant, cette méthode ne gère pas le problème des mots inconnus. De plus, MMDE apprend l'espace latent en résolvant un programme semi-défini (SDP) qui demande beaucoup de temps de calcul. Une autre méthode de réduction de dimension a été proposée [Pan et al., 2011], appelée *Transfer Component Analysis* (TCA ou SSTCA en version semi-supervisé), qui vise à pallier ces problèmes.

Plusieurs travaux mettent également en lumière que lorsque l'on peut disposer en plus d'une petite partie annotée du corpus cible, cela permet d'améliorer les résultats de manière conséquente [Daumé, 2007; Blitzer et al., 2007; Aue and Gamon, 2005].

3.2.2 *Comment évaluer la transportabilité d'un domaine à un autre ?*

Tous les travaux étudiant la portabilité d'un domaine à un autre font état de domaines plus semblables pour lesquels le transfert se passe mieux [Denecke, 2009; Blitzer et al., 2007; Aue and Gamon, 2005]. La question de savoir comment mesurer la proximité de deux domaines devient donc centrale.

Dans Ben-David et al. [2007], les auteurs développent une borne supérieure pour l'erreur de généralisation d'un classifieur entraîné sur un domaine source et testé sur un domaine cible. Cette borne comprend deux termes variables. Le premier est l'erreur effectuée sur le domaine source. Le second est une mesure de la divergence entre les distributions des domaines source et cible sous une certaine représentation. Selon la représentation choisie pour les textes (unigrammes, bigrammes, rôles sémantiques...), les distributions des traits seront différentes. Par conséquent, la divergence entre les deux domaines dépend de la représentation choisie. En choisissant une représentation très simplifiée, on peut rendre la divergence entre les deux domaines faible. Mais alors, l'erreur effectuée sur le domaine source sera très grande. Il faut donc choisir avec soin la représentation des textes pour obtenir une divergence faible entre les deux domaines tout en conservant une erreur raisonnable sur le domaine source.
Une fois la représentation définie, se pose le problème de calculer la divergence des deux distributions. Une mesure naturelle serait la distance L_1 ou variationnelle. Cependant, cette distance n'est pas calculable à partir d'un corpus fini pour des distributions à valeur réelle. C'est pourquoi Ben-David et al. [2007] utilisent ce qu'ils appellent la A-distance. Il s'agit d'une restriction de la distance variationnelle à une collection A d'ensembles de textes issus des corpus de façon à ce que chaque élément de A soit mesurable sous les deux distributions. On obtient ainsi une borne supérieure calculable pour l'erreur de généralisation du classifieur considéré.

D'un point de vue pratique, calculer la A-distance à l'aide de données réelles est comme entraîner un classifieur pour départager les textes selon qu'ils appartiennent au domaine source ou cible.

La A-distance fonctionne pour une classification de type 0/1. Les travaux de Mansour et al. [2009] introduisent la *discrepancy distance* qui peut également être utilisée pour comparer des distributions dans le cadre d'une tâche de régression.

3.3 LES TRAVAUX DE BLITZER ET AL.

Attardons nous un peu plus sur les travaux présentés dans Blitzer et al. [2007] qui sont, à notre connaissance, les premiers à avoir utilisé l'idée de projection dans un espace commun pour l'adaptation au domaine appliqué à la fouille d'opinion. Cet article décrit une heuristique pour l'adaptation au domaine appelé *Structural Correspondance Learning* (SCL). SCL utilise des données non-étiquetées provenant de deux domaines différents afin de détecter des correspondances de comportement entre des traits spécifiques au domaine source et des traits spécifiques au domaine cible.

3.3.1 *Utilisation du corpus Multi-Domain Sentiment Dataset*

Pour réaliser leur étude, les auteurs ont constitué des corpus thématiques à partir de critiques collectées sur le site internet Amazon. Ils ont utilisé quatre corpus thématiques, *DVDs*, *kitchen*, *electronics* et *books*. Les critiques sont représentées en sacs de mots en utilisant les unigrammes et les bigrammes présents. Grâce au nombre d'étoiles attribuées aux critiques, les auteurs se sont assurés que leurs corpus contiennent autant de critiques positives (quatre et cinq étoiles) que de critiques négatives (une et deux étoiles). Les textes ayant obtenu trois étoiles n'ont pas été pris en compte à cause de leur polarité ambiguë. La description plus complète du corpus *Multi-Domain Sentiment Dataset* (MDSD) se trouve à la section 5.2.1.1.
Nous avons reproduit leurs travaux et nous proposons une étude plus détaillée de l'influence de certains phénomènes et des limitations de l'approche, illustrée sur les deux corpus *DVDs* et *kitchen*. Au final, le corpus source *DVDs* contient 5586 critiques et le corpus cible *kitchen* 7945 critiques également réparties entre négatives et positives. En moyenne, les critiques du corpus *kitchen* contiennent 145 unigrammes et bigrammes, celles de *DVDs*, 269.

Les travaux des auteurs cherchent à se rapprocher de conditions réelles où l'on dispose d'un grand nombre de données non annotées à la fois pour le domaine cible et pour le domaine source, et seulement d'une petite partie de corpus source annoté. Aussi, lors de chaque expérience, on considère que l'on ne connaît les étiquettes que de 2000 critiques du corpus source : 1000 positives et 1000 négatives.
Nous avons implémenté la méthode SCL et avons étudié le sens d'adaptation de *DVDs* vers *kitchen*. Les références que nous utilisons sont les suivantes : un classifieur entraîné et testé sur le domaine source, un classifieur entraîné et testé sur le domaine cible et un classifieur entraîné sur le domaine source et testé sur le domaine cible sans ajouter les traits obtenus par SCL. Nous comparons bien sûr également nos résultats avec ceux des auteurs.

3.3.2 *Implémentation du Structural Correspondance Learning*

L'idée de la méthode SCL est d'établir des correspondances entre des mots du domaine source et des mots du domaine cible en fonction de leur comportement par rapport à des mots pivots communs aux deux domaines. Considérons le mot S qui n'apparaît que dans le corpus source et le mot C qui n'apparaît que dans le corpus cible. Un classifieur usuel entraîné sur le domaine source ne saura pas quoi faire de C. Mais si S et C, chacun dans son corpus, ont les mêmes co-occurrences avec les mots pivots communs de la même façon, on peut supposer que C équivaut à S dans le domaine cible. Le classifieur devra donc traiter C comme si c'était S.

En pratique, la première étape est donc d'identifier quels mots joueront le rôle de pivots. Les auteurs commencent par sélectionner un ensemble de mots ou bigrammes qui apparaissent fréquemment dans les deux domaines. Ces mots sont ensuite classés selon leur information mutuelle par rapport aux classes positive et négative pour les 2000 critiques du corpus source dont on connaît la polarité. L'information mutuelle de deux variables aléatoires est une quantité mesurant la dépendance statistique de ces variables. Pour deux variables X et Y prenant les valeurs x et y, elle se calcule comme suit :

$$\sum_{x,y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$$

Dans notre cas, la variable aléatoire X représente la présence ou l'absence d'un mot ou bigramme dans un texte du domaine source. La variable Y, quant à elle, indique si le texte en question possède une étiquette positive ou négative. Ainsi, dans notre cas, une information mutuelle élevée indique que la présence ou l'absence d'un mot dans texte est fortement liée à la polarité positive ou négative dudit texte.
Seuls les 1000 traits les plus informatifs sont conservés. Ces traits pivots sont donc fréquents dans les deux domaines et relativement utiles à la tâche de classification de l'opinion pour le domaine source (par exemple "a-must", "loved-it", "weak", "awful", etc.)
Les tests effectués ont mis en valeur le fait que le choix des traits pivots influence énormément les performances du classifieur. Les résultats fournis par les auteurs sont des résultats d'exactitude. Nous avons également recalculé la performance en précision pour les deux classes (positif et négatif).
Nous avons donc sélectionné des ensembles de 1000 traits pivots de trois façons différentes :

- sélection uniquement selon l'information mutuelle (MI) par rapport aux étiquettes du domaine source ;
- sélection uniquement selon la fréquence d'apparition dans les domaines source et cible ;

– combinaison des deux critères précédents comme les auteurs de l'article.

Une fois les traits pivots sélectionnés, les auteurs modélisent la corrélation entre tous les traits des deux corpus et les traits pivots en entraînant pour chaque trait pivot un classifieur linéaire appelé classifieur pivot. Ce classifieur, appris sur l'ensemble des corpus source et cible, répond à la question : "Est-ce que le mot pivot considéré à des chances d'apparaître dans ce texte sachant tous les autres mots du texte". Les vecteurs de poids de ces classifieurs pivots sont agrégés en une matrice. Celle-ci est ensuite réduite par décomposition en valeurs singulières. Les auteurs ne conservent que 50 dimensions. Ils obtiennent ainsi une matrice de projection permettant de calculer 50 nouveaux traits (à valeur réelle) pour chaque texte source et cible.
Les textes du corpus source et cible sont représentés par un vecteur contenant à la fois les traits initiaux (les unigrammes et bigrammes) et les nouveaux traits calculés à l'aide de la matrice de projection. C'est sur ces corpus étendus source et cible qu'un classifieur est entraîné et testé. De plus, dans Blitzer et al. [2007], les auteurs normalisent les nouveaux traits afin que leur norme moyenne soit équivalente à celle des anciens traits multipliée par un coefficient α. Ils obtiennent de cette façon de meilleurs résultats. Nous avons nous-mêmes expérimentalement fixé ce seuil α de pondération à 0,5.
La méthode SCL définit un cadre pour permettre l'adaptation d'un domaine à un autre quelle que soit la tâche étudiée, la spécificité de cette dernière étant intégrée à la définition des classifieurs pivots. Pour le cas de la classification d'opinion, les auteurs utilisent comme classifieur final un classifieur linéaire dont les coefficients sont obtenus par descente stochastique de gradient. Nous avons quant à nous utilisé le plus classique classifieur SVM à noyau linéaire qui donne des résultats aussi bons et même parfois meilleurs.

Par rapport à un classifieur entraîné sur un domaine source et testé sur un domaine cible sans rajouter les nouveaux traits, leur approche améliore souvent les performances (10 cas sur 12). En une occasion, ils arrivent même à dépasser les performances d'un classifieur entraîné et testé sur le domaine cible.
Le tableau 1 présente les résultats obtenus avec notre implémentation de la méthode *SCL* pour un classifieur entraîné sur *DVDs* et testé sur *kitchen* ainsi que les références présentées plus haut.

Nous observons quelques différences de résultats entre l'article original et notre implémentation, notamment pour la référence domaine source sur domaine cible. Ces différences s'expliquent par l'utilisation d'un classifieur SVM à noyau linéaire dans notre cas, alors que les auteurs utilisent une descente de gradient stochastique pour déterminer

	Blitzer et al.	Exactitude (*Accuracy*)	Précision classe positive	Précision classe négative
Réf. source->cible	74,0	78,5	79,4	77,6
Réf. source->source	82,4	81,8	80,3	83,4
Réf. cible->cible	87,7	87,7	88,4	87,0
Pivots : fréquence	.	79,8	80,9	78,7
Pivots : MI	.	79,6	85,0	75,7
Pivots : mixte	79,4	79,9	83,9	76,7
Pivots : mixte pond.	81,4	80,7	82,5	79,13

TABLE 1: Résultats pour un classifieur entraîné sur *DVDs* et testé sur *kitchen*

les coefficients de leur classifieur linéaire. Nous observons cependant également une augmentation des résultats grâce à la méthode SCL.

Les pivots sélectionnés uniquement par la fréquence amènent une petite amélioration par rapport à la référence sans toutefois changer l'écart de performance entre la classe positive et la classe négative. Les pivots sélectionnés uniquement par MI, quant à eux, favorisent bien plus la classe positive. En combinant les deux critères de sélection, on arrive à réduire un peu cette différence de performance entre les deux classes, d'autant plus si l'on pondère la contribution des nouveaux et des anciens traits.
On observe donc ici une sensibilité des résultats à la collection des mots pivots. Cette sensibilité est notée par les auteurs et nous la retrouverons également dans la suite de nos travaux (section 5.3.3).

3.4 LIMITATIONS DU STRUCTURAL CORRESPONDANCE LEARNING

Nous avons donc vérifié l'utilité de la matrice de projection créée par la méthode SCL pour la classification des opinions. Cependant, elle peut également réaliser de mauvais alignements. Cela peut notamment arriver lorsqu'un des corpus est plus hétérogène que l'autre. Par exemple le corpus *DVDs*, bien que rassemblant des textes d'un même domaine, fait référence à plusieurs sujets qui sont les sujets des films. Les mots descriptifs des sujets des films ne sont pas informatifs pour notre tâche de classification de l'opinion. Ils apparaissent peu fréquemment en proportion du corpus et risquent d'être mis en corrélation avec des mots du second domaine peu fréquents mais utiles à la classification de l'opinion. Lorsque le classifieur est adapté du domaine hétérogène vers le domaine homogène, il manque donc les informations contenues dans les mots peu fréquents et informatifs du domaine cible. Dans l'autre sens, le classifieur va attribuer des poids à

des mots qui ne sont pas informatifs pour la classification d'opinions. Ceci explique que l'on puisse avoir un transfert important dans un sens et faible dans l'autre.

3.4.1 *Créer des clusters pour rester interprétable*

L'utilisation d'une matrice de projection obtenue par une décomposition en valeurs singulières rend l'interprétation des résultats plus difficile car les traits finaux ne sont plus des unigrammes ou des bigrammes. Nous avons vérifié dans une expérience si l'on pouvait se passer de cette décomposition afin de garder le lien direct avec les mots pivots. Malheureusement, la suppression de cette étape amène des dégradations de 3 % d'exactitude en moyenne par rapport à la référence au lieu d'une amélioration.
Dans cette thèse, nous voulons rester à un niveau interprétable, afin que nos méthodes puissent servir à expliquer la classification des textes et à étendre des ressources lexicales. Il nous faut donc garder des traits liés aux mots de façon directe. Pour ce faire, nous proposons de projeter dans l'espace commun non pas les textes entiers mais simplement les mots. Il y a alors plusieurs façon de donner une information polaire aux mots projetés. Par exemple, il est possible d'utiliser l'hyperplan séparateur du classifieur afin de donner un score d'opinion aux termes cibles qui serait la distance à cet hyperplan séparateur. Il est en effet vraisemblable que les mots réellement polarisés auront une grande distance à cet hyperplan. Une autre possibilité, en suivant les travaux de Pan et al. [2010], est de regrouper en clusters, ou classes de mots, les mots projetés. Nous avons retenu cette voie car les clusters ainsi créés ont beaucoup d'applications potentielles que nous évoquons plus en détail au paragraphe suivant.

3.4.2 *Utilisations possibles des clusters de mots*

Une fois créés, les clusters de mots peuvent ensuite être utilisés comme traits d'apprentissage pour les classifieurs automatiques d'opinion. Comme les mots du domaine source utilisés lors de leur création sont les mots les plus utiles pour la tâche de classification de l'opinion, on peut espérer qu'une grosse majorité de l'information d'opinion soit contenue dans ces clusters. Aussi, il est plausible qu'une représentation des textes en sacs de clusters donne des résultats aussi bons, voire meilleurs, qu'une représentation en sacs de mots et ce, avec beaucoup moins de traits. Ce nombre de traits réduits peut être un avantage lors du traitement d'un grand nombre de données, en particulier si la chaîne de traitement choisie utilise de lourds calculs matriciels.

De plus, la création de clusters de mots a possiblement de nombreux

avantages grâce à la conservation explicite des mots. Cette conservation permet l'interprétation des clusters et d'envisager la création de ressources qui pourront être utilisées de manière indépendante dans d'autres applications, par exemple de la fouille d'opinion utilisant des analyses plus linguistiques. En effet, les clusters de mots rassemblent *a priori* des mots au comportement semblable vis-à-vis de la tâche de fouille d'opinion au niveau du texte. Il pourrait être tout à fait intéressant de voir si cela est le signe d'un comportement semblable d'un point de vue morpho-syntaxique.
Enfin, l'application certainement la plus importante est le transfert de l'information polaire du domaine source au domaine cible et la création d'un dictionnaire de polarité propre au domaine cible. Nous avons vu à la partie 2.2.1.1 l'importance de tels dictionnaires. Or, si un cluster contient des mots du domaine source que l'on sait être positifs dans le domaine source, les mots du domaine cible présents dans ce même cluster ayant le même comportement, on pourra en déduire qu'ils sont eux-même positifs dans le domaine cible. Et inversement pour les termes négatifs. Ces clusters de mots permettent donc la création de dictionnaires polaires adaptés à un nouveau domaine.

3.5 CONCLUSION

Les travaux sur la projection dans un espace commun visent à réduire la distance entre deux domaines en terme distributionnel. Cela permet notamment de pallier le problème des mots du domaine cible qui sont inconnus dans le domaine source. Cependant, à partir du moment où l'on garde le lien direct avec les mots, on voit apparaître un autre problème : des mots, ou groupes de mots, qui n'ont pas la même polarité dans le domaine source ou le domaine cible. En effet, si de tels mots existent, ils vont perturber la création des clusters et surtout propager des erreurs lors de leur interprétation. Nous avons plus précisément mis en évidence ce problème dans le chapitre suivant (4) lors d'une courte étude de quelques clusters. La suite de la thèse sera ensuite consacrée à l'étude plus précise de ce phénomène.

4

CRÉATION DE CLUSTERS DE MOTS

4.1 EXAMEN PRÉLIMINAIRE SUR DES CLUSTERS PAR PROJECTION

Comme nous l'avons vu au chapitre précédent, créer des clusters de mots regroupant des mots polaires appartenant à deux domaines différents peut être très utile pour l'adaptation au domaine pour la fouille d'opinion. Dans ce chapitre, nous présentons une analyse que nous avons effectuée sur l'obtention de tels clusters de mots. Après avoir présenté la méthode utilisée pour créer ces clusters, nous décrivons les mesures mises au points afin d'évaluer leurs intérêts. L'examen de ces mesures nous amène ensuite à mettre en lumière la nécessité de s'intéresser à la détection des mots ou groupes de mots qui indiquent une certaine polarité dans un domaine particulier et une autre polarité dans un domaine différent. Il s'agit de ce que nous appelons les *marqueurs multi-polaires*, dont l'étude est au centre de cette thèse.

4.1.1 *Méthode générale*

Afin de créer ces clusters, nous nous inspirons des travaux de Blitzer et al. [2007], présentés précédemment à la partie 3.3, et utilisons également des mots pivots communs aux deux domaines et utiles à la classification de l'opinion. Comme vu précédemment, l'idée est d'établir des correspondances entre des mots du domaine source et des mots du domaine cible en fonction de leur comportement par rapport à des mots pivots communs aux deux domaines.
Pour chaque mot des corpus, nous calculons un vecteur de correspondance avec les mots pivots qui, une fois agrégés forment une matrice de correspondance (section 4.1.3). Comme décrit précédemment, Blitzer et al. [2007] utilisent ensuite cette matrice de correspondance afin de projeter à la fois les textes du domaine source et les textes du domaine cible dans un espace commun. Ils entraînent ensuite directement leur classifieur dans cet espace abstrait commun.
A ce stade, nous nous différencions de la méthode *Structural Correspondence Learning*. En effet, nous utilisons la matrice de projection pour projeter directement les mots dans l'espace commun et non plus les textes. Ces mots peuvent ensuite être regroupés en clusters dans cet espace commun (section 4.1.4). Comme discuté dans la partie 3.4.2, ces clusters peuvent servir de traits pour des classifieurs ou bien propager des informations du domaine source vers le domaine cible pour la création de lexiques thématiques d'opinion.

4.1.2 *Sélection des mots pivots*

Pour cette expérience, nous avons utilisé les mêmes domaines que précédemment, *DVDs* et *kitchen* issus du *Multi-Domain Sentiment Dataset*. *DVDs* comprend en tout 5586 critiques et *kitchen* 7945. Pour chaque domaine nous nous astreignons à utiliser les annotations de seulement 2000 critiques (1000 positives et 1000 négatives). Les autres critiques seront utilisées pour établir des statistiques de co-occurrence.

Nous avons vu au chapitre précédent que le choix des mots pivots avait une grande importance sur la nature des résultats. Afin de pouvoir se comparer aux travaux de Blitzer et al. [2007], nous avons, tout comme eux, constitué des collections de 1000 mots pivots. Nous commençons par sélectionner des mots apparaissant dans les deux domaines. Pour ce faire, nous fixons un seuil minimal du nombre d'occurrences dans chacun des corpus. Par exemple, si le seuil est de 5, un mot ne sera considéré comme candidat pivot que s'il apparaît au moins cinq fois dans le corpus source et cinq fois également dans le corpus cible. Ce seuil est en réalité fixé automatiquement afin d'être le plus élevé possible tout en gardant au moins 1000 candidats pivots parmi les mots les plus fréquents. De plus, si la proportion d'apparition d'un mot est plus que deux fois plus grande dans un des deux corpus, ce mot n'est pas pris en compte. En effet, un trop fort déséquilibre en terme de fréquences participe à la distance entre les deux domaines telle que décrite par Ben-David et al. [2007, 2010] (voir section 3.2.2). Ce mot n'a vraisemblablement pas la même signification dans les deux corpus, ou tout du moins pas la même utilisation, aussi ne peut-il pas jouer le rôle de pont entre les deux corpus.

Ces mots pivots doivent à la fois être indépendants du domaine et utiles à la tâche de classification de l'opinion. Afin d'assurer ces caractéristiques, nous utilisons des mesures d'information mutuelle, comme introduit à la section 3.3.2. Ainsi, nos mots pivots candidats doivent avoir une information mutuelle faible par rapport à l'appartenance d'un texte à un domaine particulier. A l'inverse, ils doivent avoir une information mutuelle forte quant à la polarité des textes annotés du domaine source. Nous appliquons un seuil de sélection sur l'une de ces mesures d'information mutuelle qui permette de conserver au moins 1000 candidats pivots. Puis les mots ou bigrammes retenus sont classés en fonction de la seconde information mutuelle et seuls les 1000 premiers candidats sont conservés. Ainsi, selon qu'on utilise d'abord l'une ou l'autre des informations mutuelles, nous ne sélectionnons pas, au final, exactement la même liste de mots pivots.

4.1.3 *Création de la matrice de correspondance*

Après avoir sélectionné les mots pivots afin qu'il soient communs aux deux domaines et utiles à la tâche de classification de l'opinion, il faut ensuite choisir les autres mots des vocabulaires source et cible qui seront reliés via les mots pivots dans une matrice de correspondance avant d'être répartis en clusters.
En toute rigueur, il faudrait créer cette matrice de correspondance pour l'intégralité du vocabulaire des deux corpus. En pratique, prendre en compte tout le vocabulaire pose des contraintes matérielles. C'est pourquoi, pour cette expérience, nous avons utilisé une sélection des mots les plus utiles pour la classification de l'opinion grâce à un calcul d'information mutuelle effectué sur la partie des corpus dont on utilise l'annotation, que ce soit pour source et cible. Nous avons ainsi créé plusieurs sous-ensembles de mots de différentes tailles : 100 mots, 1865 mots ou bien 4703 mots (ou bigrammes).
Si l'on veut par la suite pousser plus loin cette expérimentation sur les clusters, il sera possible de sélectionner les mots à projeter sans utiliser d'annotation dans le domaine cible. En effet, comme le but de cette projection est d'importer le savoir du domaine source sur les mots inconnus du domaine cible, il est nécessaire de ne prendre en compte que les mots du domaine source qui ont une claire orientation en terme de polarité. Et pour la même raison, seuls les mots du domaine cible qui sont inconnus ou très peu présents dans le domaine cible seraient alors à prendre en compte.

Après avoir déterminé la partie du vocabulaire que nous souhaitons projeter, il faut à présent créer une matrice de lien entre ces mots, issus des corpus source et cible, et les mots pivots préalablement sélectionnés. Nous avons testé deux manières de calculer cette matrice que nous appelons matrice de correspondance.
Une première méthode consiste à simplement utiliser les co-occurrences au niveau du texte entre les mots des deux corpus et les mots pivots. Cette approche est celle adoptée par Pan et al. [2010]. Nous effectuons ensuite une réduction de la matrice en calculant sa décomposition en valeurs singulières.
Nous avons également testé une autre matrice de correspondance inspirée des travaux de Blitzer et al. [2007] avec une agrégation de classifieurs pivots linéaires mais les résultats obtenus étaient moins bons. Les résultats présentés ci-dessous sont donc ceux obtenus avec la matrice de co-occurrences ayant subie une SVD.

4.1.4 *Regroupement des mots du vocabulaire par Markov Clustering*

La matrice de correspondance construite à l'étape précédente, nous permet d'avoir des représentations des mots de vocabulaire source et cible dans un espace commun à plusieurs dimensions. Blitzer et al. [2007] utilisaient leur matrice de correspondance afin de projeter des textes entiers. Ici, nous restons au niveau des mots.
Ensuite, nous utilisons un algorithme de clustering pour regrouper les mots en clusters dans le nouvel espace. Pour cette étude, nous avons utilisé un algorithme classique, le Markov Clustering, permettant de ne pas fixer *a priori* le nombre de clusters.
Pour utiliser le Markov Clustering, il est nécessaire de calculer la similarité entre tous les mots deux à deux. Nous utilisons pour cela une mesure cosinus classique en ne conservant que les similarités au dessus d'un certain seuil k (testé de 0.1 à 0.9 avec un pas de 0.2). Le paramètre k influence la taille des clusters obtenus qui elle-même a une incidence sur la qualité de ces clusters. Avec une haute valeur de k, les clusters sont plus petits et plus nombreux. Plus la valeur de k diminue, plus les clusters obtenus contiennent de mots mais en contrepartie, ils sont moins nombreux comme le montre la figure 1.

Nous verrons à la section 4.3, que les petits clusters ont plus tendance à regrouper des mots de même polarité que les gros clusters. Par contre, ces derniers contiennent plus souvent à la fois des mots du domaine source et des mots du domaine cible. Ces deux qualités sont nécessaires pour transférer efficacement l'information du domaine source au domaine cible. C'est pourquoi les clusters de taille moyenne se révèlent les plus intéressants. Les résultats présentés plus bas sont obtenus avec le seuil k réglé de façon à obtenir le plus grand nombre de clusters de taille moyenne. Le tableau 2 présente les trois clusters de moyenne taille obtenus avec le vocabulaire de 100 mots ainsi que le cluster de plus grande taille. Les clusters obtenus par projection du plus grand vocabulaire sont présentés en annexe B.

4.2 UTILISATION DES CLUSTERS DE MOTS

4.2.1 *De nouveaux traits d'apprentissage*

Comme l'ont fait Blitzer et al. [2007], il est également intéressant de voir si les clusters ainsi créés peuvent servir de nouveaux traits d'apprentissage pour les classifieurs.
Nous avons entraîné deux classifieurs à séparer les critiques en positif et négatif. Le premier classifieur est entraîné sur les critiques représentées en sacs de mots d'unigrammes et bigrammes normalisés en nombre d'occurrences. Le second classifieur utilise une nouvelle représentation où chaque unigramme ou bigramme est remplacé par

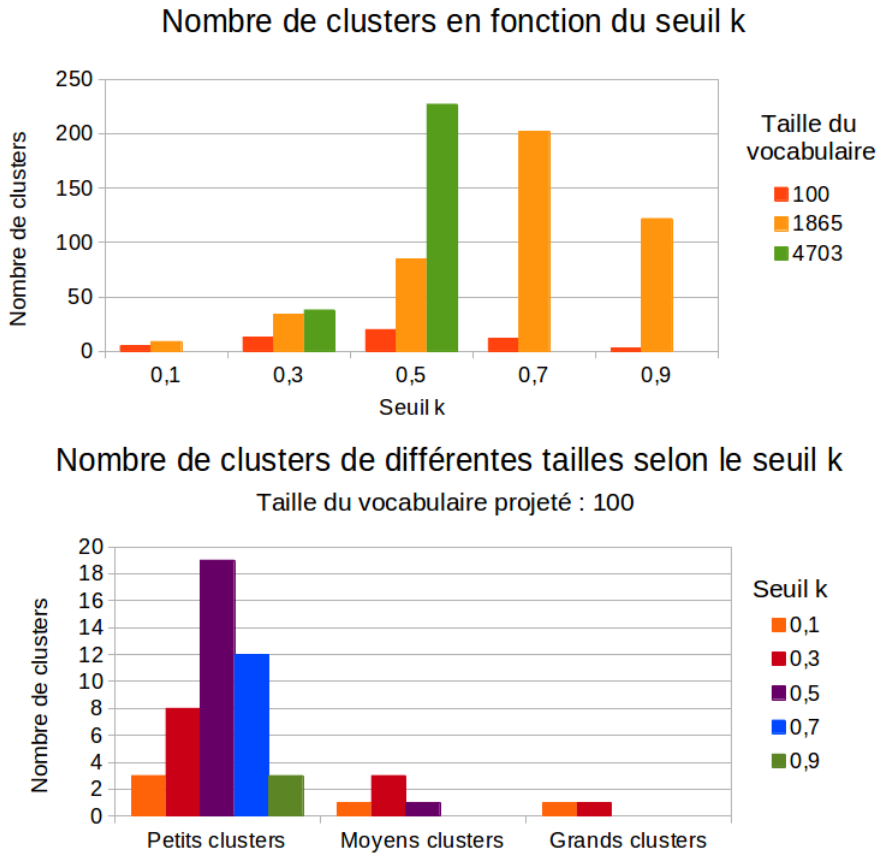


FIGURE 1: Haut : Évolution du nombre de clusters en fonction du seuil k. Bas : Répartition de la taille des clusters pour différentes valeurs de k pour un lexique de 100 mots. Petits clusters : de 2 à 4 mots ; moyens clusters : de 5 à 19 mots ; grands clusters : 20 mots ou plus.

l'identifiant du cluster auquel il appartient. Les unigrammes ou bigrammes qui n'appartiennent à aucun cluster ne sont pas utilisés.

Les résultats présentés dans le tableau 3, montrent que la représentation en clusters n'atteint pas les performances de la représentation en sacs de mots. Par contre, l'exactitude (*accuracy*) de la nouvelle représentation augmente avec la taille du vocabulaire projeté. De plus, elle n'utilise en définitive que très peu de traits (202 clusters au lieu de 288191 unigrammes et bigrammes). Ainsi, les clusters créés par projection semblent utiles au regard de la tâche de classification de l'opinion.

4.2.2 *Transfert d'information de polarité*

Comme discuté à la section 3.4.2, l'application certainement la plus importante est le transfert de l'information polaire du domaine source au domaine cible et la création d'un dictionnaire de polarité propre au domaine cible. Pour pouvoir transférer de l'information du do-

trying-to trying waste-your don't-waste waste-of a-waste waste total-waste money money-on unless unless-you your-money save-our save my-money disappointment a-disappointment like-a more-like looks-like garbage time-and horrible reviews yourself-a not did-not away stay-away not-worth terrible didn't fails dull unfortunately of-time instead joke not-even barely worse huge they're fact-that don't disappointed only no was
minutes supposed-to supposed i-turned turned-it it-off turned to-turn went went-to hours might our not-good poorly maybe below enough-to
best the-best highly highly-recommend i-highly excellent an-excellent not-recommend would-not to-your add-to wonderful poor within-the awesome rating disappointing terrific i-find
love i-love love-this fabulous fantastic you-will wooden

TABLE 2: Clusters obtenus avec le vocabulaire de 100 mots

Taille de vocabulaire projeté	100	1865
Représentation en mots	78.29	78.29
Représentation en clusters	69.83	72.55

TABLE 3: Scores d'exactitude obtenus sur une tâche de classification de l'opinion pour différentes tailles de vocabulaire projeté.

maine source au domaine cible sans propager de biais, les clusters doivent bien sûr contenir des mots des deux domaines et des mots d'une même polarité. Dans la partie suivante, nous nous intéressons à l'évaluation des clusters afin de mesurer leur intérêt pour cette tâche de transfert.

Il est ici intéressant de remarquer que puisque les textes des corpus sont représentés en bigrammes et unigrammes, certaines expressions plus longues se retrouvent découpées. Afin de les reconstruire de manière automatique, nous proposons de reconstruire à partir des bigrammes appartenant à un même cluster les plus grandes suites possibles. Ensuite, nous vérifions si cette suite de mots est bel et bien présente dans le corpus. Voici les suites de mots reconstituées pour l'expérience à petite échelle :

- *don't waste your money on*
- *total waste of time*
- *save your money*
- *i highly recommend*
- *would not recommend*
- *i love this*
- *i turned it off*

– *went to turn*
– *enough to turn*

Les deux dernières suites reconstruites ne semblent pas correctes mais les sept autres sont bien des expressions d'opinion qui mériteraient d'être dans un dictionnaire évaluatif. Certaines sont même assez complexes et *i turned it off* ne s'emploie que dans certains domaines.

Ainsi, l'utilisation de nos clusters d'unigrammes et bigrammes permet de transférer de l'information polaire du domaine source au domaine cible non seulement pour des unigrammes et bigrammes mais également pour des expressions multi-mots plus complexes, qui font souvent défaut aux lexiques d'opinion, certaines étant de plus bien spécifiques à un domaine donné.

4.3 COMMENT ÉVALUER LA QUALITÉ D'UN CLUSTER

Un cluster optimal ne contiendrait que des mots d'une même polarité mais mélangeant des mots du domaine source et des mots du domaine cible. Ainsi il serait aisé de transférer la connaissance du domaine source sur le domaine cible. Afin d'évaluer ces qualités de nos clusters, nous avons défini deux mesures propres : la mixité et la pureté.

4.3.1 *Mesure de mixité*

4.3.1.1 *Définition de la mixité*

La mixité mesure si un cluster contient des mots à la fois du domaine source et du domaine cible. Nous souhaitons en effet créer des ponts entre les vocabulaires des deux domaines. Nous avons donc mis en place la formule suivante afin d'évaluer la mixité d'un cluster particulier :

$$M_{cluster} = \frac{(S + C) - |S - C|}{S + C}$$

avec S le nombre de mots ou bigrammes du cluster appartenant au domaine source et C le nombre de mots ou bigrammes du cluster appartenant au domaine cible. Ainsi, si tous les mots appartiennent à un même domaine, le cluster a une mixité de 0. A l'inverse, une mixité de 1 indique une parfaite répartition (50 % - 50 %) entre les deux domaines.

La difficulté réside dans le fait d'attribuer un mot à un domaine particulier. S'il existe bien évidemment des mots qui n'apparaissent que dans un seul des deux domaines, beaucoup de mots sont en réalité mixtes. Le nombre de ces mots mixtes va diminuer suite à la détection des marqueurs multi-polaires.

Nous avons fait le choix d'attribuer un domaine à un mot de manière nette. Si le pourcentage d'apparition d'un mot dans le corpus source dépasse un certain seuil, il est considéré comme faisant partie du domaine source, et inversement. Les mots dont les pourcentages d'apparition dans les deux domaines ne dépassent pas le seuil, c'est-à-dire qu'ils apparaissent de manière équitable dans les deux domaines, ne sont pas pris en compte.
Il est également possible de fixer le seuil de sélection à 50 %, auquel cas, un domaine est attribué à chaque mot.

4.3.1.2 *La mixité des clusters*

Comme le montre le tableau 4, les petits clusters sont très peu mixtes et ne peuvent donc pas être utilisés pour la construction d'un vocabulaire bi-domaine. Heureusement, la mixité est meilleure pour les autres clusters, surtout ceux de taille moyenne.

M_{global}	M_{petit}	M_{moyen}	M_{grand}
14,9 %	6,3 %	35,6 %	22,2 %

TABLE 4: Mixité de clusters de différentes tailles pour une taille de vocabulaire projeté de 100.

4.3.2 *Mesure de Pureté*

La pureté mesure la propension d'un cluster à ne contenir que des mots de la même polarité. Cette mesure s'appuie donc sur une définition précise de la polarité. De façon pratique, la polarité d'un mot peut être mesurée de différentes façons. Nous présentons tout d'abord deux façons simples d'étudier la polarité. La première est manuelle et la seconde utilise les scores d'un dictionnaire polaire. Ces deux façons de faire donnent des résultats intéressants et faciles à interpréter mais sont difficilement applicables à grande échelle. Aussi, nous présentons plus bas une définition de la polarité des mots et bigrammes évaluée directement à partir des corpus. Dans tous les cas, un score de polarité pour un mot w_i, noté $pol(wi)$, variera de -1 pour les mots strictement négatifs à +1 pour les mots strictement positifs.

4.3.2.1 *Polarité évaluée manuellement*

Pour le premier type d'évaluation de la polarité, nous avons classé manuellement les mots ou bigrammes des clusters en cinq catégories : positif, probablement positif selon le contexte, non défini, probablement négatif selon le contexte, négatif.

Le graphique 2 montre le résultat de cette évaluation sur les trois clusters de moyenne taille et le cluster de grande taille détectés sur le vocabulaire de 100 mots. Nous observons que pour les clusters

de taille moyenne, la pureté est bien respectée. Ces clusters peuvent donc être utilisés pour transmettre la polarité du domaine source au domaine cible. Par contre, on voit que le grand cluster, qui est principalement négatif, contient également des mots très fortement positifs. Il ne peut donc pas être utilisé directement pour le transfert d'information.

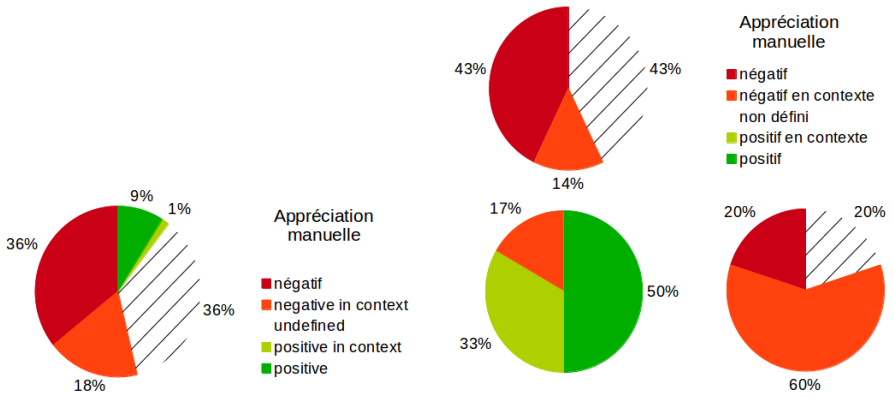


FIGURE 2: Évaluation manuelle de la polarité des mots constituant le grand cluster (gauche) et les clusters de taille moyenne (droite) obtenus lors de la projection du vocabulaire de 100 mots.

Afin d'obtenir un score de polarité pour chaque mot compris entre -1 et 1, nous pouvons attribuer des scores selon nos jugements. Un mot jugé très négatif aurait un score de -1 et un mot négatif selon le contexte un score de -0,5. Pour les mots jugés positifs, ces scores seraient de +1 et +0,5. Les mots neutres ou pour lesquels une estimation de la polarité n'est pas possible auraient un score de 0.

L'inconvénient de cette méthode est naturellement qu'elle n'est pas praticable à grande échelle.

4.3.2.2 *Polarité évaluée à l'aide de lexiques*

Une autre possibilité d'évaluer la polarité des mots d'un cluster est d'utiliser des lexiques polaires qui donnent un score de polarité à un sens d'un mot. Nous avons ici utilisé le lexique SentiWordNet [Esuli and Sebastiani, 2006; Baccianella et al., 2010]. Afin d'obtenir une mesure de polarité allant de -1 à +1, nous avons considéré la moyenne des scores SentiWordNet pour les mots apparaissant dans le lexique. Comme nous le voyons à la figure 3, les résultats vont dans le même sens que l'évaluation de polarité manuelle : les clusters de taille moyenne sont purs alors que le cluster de grande taille ne l'est pas.

Cette façon de calculer la polarité est automatisable et donc utilisable à grande échelle mais pose deux problèmes majeurs. Tout d'abord, il faudrait au préalable effectuer une désambiguïsation de sens afin de sélectionner le score adéquat. L'utilisation comme ici d'un score moyen risque d'introduire de mauvaises estimations de

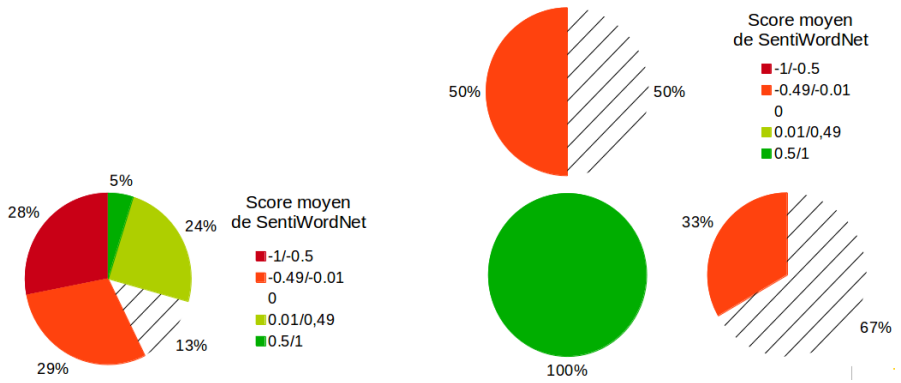


FIGURE 3: Évaluation à l'aide de SentiWordNet de la polarité des mots constituant le grand cluster (gauche) et les clusters de taille moyenne (droite) obtenus lors de la projection du vocabulaire de 100 mots.

la polarité réelle. Ensuite, beaucoup de mots de nos clusters ne se retrouvent pas dans de tels lexiques, à commencer par les bigrammes. De plus, comme nous l'expliquerons plus en détail dans la partie 5.1, nous nous intéressons à ce que nous appelons marqueurs de polarité au niveau du texte, qui ne sont pas nécessairement porteur de polarité en eux-même et donc absents des lexiques pré-établis. Ainsi, puisque cette méthode ne donne pas de score de polarité à tous les objets d'un cluster, elle n'est pas adaptée pour le calcul global de la pureté d'un cluster.

4.3.2.3 *Polarité évaluée à partir des corpus*

En utilisant les définitions de la pureté présentées plus haut, nous avons montré que les clusters de taille petite et moyenne obtenus en projetant une petite partie du vocabulaire sont bel et bien purs. Ils seraient donc tout à fait utilisables pour transférer de l'information polaire du domaine source au domaine cible. Cependant, l'évaluation manuelle à grande échelle des clusters n'est pas possible et l'évaluation à l'aide de lexiques pose notamment le problème des mots inconnus. Aussi, nous avons opté pour une définition de la polarité d'un mot ou bi-gramme qui peut se calculer automatiquement à l'aide des corpus annotés d'où ils sont extraits.

La polarité d'un mot varie de -1 à 1 et représente son pourcentage d'apparition dans les critiques positives et les critiques négatives. Par exemple, si un mot apparaît à égalité dans des critiques positives et des critiques négatives, il aura un score de polarité de 0. La polarité du mot w_i se calcule donc ainsi :

$$\mathrm{pol}(w_i) = \frac{P - N}{P + N}$$

avec P le nombre d'apparitions du mot w_i dans des critiques étiquetées positives et N le nombre d'apparitions dans des critiques étiquetées négatives.

4.3.2.4 *La pureté des clusters*

Une fois la polarité d'un mot ou bigramme définie, nous pouvons en venir à la mesure de la pureté d'un cluster, qui mesure la tendance des mots de ce cluster à être de la même polarité.

La constitution des clusters nécessite la présence de deux corpus, l'un source et l'autre cible annotés en positif ou négatif au niveau du texte. Dans des conditions réelles d'utilisation, seul le corpus source est annoté de cette manière, le corpus cible ne comportant aucune annotation. Pour que notre mesure de pureté soit utilisable en conditions réelles, il faut que la pureté calculée d'un cluster soit à peu près la même selon qu'on utilise dans le calcul à la fois les annotations des corpus source et cible, ce qui donne la polarité réelle du cluster, ou bien seulement celles du corpus source, puisque nous ne disposons pas d'annotation sur le corpus cible en situation réelle. Nous avons testé plusieurs formulations pour le calcul de la polarité avant d'en choisir une qui réponde à ce critère. La pureté globale du cluster est ainsi calculée en ôtant l'écart type des valeurs de polarité des mots qui le composent de 1. Ce qui donne la formule ci-dessous pour un cluster c :

$$Pur(c) = 1 - \sigma_{pol}$$

soit :

$$Pur(c) = \sqrt{\sum_{w_i \in c} \left(pol(w_i - \frac{1}{|c|} \sum_{w_j \in c} pol(w_j) \right)^2}$$

Nous avons donc utilisé cette mesure de pureté afin de comparer la pureté des clusters calculée avec à la fois les annotations des domaines source et cible (polarité réelle) et celle calculée uniquement avec les annotations du domaine source (utilisation réelle).

Comme le montre la figure 4, les points sont globalement sur la médiatrice, ce qui indique que notre définition de la pureté est efficace pour estimer la pureté d'un cluster uniquement à partir des annotations sources.
On observe trois type d'erreurs :

- L'estimation est faussement de 0 :
  aucun mot du cluster n'apparaît dans le corpus source
- L'estimation est faussement de 1 :
  un seul mot du cluster apparaît dans le corpus source

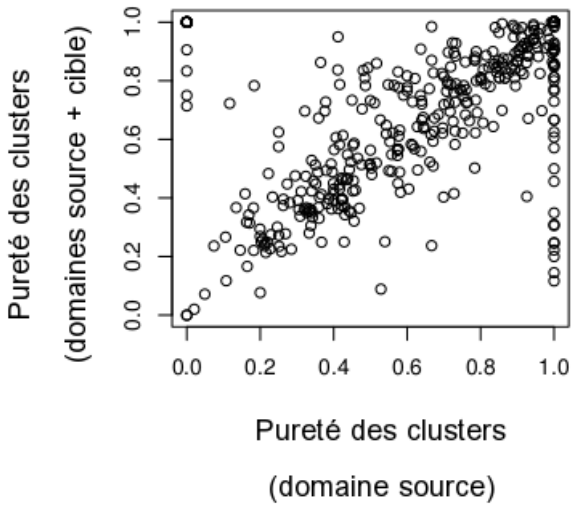


FIGURE 4: Comparaison de la pureté calculée des clusters en utilisant à la fois les annotations des corpus source et cible ou bien seulement celles du corpus source (cas réel).

– Écart à la médiatrice :
  le cluster contient des mots qui changent de polarité selon le domaine

Les deux premiers types d'erreurs ne nous intéressent pas car pour réaliser un transfert d'information de source à cible, il faut que le cluster contienne des informations sur le domaine source, c'est à dire plusieurs mots du domaine source. Ces clusters ne seront donc de toutes façons pas utilisés pour le transfert d'information.
Le dernier type d'erreur met en évidence le problème de ce que nous appelons les marqueurs multi-polaires. Il s'agit de mots ou bigrammes existant dans les deux domaines mais étant dans chaque cas le marqueur d'une polarité différente au niveau du texte. Ainsi, le calcul de la polarité d'un de ces marqueurs ne donnera pas le même résultats selon qu'on utilise les annotations des deux domaines ou de seulement l'un des deux domaines. Par conséquent, le calcul de la pureté d'un cluster contenant des marqueurs multi-polaires sera faussé. Afin de pallier ce problème, les marqueurs multi-polaires devraient être considérés comme deux mots différents selon qu'ils apparaissent dans le corpus source et le corpus cible lors de la projection du vocabulaire et de la création des clusters. Le chapitre 5 présente des façons de détecter ces marqueurs. Cette détection devrait être réalisée au préalable à la création de clusters, afin de s'assurer que la mesure de pureté soit non biaisée et que les clusters ainsi sélectionnés puissent bien servir au transfert de polarité d'un domaine à un autre.

4.4 CONCLUSION

Nous avons étudié la possibilité d'utiliser des mots pivots afin, non pas de projeter les textes issus de domaines différents dans un espace commun, mais de regrouper en clusters des mots du domaine source et des mots du domaine cible ayant le même comportement vis à vis de la tâche de classification de l'opinion au niveau du texte. Notre analyse utilisant un vocabulaire réduit a montré que cet objectif était tout à fait atteignable. Parmi les clusters obtenus, ceux de taille moyenne se révèlent les plus intéressants. Nos résultats suggèrent également que de tels clusters peuvent être utiles en tant que traits d'apprentissage pour les classifieurs ou bien servir à reconstruire des expressions polaires multi-mots.

Néanmoins, à plusieurs moments de notre analyse, nous avons mis en lumière le problème de ce que nous appelons *marqueurs multi-polaires*. Il s'agit de mots ou groupes de mots qui indiquent une certaine polarité dans un domaine donné mais pas dans un autre. Une présentation plus complète de ces marqueurs est faite au chapitre 5.
Tout d'abord, les pivots sont choisis en se référant aux annotations du domaine source. Or, si un mot pivot est négatif dans le domaine source mais positif dans le domaine cible, cela va entraîner des erreurs de détection. En effet, l'hypothèse fondamentale est que les pivots ont le même comportement à la fois dans le corpus source et dans le corpus cible. Ensuite, ces marqueurs multi-polaires vont perturber l'évaluation des clusters. En effet, un mot qui se comporte différemment selon les domaines ne devrait pas être considéré comme commun au deux domaines. Il devrait au contraire être différencié en deux mots distincts. Les marqueurs multi-polaires vont donc influencer le calcul de mixité, tout comme, nous l'avons vu, celui de pureté.
La suite de cette thèse sera consacrée à l'étude de ce phénomène, d'abord d'un point de vue théorique aux chapitres 5, 6 et 7 puis d'un point de vue pratique aux chapitres 8, 9 et 10.

Deuxième partie

LES MARQUEURS MULTI-POLAIRES

Dans cette partie, nous explorons le concept de marqueur multi-polaire. Nous appelons marqueurs multi-polaires tout mot ou bi-gramme qui apparaît dans plusieurs domaines, qui est un indice de polarité dans au moins un domaine, et dont la valeur de polarité est différente selon le domaine.

5

CONCEPT ET DÉTECTION DES MARQUEURS MULTI-POLAIRES

5.1 DÉFINITION DE LA NOTION DE MARQUEURS MULTI-POLAIRES

Pour les classifieurs d'opinion statistiques, la représentation du texte en sacs de mots est très souvent une bonne base qui, suivant la tâche, peut être difficile à dépasser. Il est d'ailleurs rare que les sacs de mots ne fassent pas partie des traits utilisés. C'est pourquoi, nous avons choisi de nous intéresser aux représentations en sacs de mots et donc aux mots eux-mêmes.
Cependant, la polarité d'un mot ou d'une expression peut varier en fonction du contexte. Depuis quelques années, l'intérêt pour lever l'ambiguïté sur la polarité des mots ambigus s'est amplifié [Wu and Jin, 2010]. Presque tous les schémas d'annotation existant pour la polarité permettent de noter cette ambiguïté [Su and Markert, 2008; Wilson et al., 2005]. Plusieurs travaux notent en effet qu'apparaître dans des domaines différents entraîne parfois des changements de polarité pour certains mots. Dans leur travail sur la polarité contextuelle, Wilson et al. [2005] incluent le sujet et le domaine comme indicateurs possibles de variation de polarité. De plus, Su and Markert [2008] remarquent dans leur étude que des préférences de polarité existent selon le domaine ou le sujet du texte. Leur corpus contient 32,5 % de mots à la polarité ambiguë. La simple désambiguïsation de sens ne parvient pas à résoudre complètement cette ambiguïté ce qui montre bien qu'il s'agit d'un phénomène plus global.

Comme nous étudions la classification de l'opinion au niveau des textes entiers, nous arrivons à la définition suivante :

MARQUEUR MULTI-POLAIRE Est appelé "marqueur multi-polaire" un mot ou une expression qui, de manière récurrente dans un domaine particulier, est un indicateur de l'opinion de l'auteur sur l'objet général du texte. De plus, cette indication de polarité varie en fonction du domaine.

Ainsi, un marqueur multi-polaire n'a pas nécessairement de polarité intrinsèque. Cela peut être un mot ou un groupe de mots intrinsèquement objectif qui apparaît néanmoins régulièrement dans des contextes polaires pour un domaine particulier.
Par exemple, savoir qu'un restaurant sert du pâté en croûte ne donne pas d'indication sur la qualité de la nourriture, par contre, dire d'un tableau qu'il est une véritable croûte est mauvais signe pour l'exposition qui le présente. Ce phénomène de marqueurs multi-polaires

entre deux domaines peut également se retrouver entre des sous-domaines plus proches comme les films comiques et les films dramatiques. En effet, "être mort de rire" est une excellente chose pour les premiers mais pas pour les seconds.

5.1.1 *Liens entre subjectivité, polarité, adjectifs, lexiques et marqueurs multi-polaires*

Notre définition des marqueurs multi-polaires ne s'étend pas seulement aux seuls mots ou groupes de mots subjectifs.
Les expressions subjectives sont des mots ou des groupes de mots utilisés pour exprimer des états mentaux comme la spéculation, l'évaluation, le sentiment ou la conviction [Wiebe et al., 2005; Wiebe and Mihalcea, 2006; Wilson, 2008; Akkaya et al., 2009]. Ils sont appelés "état privés", c'est à dire que ce sont des états internes qui ne peuvent pas être directement observés par les autres [Quirk and Crystal, 1985]. La polarité d'un mot ou d'un sens particulier d'un mot, au contraire, fait référence à l'opinion positive ou négative qu'a un agent sur un objet particulier. Ces deux notions ne sont bien sûr pas indépendantes et la plupart des sens subjectifs des mots ont une polarité claire. Néanmoins, une expression polarisée peut également apparaître dans un contexte neutre [Wilson et al., 2009]. De plus, une polarité peut être associée à des mots ou des sens de mots objectifs. Su and Markert [2008] donnent l'exemple du mot *tuberculose* : ce mot ne décrit pas un état privé, on peut le vérifier de manière objective et sa présence dans une phrase ne force pas cette dernière à être porteuse d'opinion. Mais pour la plupart des gens, ce mot porte tout de même une forte connotation négative. De même, nous ne considérons pas que le fait d'être polaire soit réservé aux mots ou expressions ayant été au préalable classés comme subjectifs.

Notre définition des marqueurs multi-polaires se rapproche des travaux sur les concepts de polarité contextuelle ou ciblée [Wilson et al., 2005, 2009; Fahrni and Klenner, 2008]. Néanmoins, ces travaux ont presque toujours un *a priori* sur la forme que doivent prendre les mots étudiés. Fahrni and Klenner [2008] se focalisent sur la détermination de la polarité ciblée des adjectifs. Un nom spécifique à un domaine est souvent modifié par un adjectif qualificatif. D'après les auteurs, les adjectifs n'ont pas de polarité *a priori* mais une polarité ciblée. Dans certains cas, un même adjectif peut changer de polarité en fonction du nom qu'il accompagne. Les auteurs utilisent Wikipédia pour la détection automatique des mots qui peuvent potentiellement être la cible d'une opinion pour un domaine donné. Une méthode de *bootstrap* est ensuite utilisée afin de déterminer la polarité ciblée des adjectifs associés à ces mots. Ils obtiennent de bons résultats mais s'intéressent uniquement aux adjectifs. Wilson et al. [2005], quant à eux,

ne se restreignent pas aux adjectifs mais travaillent uniquement sur des segments de texte contenant des mots prédéterminés (des mots d'un lexique ayant au moins un sens subjectif). Ils se placent au niveau du segment et déterminent d'abord si une expression est neutre ou polaire avant de désambiguïser la polarité de ces expressions polaires en utilisant des règles manuelles et des traits structurels. Leur lexique couvre 75 % des segments polaires de leur corpus. Avec ce type d'approche, si on veut augmenter les performances, il est nécessaire d'augmenter le lexique.
À l'inverse de ces travaux, nous ne souhaitons pas faire d'*a priori* sur les mots ou groupes de mots concernés par le phénomène des marqueurs multi-polaires. C'est pourquoi nous avons donc choisi de les sélectionner automatiquement et de les classifier en une seule étape.

5.1.2 *Approches complémentaires*

Lorsqu'il s'agit de traiter de l'adaptation au domaine pour la fouille d'opinion, beaucoup de travaux utilisent un lexique pré-existant donnant la polarité *a priori* des mots. Ces lexiques sont améliorés, par exemple en pondérant les différentes polarités possibles d'un mot en fonction du domaine [Choi and Cardie, 2009]. Ces lexiques particuliers peuvent alors être utilisés dans des classifieurs à base de règles pour classer la polarité des textes entiers [Ding et al., 2008].
Les études utilisant des classifieurs à base de corpus s'intéressent, quant à elles, principalement à la représentation des données [Glorot et al., 2011; Huang and Yates, 2012]. L'erreur d'adaptation d'un classifieur dépend en effet de sa performance sur le domaine source ainsi que de la distance entre les distributions des mots dans les domaines source et cible [Ben-David et al., 2007]. Avec une bonne projection, un lien peut être établi entre les mots du domaine cible qui n'existent pas dans le domaine source et les autres mots [Pan et al., 2010; Blitzer et al., 2007]. Cependant, si un mot a une polarité différente dans le domaine source et le domaine cible, cela va introduire une erreur d'adaptation. Ainsi, la détection des marqueurs multi-polaires est complémentaire à ces approches et leurs améliorations respectives peuvent être combinées.

5.2 SÉLECTION SUPERVISÉE DES MARQUEURS MULTI-POLAIRES

Pour permettre la détection des marqueurs multi-polaires, nous faisons l'hypothèse que nous disposons de corpus de textes à la fois dans le domaine source et dans le domaine cible. Afin d'étudier au mieux l'influence des mots multi-polaires sur la classification de l'opinion, nous supposons que chaque corpus contient une sous-partie an-

notée. Ainsi, nous pourrons mettre en évidence l'apport de la prise en compte des marqueurs multi-polaires pour la fouille d'opinion multi-domaines en utilisant les marqueurs multi-polaires détectés de manière supervisée. Nos expériences sur la détection des marqueurs multi-polaires en n'utilisant aucune annotation cible sont présentées à la partie 5.3.

5.2.1 *Description de la méthode de sélection*

Pour chaque paire de corpus thématiques, nous utilisons les critiques étiquetées. Pour chaque mot candidat, nous regardons si sa distribution dans les critiques positives et négatives est statistiquement différente selon les domaines. Certains mots peuvent changer de polarité à l'intérieur du même domaine mais, pour détecter un marqueur multi-polaire tel que défini plus haut, nous ne nous intéressons qu'à la polarité au niveau global d'un domaine thématique.
Disposant d'un corpus thématique annoté en positif/négatif au niveau du texte, il est possible d'obtenir un score de positivité indiquant sa propension à être employé dans des critiques positives et par extension, sa polarité. Ce score de positivité est défini comme suit :

$$\text{positivité}_{mot} = \frac{\text{Nombre de critiques positives contenant ce mot}}{\text{Nombre total de critiques contenant ce mot}}$$

Un score de 1 (resp. 0) signifie que dans ce domaine, le mot n'apparaît que dans des critiques positives (resp. négatives). Un écart de 0.5 est donc très significatif, faisant passer un mot de neutre à fortement polarisé.
Ce score dépend du corpus sur lequel il est calculé. Par définition, pour un marqueur multi-polaire, la positivité calculée sur le corpus source va être significativement différente de la positivité calculée sur le corpus cible.
Un mot ou un bigramme est considéré comme un candidat possible pour être un marqueur multi-polaire s'il apparaît au moins 5 fois dans le corpus source et cinq fois dans le corpus cible. Nous comparons ensuite le nombre d'apparitions d'un mot candidat dans les critiques positives et négatives du domaine source avec son nombre d'apparitions dans les critiques positives et négatives du domaine cible. Nous effectuons cette comparaison à l'aide d'un test du χ^2 avec un risque de première espèce (i.e. risque de faux positif) de 1 %. Le test du χ^2 est un test statistique permettant notamment de tester l'homogénéité de deux variables aléatoires. Il s'agit ici de se demander si deux listes de nombres de même effectif total N peuvent dériver de la même loi de probabilité. Pour chaque domaine, nous disposons du nombre de critiques positives contenant un mot candidat, du nombre de critiques positives ne le contenant pas, du nombre de critiques négatives le contenant et du nombre de critiques négatives ne le contenant pas.

Le test indique si ces deux distributions peuvent être des réalisations de variables aléatoires possédant la même loi. Lorsque le corpus est équilibré, le test d'homogénéité peux être appliqué directement. Dans le cas contraire, il est nécessaire d'utiliser la formulation du test du χ^2 en test d'indépendance.
Dans tous les cas, si les distributions du mot dans les corpus source et cible sont significativement différentes, il est considéré comme marqueur multi-polaire.

5.2.1.1 *Description des corpus utilisés*

Nous avons réalisé la détection de marqueurs multipolaires pour l'anglais et le français. Pour l'anglais, nous avons utilisé les corpus *Multi-Domain Sentiment Dataset* (MDSD), collectés par [Blitzer et al., 2007] et présentés dans le tableau 5. Il s'agit de quatre corpus thématiques (*DVDs*, *kitchen*, *electronics* et *books*) contenant des critiques collectées sur le site internet Amazon. Des exemples des critiques se trouvant dans ces corpus sont présentés à l'annexe A.
Chacun des corpus thématiques contient 1000 critiques positives et 1000 critiques négatives que nous utilisons pour la détection des marqueurs multi-polaires. Ces corpus contiennent également un certain nombre de critiques supplémentaires qui seront utilisées pour le test des expériences présentées dans les parties suivantes (de 3586 à 5945 selon le corpus). Les textes sont représentés en sacs de mots de bigrammes et unigrammes des formes fléchies des mots pleins. Leurs nombres d'occurrences sont pondérés par la taille du texte.

Domaines		Books	DVDs	Electronics	Kitchen
Nombre de textes positifs		1000	1000	1000	1000
Nombre de textes négatifs		1000	1000	1000	1000
Nombre de textes test		4465	3586	5681	5945
Taille vocabulaire (uni/bigrammes)		448 713	282 108	232 892	204 580
Taille moyenne des textes		261	269	174	145
Taille moyenne des textes (tokens)		228	235	152	129

TABLE 5: Présentation du corpus anglais MDSD. Les tailles des vocabulaires et des textes sont exprimées en nombre d'unigrammes et bigrammes.

Pour le français, nous avons utilisé les corpus *JeuxVideo* et *AvoirAlire* issus du Défi Fouille de Textes 2007 (DEFT) [Grouin et al., 2007]. Ces corpus contiennent des critiques issues des sites *avoir-alire.com* et *jeuxvideo.com*. Elles sont réparties en trois classes, positif, neutre et négatif mais nous ne considérons ici que les classes positif et négatif. Comme le corpus *AvoirAlire* contient des critiques de différents domaines (films, musiques, livres, pièces de théâtre...), une séparation manuelle selon ces sous-domaines a été effectuée. Les critiques de ces corpus sont majoritairement étiquetées positif ou neutre.

Polarité du texte	BDs	Musique	Films	Livres	Théâtre	Inclassables
Négatif	23	15	427	19	29	3
Neutre	118	101	490	200	98	19
Positif	246	225	706	538	162	41

TABLE 6: Séparation manuelle des textes du corpus DEFT en différents domaines.

Comme nous pouvont le constater au tableau 6, seule la sous-partie *films* contient suffisamment de critiques négatives pour représenter un corpus d'apprentissage équilibré. Pour la détection des marqueurs multi-polaires, nous avons donc utilisé des corpus constitués de critiques sélectionnées au hasard dans la sous-partie *films* de *AvoirAlire* ainsi que dans *JeuxVideo* afin de constituer deux corpus thématiques équilibrés, présentés dans le tableau 7. Chacun contient 420 critiques positives et 420 critiques négatives. Le reste des critiques est utilisé pour le test lors des expériences des parties suivantes (293 textes pour *films*, 1446 pour *jeux vidéo*). Comme pour l'anglais, les textes sont représentés en sacs de mots pondérés de bigrammes et unigrammes des formes fléchies.

Domaines		Jeux video	Films
Nombre de textes positifs		420	420
Nombre de textes négatifs		420	420
Nombre de textes test		1446	293
Taille vocabulaire (uni/bigrammes)		630 503	236 137
Taille moyenne des textes		2226	757
Taille moyenne des textes (tokens)		1482	578

TABLE 7: Présentation du corpus français issu de DEFT. Les tailles des vocabulaires et des textes sont exprimées en nombre d'unigrammes et bigrammes.

Les corpus anglais et français sur lesquels nous avons travaillé sont tous les deux des corpus de critiques. Pour le corpus anglais, il s'agit d'avis clients laissés par des personnes ayant acheté le produit dont on parle. A l'opposé, les critiques du corpus français sont écrites par des critiques de métier. Ainsi, les critiques en anglais sont en général plus courtes, utilisent un vocabulaire plus courant tout en contenant potentiellement plus de néologismes ou de fautes de grammaire et orthographe que les critiques en français.

5.2.2 *Exemples de marqueurs multi-polaires*

Le tableau 8 présente quelques marqueurs détectés comme changeant de polarité entre deux domaines dans le corpus anglais MDSD. Rappelons que pour chaque domaine, un mot a un score de positivité qui correspond à son nombre d'occurrences dans des critiques positives par rapport à son nombre total d'occurrences dans le domaine. Un score de 0 indique un mot complètement négatif et un score de 1 un mot complètement positif. On remarque que les mots détectés sont variés, ce sont aussi bien des noms, des adjectifs ou des verbes, conjugués ou non et ils se comportent différemment. Une liste plus exhaustive des marqueurs détectés est présentée à l'annexe D.
En analysant les mots détectés, nous avons établi une classification empirique des différents comportements des marqueurs multi-polaires, présentée au chapitre 7.

	region	*I loved*	*worry*	*compare*	*return*
Domaine *electronics*	0.154	0.091	0.929	0.846	0.055
Domaine *books*	0.818	0.735	0.3	0.263	0.633

TABLE 8: Pourcentage de présence de cinq exemples de marqueurs dans les critiques positives pour deux domaines.

Nous avons ainsi en moyenne détecté 400 marqueurs multi-polaires sur l'anglais et 1000 sur le français. Ce décalage est vraisemblablement dû au fait que le vocabulaire français est, dans notre exemple, plus étendu que le vocabulaire anglais. Ceci s'explique d'une part parce que le corpus français considéré est de nature légèrement différente : les auteurs des textes étant des critiques de métier, ils ont vraisemblablement un vocabulaire plus riche que les auteurs des critiques du site Amazon. D'autre part, cette détection s'appuie sur les formes fléchies des mots, qui sont plus nombreuses en français du fait d'une morphologie flexionnelle plus riche. Notons que l'intégralité des marqueurs détectés selon cette méthode ne sera pas forcément utilisée par les classifieurs automatiques d'opinion. Il s'agit essentiellement d'indicateurs pour repérer d'éventuelles difficultés dans l'adaptation d'un domaine à un autre.

Dans un second temps, nous avons vérifié que ces mots étaient effectivement utilisés par des classifieurs automatiques d'opinion.

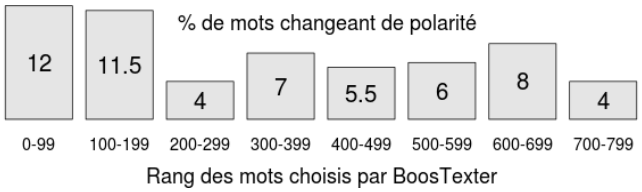


FIGURE 5: Nombre de mots changeant de polarité parmi les mots sélectionnés par boostexter par tranche de 100 mots

Nous avons utilisé une méthode de boosting, BoosTexter [Schapire and Singer, 2000], car il est possible de voir quels mots sont choisis comme classifieurs faibles. Plus un mot est sélectionné tôt, plus il est jugé discriminant pour la tâche de classification. Pour chaque paire source-cible, nous nous sommes intéressée aux rangs des mots changeant de polarité et retenus par BoosTexter comme classifieurs faibles. La figure 5 présente le nombre moyen de mots changeant de polarité par tranche de 100 classifieurs faibles. Parmi les 100 premiers classifieurs faibles, ainsi que les 100 suivants, 10 à 12 % changent de polarité. Ce sont donc autant d'erreurs de transfert possibles.

Le tableau 9 donne des exemples de mots ou bigrammes détectés comme marqueurs multi-polaires entre deux domaines avec une différence de positivité d'au moins 0,5 qui sont effectivement utilisés par le classifieur BoosTexter. C'est pour cette raison que les marqueurs ne sont pas les mêmes suivant le sens d'adaptation. En effet, un mot va servir de classifieur faible pour un sens donné d'adaptation s'il est polaire pour le domaine source. Une erreur peut alors se propager si ce même terme est non-polaire dans le corpus cible ou bien polaire mais de manière opposée. Ce dernier cas, le passage de très positif à très négatif et inversement, est en réalité peu fréquent. Le nombre de marqueurs multi-polaires utilisés par BoosTexter dans les deux sens d'adaptation va donc varier. La propagation des erreurs n'a donc pas la même ampleur selon le sens d'adaptation. Cela se retrouve dans les expériences pratiques effectuées au chapitre 8.
On remarque également que certains éléments sont intuitivement des syntagmes marquant clairement une opinion tels que "loved it" ou bien "smart", alors que d'autres sont plus surprenant comme "found the" ou bien "returning". L'étude de la pertinence des marqueurs sélectionnés est effectuée dans les chapitres 6 et 7. Nous verrons que certains marqueurs font par exemple partie de tournures de phrases habituelles dans certains domaines et pas dans d'autres ou bien font

Books → DVDs	profound, one with, asking
Books → Electronics	been able, the house, someone who, loved it, internal, easier, one with, thin
Books → Kitchen	loved this, loved it, the house, no matter, easier
DVDs → Books	smart, had never
DVDs → Electronics	one day, i loved, b, found the, case and
DVDs → Kitchen	i loved, times and, times i
Electronics → Books	region, return, someone who
Electronics → DVDs	returning, them but, draw, value, region, expected, broken, today, theater
Electronics → Kitchen	suction, places, floor, delivery
Kitchen → Books	or other, the cover, return, comfortable, pain
Kitchen → DVDs	returning, based on, followed, so far, pain, broken
Kitchen → Electronics	places
Jeux vidéo → Films	cruel, rome, se prend, pauvres, sont à, lenteur
Film → Jeux vidéo	gentils, ludique, répétition, du héros, et fait, de créer, intéressante, de effets, fbi

TABLE 9: Mots utilisés par un classifieur qui sont détectés comme fortement multi-polaire (différence de positivité de plus de 0,5)

référence à des utilisations jugées normales pour certains types d'objets et anormales pour d'autres.

5.3 VERS UNE SÉLECTION NON SUPERVISÉE

La sélection de marqueurs multi-polaires en utilisant des corpus annotés pose le problème de l'obtention des ces annotations. A l'heure actuelle, il existe plusieurs méthodes pour obtenir des corpus thématiques globalement annotés au niveau du texte [Pak and Paroubek, 2010; Giesbrecht, 2010; Généreux et al., 2008]. Par exemple, sur les plate-formes dédiées à la critique de produits commerciaux, les utilisateurs peuvent souvent rajouter une note qui est également récupérable. Il apparaît néanmoins nécessaire de corriger les biais des critiques : certaines études montrent en effet que si on demande à des humains de deviner le score mis par l'auteur de la critique, l'accord inter annotateur devient très bas si l'échelle a plus que trois niveaux [Grouin et al., 2007]. Souvent, les textes avec des scores moyens sont écartés et seuls les textes avec des notes extrêmes sont gardés pour former deux classes opposées, l'une positive et l'autre négative. Lorsque les textes sont associés à des étiquettes explicites (*tags*), il est également possible de se constituer un corpus annoté selon la polarité en sélectionnant les textes associés à des mots dont la polarité est connue. Cela est notamment le cas des corpus de tweets avec l'utilisation des étiquettes dites *hashtags* [Davidov et al., 2010; Kouloumpis et al., 2011].
Cependant, tous les domaines ne disposent pas de telles annotations, encore moins dans toutes les langues. De plus, on peut vouloir s'intéresser non pas aux textes entiers mais seulement à des thèmes précis qui nécessitent de ne prendre en compte qu'une seule partie des textes. Or obtenir automatiquement des annotations pour des sous-thèmes de textes est beaucoup plus ardu. Il existe donc de nombreux cas de figure pour lesquels l'acquisition de données annotées n'est pas aisée. C'est pourquoi il est intéressant de prendre pour contrainte l'absence d'annotation dans le corpus cible, afin de traiter ces cas de figure.
La question que nous nous posons à présent est donc de savoir comment détecter des marqueurs multi-polaires sans utiliser d'annotations sur le corpus cible.

5.3.1 *Description d'une méthode de sélection semi-supervisée*

Nous proposons d'utiliser une liste de mots ou bigrammes, que nous appellerons mots pivots, présents à la fois dans le domaine source et le domaine cible, utiles à la classification de l'opinion dans le domaine source et de polarité stable. Ces mots pivots servent à

comparer la distribution des autres mots dans les domaines source et cible. Supposons que les mots "bien" et "mauvais" fassent partie des mots pivots. Ils apparaissent dans le domaine source et dans le domaine cible. Par hypothèse, comme ils sont utiles à la classification des textes du domaine source selon l'opinion de leur auteur, "bien" va plus se retrouver dans les critiques étiquetées positives du domaine source et "mauvais" va plutôt se retrouver dans les critiques étiquetées négatives du domaine source. L'hypothèse fondamentale est que ces mots pivots indiquent la même polarité dans les deux domaines. On suppose alors, sans avoir aucunement accès à des annotations du domaine cible, que, comme pour le domaine source, "bien" va plus se retrouver dans les critiques positives du domaine cible et "mauvais" va plutôt se retrouver dans les critiques négatives.

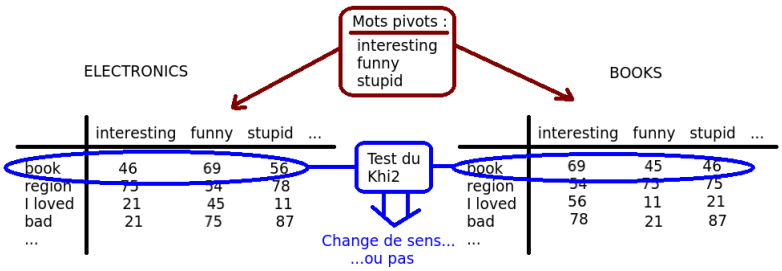


FIGURE 6: Procédure de détection des marqueurs multi-polaires à l'aide d'une collection de mots pivots.

Considérons maintenant le mot candidat "navet". Si ce mot est un marqueur de positivité dans le domaine source, il va plutôt apparaître dans les textes positifs et donc apparaître en même temps que "bon" mais pas avec "mauvais". Si ce mot est également un marqueur de positivité dans le domaine cible, il va de nouveau apparaître aux côtés de "bon" mais pas de "mauvais". Ainsi, si "navet" est un marqueur de positivité dans les deux domaines, son profil de co-occurrence par rapport à "bon", "mauvais" et les autres mots pivots va être similaire dans le domaine source et le domaine cible. Au contraire, si "navet" est un marqueur de négativité dans le domaine cible, il va apparaître aux côté de "mauvais" mais pas de "bon". Dans ce cas, le profil de co-occurrence de "navet" par rapport aux mots pivots sera très différent dans le domaine source et dans le domaine cible. Si "navet" est neutre dans le domaine cible, ses profils de co-occurrence par rapport aux mots pivots dans le domaine source et dans le domaine cible seront également différents. C'est donc en comparant statistiquement ces profils de co-occurrence par rapport aux mots pivot dans le domaine source et dans le domaine cible que l'on détermine si un mot ou bigramme candidat est effectivement ou non un marqueur

multi-polaire. Cette comparaison est effectuée à l'aide d'un test du χ^2, comme présenté à la figure 6.

5.3.2 *Sélection des mots pivots*

La difficulté réside donc dans la sélection de mots pivots, dont on veut être sûr qu'ils indiquent une polarité stable à travers les deux domaines sans utiliser aucune annotation dans le domaine cible. Nous proposons de sélectionner les mots pivots en deux temps. Une présélection est d'abord effectuée pour que les mots choisis ne soient pas trop représentatifs d'un domaine et qu'ils soient utiles à la classification de l'opinion pour le domaine source, ensuite un processus itératif permet d'épurer cette liste des mots pouvant changer de polarité.
Tout d'abord, nous calculons l'information mutuelle entre la présence et l'absence d'un mot dans une critique et l'appartenance au domaine source ou cible. Cette information mutuelle doit être faible afin que ce mot ne soit pas représentatif d'un domaine. En utilisant les étiquettes du domaine source, nous calculons également l'information mutuelle entre la présence d'un mot dans une critique source et l'étiquette positive ou négative de celle-ci. Cette fois-ci, l'information mutuelle doit être élevée, afin que ces mots soient utiles pour la détection de la polarité. Nous avons réalisé deux types de présélection, en donnant la priorité à l'une ou l'autre de ces informations mutuelles.
Une fois les mots pivots candidats sélectionnés, nous effectuons la procédure de détection des mots changeant de polarité sur les mots pivots candidats eux-même (figure 7). Nous éliminons ainsi de la liste le mot le plus susceptible de changer de polarité. Puis nous recommençons jusqu'à ce que plus aucun mot de la liste ne soit considéré comme changeant de polarité. La liste finalement obtenue est présentée à l'annexe C.

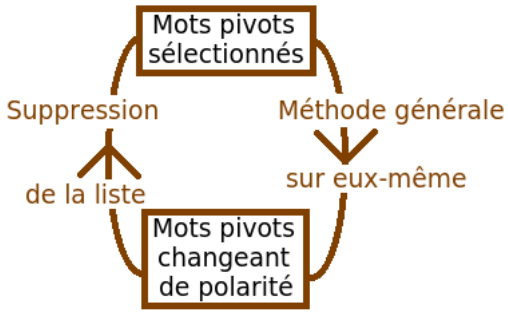


FIGURE 7: Procédure d'épuration de la collection de mots pivots.

Une fois en possession d'une liste de mots pivots propres à une paire de domaines, il ne reste plus qu'à calculer les profils de co-

occurrence des mots candidats par rapport à cette liste.

5.3.3 *Résultats et pistes d'amélioration*

5.3.3.1 *Comparaison avec la méthode supervisée*

Pour évaluer l'approche proposée, nous comparons les mots obtenus automatiquement par notre méthode avec ceux obtenus en utilisant les étiquettes source et cible, comme décrit dans la section 5.2. Nous testons les résultats pour deux jeux de mots pivots, présélectionnés soit en privilégiant l'information mutuelle source-cible, soit l'information mutuelle positif-négatif. Comme le montre le tableau 10, la méthode automatique sélectionne beaucoup plus de mots que la méthode supervisée (433 mots), ce qui explique le faible score pour la précision. Ce score peut être amélioré sans baisser le rappel en ne prenant que les mots pour lesquels on est le plus sûr du changement de sens, en fixant un seuil sur la valeur-p du test du χ^2 (précision max).

	nb mots	précision	rappel	précision max
S-C	1828 mots	14.50 %	61.20 %	16.11 %
P-N	1604 mots	16.52 %	61.20 %	18.39 %

TABLE 10: Comparaison des mots sélectionnés par la méthode automatique avec ceux sélectionnés par la méthode supervisée selon deux jeux de mots pivots, présélectionnés en privilégiant soit l'information mutuelle source-cible (S-C), soit l'information mutuelle positif-négatif (P-N). Domaine source : *DVDs* ; domaine cible : *kitchen*.

Beaucoup de marqueurs multi-polaires sont donc sélectionnés en premier par notre méthode, ce qui est très encourageant. L'étude et la mise au point de seuils de sélection optimaux est donc une piste privilégiée pour le développement futur de ces travaux. On remarque également que la méthode de présélection des mots pivots joue un rôle important dans les performances.

5.3.3.2 *Étude de la collection des mots pivots*

Nos expériences ont montré que les mots ou bigrammes détectés comme marqueurs multi-polaires varient beaucoup en fonction de la collection de mots pivots utilisée. Nous avons donc étudié la composition de la collection donnant de meilleurs résultats afin de dégager des pistes d'amélioration.

Afin de vérifier la pertinence des pivots sélectionnés après épuration, nous les avons triés selon leur positivité, calculée sur le domaine

source, et regardé s'ils apparaissaient dans les mêmes textes. Tout d'abord, nous remarquons que sur les 244 mots pivots restant après épuration, 117 sont négatifs et 127 positifs. La collection obtenue est donc équilibrée bien qu'à aucun moment du processus de sélection cela ne soit un objectif explicite. De plus, sur les 117 mots pivots négatifs, 81 ont une positivité de 0, c'est à dire que dans le corpus source, ils apparaissent uniquement dans des critiques négatives. Les 36 mots pivots négatifs restant ont une positivité entre 0 et 0,3. Quant aux 127 mots pivots positifs, 77 ont une positivité de 1, apparaissant uniquement dans les critiques positives du domaine source. Les 50 mots pivots positifs restant ont une positivité entre 1 et 0,7. Ainsi, comme prévu, les mots pivots obtenus sont tous fortement valués, aucun ne se situant dans la zone plus neutre de 0,3 à 0,7.

Si la plupart des mots pivots positifs (resp. négatifs) apparaissent uniquement dans les mêmes textes que les autres mots pivots positifs (resp. négatifs), certains apparaissent également avec les mots pivots de l'autre polarité. Il s'agit en général des mots pivots ayant une positivité autre que 0 ou 1. Il est par conséquent fort probable que ces mots pivots apparaissent également avec des marqueurs candidats d'une positivité opposée à la leur. Cela n'entraîne pas nécessairement des erreurs dans la détection des marqueurs multi-polaires mais donner un rôle différent aux mots pivots tranchés (positivité de 0 et 1) et aux autres pourrait être une bonne idée.

Il est également important de remarquer que tous les mots pivots d'une même polarité n'appa raissent pas tous ensemble. Ainsi, un marqueur candidat pourrait apparaître dans le corpus source uniquement avec certains mots pivots positifs et dans le domaine cible avec des mots pivots différents mais également positifs. Un tel marqueur candidat serait alors détecté comme marqueur multi-polaire alors qu'il reste en réalité positif dans les deux corpus. Cela pourrait expliquer le surplus de marqueurs multi-polaires détectés par notre méthode semi-supervisée.
Ainsi, plutôt que de considérer le profil de co-occurrence avec chacun des mots pivots il serait judicieux de rassembler les mots pivots en groupes. Le profil de co-occurrence serait alors constitué des apparitions communes avec tous les mots pivots négatifs stricts (positivité de 0), des apparitions communes avec tous les mots pivots négatifs légers, des apparitions communes avec tous les mots pivots positifs légers et des apparitions communes avec tous les mots pivots positifs stricts (positivité de 1).

5.3.3.3 *Différentes co-occurrences*

Il est également intéressant de s'interroger sur le calcul du score d'association entre les marqueurs candidats et les mots pivots. Les ré-

sultats présentés ci-dessous ont été obtenus en calculant la co-occurrence au niveau du texte. Nous avons également essayé de calculer cette co-occurrence sur des fenêtres glissantes de différentes tailles mais les marqueurs multi-polaires obtenus n'étaient pas meilleurs.
Une autre tentative a été d'estimer ces scores avec d'autres mesures permettant de capter des proximités sémantiques différentes, telles que les méthodes de factorisation de contexte [Mikolov et al., 2013; Erk, 2012; Van de Cruys et al., 2011]. Nous avons par exemple utilisé le modèle neuronal de Mikolov et al. [2013] en entraînant un réseau de neurones différent pour le domaine source et pour le domaine cible. Nous avons alors utilisé la mesure cosinus entre les représentations des marqueurs candidats et des mots pivots afin d'établir l'équivalent du profil de co-occurrence. Comme nous entraînons deux réseaux de neurones différents, un sur le corpus du domaine source et l'autre sur le corpus du domaine cible, les profils de co-occurrence obtenus pour chaque marqueur candidat sont différents pour le domaine source et le domaine cible. Nous pouvons alors les comparer avec un test du χ^2 comme précédemment. Cette façon de calculer les niveaux de lien entre les marqueurs candidats et les mots pivots n'a pas réussi à obtenir de meilleurs résultats, quelque soit la largeur de la fenêtre utilisée pour le calcul des réseaux de neurones.
Il pourrait cependant être intéressant de comparer directement les représentations obtenues pour chaque marqueur candidat avec les réseaux de neurones entraînés sur les corpus source et cible sans passer par la distance cosinus avec les mots pivots. Cette façon de procéder aurait l'avantage d'être entièrement non-supervisée puisqu'il n'y aurait nul besoin des étiquettes du domaine source pour sélectionner les mots pivots.

5.4 CONCLUSION

Les marqueurs multi-polaires sont donc des mots ou groupes de mots qui indiquent, selon le domaine, une polarité au niveau du texte. Il est facile de les détecter lorsque l'on dispose de corpus de deux domaines différents, chacun annotés au niveau des textes. Nous avons proposé une méthode afin de détecter ces marqueurs sans utiliser d'annotation dans le domaine cible, ce qui se rapproche des cas réels d'utilisation. Nous reprenons l'idée des mots pivots ayant un comportement semblable entre les deux domaine. Il est dans ce cas nécessaire de s'assurer que les mots pivots aient une polarité stable. Nous réalisons cela par une auto-épuration de la collection de mots candidats. Le rappel des résultats obtenus valide l'intérêt de notre approche. Cependant, la très faible précision ne permet actuellement pas une utilisation en situation réelle. Nous avons proposé plusieurs possibilités d'amélioration, notamment en différenciant le rôle des mots pivots à polarité stricte et non-stricte, ou bien en utilisant des représentations

vectorisées des mots candidats. Ces développements devront être étudiés plus en détail dans les prolongements de cette thèse.

Nous avons ensuite cherché à savoir si les marqueurs ainsi sélectionnés étaient ressentis comme multi-polaires par des êtres humains. Le chapitre suivant présente l'expérience d'annotation que nous avons menée afin d'étudier cette question.

6

MISE EN PLACE D'UNE EXPÉRIENCE D'ANNOTATION MULTI-ANNOTATEURS

Dans le chapitre précédent, nous avons présenté le concept de marqueur multi-polaire. Il s'agit de mots ou groupes de mots qui indiquent, selon le domaine, une polarité différente au niveau du texte. Nous avons montré que les marqueurs détectés en utilisant des annotations des domaines source et cible au niveau du texte sont effectivement utilisés par des classifieurs d'opinion statistiques. Une analyse manuelle des marqueurs ainsi sélectionnés a mis en lumière plusieurs phénomènes linguistiques différents.
Nous souhaitons à présent voir si des annotateurs humains ressentent également ces marqueurs sélectionnés automatiquement comme multi-polaires. Nous faisons l'hypothèse suivante : si le marqueur sélectionné est bel et bien porteur d'indications quant à la polarité du texte, un annotateur humain devrait pouvoir déduire cette polarité de la phrase contenant ce marqueur.

6.1 CRÉATION D'UNE INTERFACE D'ANNOTATION

Le test du χ^2 qui permet de détecter les marqueurs multi-polaires utilise une annotation en terme de polarité au niveau global du texte. Le but de notre campagne d'annotation est de vérifier que l'information sur l'avis de l'auteur, la polarité, se retrouve bien autour du marqueur, c'est à dire dans la phrase qui le contient, et non pas ailleurs dans la critique.
Pour séparer les critiques en phrases, nous avons utilisé la bibliothèque logicielle en Python NLTK avec les séparateurs par défaut, pour le français ainsi que pour l'anglais. L'interface d'annotation a, quant à elle, été réalisée en utilisant le *framework* Web Python Django. Les instructions données aux annotateurs étaient les suivantes :

Objectif :

Vous allez voir des phrases extraites de critiques de produits. Il peut s'agir de livres, de films, d'objets électroniques ou du quotidien. L'objectif est de dire si, à la lecture de cette phrase, l'auteur de la critique a, selon vous, écrit une critique ***positive****,* ***négative*** *ou* ***neutre****.*
Autrement dit, nous nous intéressons à la ***polarité globale*** *de la critique dont les phrases proposées ne sont que* ***des extraits****. Pour vous aider dans l'annotation, le type d'objet ainsi que son nom sera indiqué.*

ATTENTION *: Certains extraits ne donnent* ***pas d'information*** *sur la polarité de la critique. Dans ce cas, vous cocherez la case "****ne sais pas****".*

Faites également attention à ***l'objet dont parle l'extrait*** *qui peut être* ***différent de l'objet général*** *de la critique (qui vous est précisé). Par exemple, un extrait peut faire* ***l'éloge d'un produit concurrent*** *ou bien parler du premier tome d'une série alors que la critique porte sur le second. Une phrase positive peut donc parfois être le signe d'une* ***critique négative*** *selon l'objet précis dont elle parle.*

Afin de déterminer l'état d'esprit de l'auteur, vous êtes encouragés à vous appuyer sur vos ***connaissances de la vie courante****. Par exemple, si une phrase parle de délais de livraison trop longs, même si ce n'est pas une critique directe de l'objet, c'est plutôt* ***mauvais signe*** *pour la satisfaction de l'auteur et la* ***polarité globale*** *de la critique sera* ***vraisemblablement négative****.*

Ces instructions sont accompagnées des phrases exemples présentées à la figure 8 qui permettaient de se familiariser avec l'apparence de l'interface.

L'annotateur passe ensuite à la phase d'annotation proprement dite. Dix phrases à annoter lui sont présentées. Une fois qu'il a validé son choix pour ces dix phrases, ses réponses sont enregistrées dans la base de donnée. Après de très chaleureux remerciements, il lui est proposé soit de continuer à annoter des phrases dix par dix, soit de se déconnecter.

Les phrases présentées sont choisies dans la base de données parmi celles que l'annotateur n'a pas encore annotées. L'algorithme de choix va préférentiellement présenter des phrases qui ne possèdent qu'une seule annotation afin d'être sûr d'avoir au moins deux annotations de deux annotateurs différents par phrase. Ensuite, les phrases sont présentées aléatoirement.

Des exemples de phrases proposées lors de cette tâche d'annotation sont présentés à l'annexe E.

6.2 VALIDATION DU GUIDE D'ANNOTATION

6.2.1 *Sélection des phrases*

Afin de valider notre guide d'annotation, nous avons réalisé une première étude sur deux marqueurs multi-polaires détectés entre les domaine *electronics* et *books* du corpus anglais. Ces marqueurs ont été choisis car ils présentaient une très large différence de polarité entre les deux domaines. Il s'agit du mot *return* et du bigramme *I loved*.

FIGURE 8: Phrases fournies en exemple aux annotateurs.

Après analyse, ces marqueurs sont notamment respectivement des représentants des catégories *changement de sens* et *changement d'utilisation temporelle*, que nous définissons à la section 7.3. Nous avons travaillé sur 150 phrases, 44 contenant *I loved* et 106 *return*.
Pour cette pré-étude, nous disposions de 5 annotateurs, dont 4 n'étaient pas spécialistes du traitement automatique des langues. Les annota-

FIGURE 9: Quelques récompenses pour les annotateurs courageux.

teurs sont désignés avec les lettres A, B, C, D et E, l'annotateur familier avec le domaine étant l'annotateur A.

6.2.2 *Mesures d'accord entre deux annotateurs*

Il existe plusieurs façon d'obtenir une estimation du taux d'accord entre les annotateurs pour une tâche d'annotation [Fort, 2012]. Nous présentons dans cette section quelques scores inter-annotateurs classiques avant d'expliquer le score personnalisé que nous avons mis en place pour notre travail à la section 6.2.3. L'analyse des accords obtenus et des résultats sont présentés à partir de la section 6.2.4.

6.2.2.1 *Accord observé*

La mesure la plus évidente d'accord inter-annotateurs est le pourcentage d'accord ou accord observé (A_o), qui correspond à la proportion d'éléments sur lesquels les annotateurs sont d'accord, autrement dit, le nombre total d'éléments (i) " annotables " pour lesquels il y a accord (agr), divisé par le nombre total d'éléments annotables. Il est défini comme suit :

$$A_o = \frac{1}{|I|} \sum_{i \in I} agr_i$$

Ce coefficient ne prend cependant pas en compte le hasard, qui peut influencer les résultats. Ainsi, un schéma d'annotation comprenant un petit nombre de catégories donnera de meilleurs accords observés, uniquement du fait du hasard. En outre, cette mesure d'accord ne compense pas la distribution des éléments dans les catégories : une

catégorie prépondérante va très largement influencer l'accord observé [Artstein and Poesio, 2008].

6.2.2.2 π *de Scott*

Le coefficient π [Scott, 1955], considère que les distributions réalisées par les annotateurs par hasard sont équivalentes, mais il suppose que la répartition des éléments (i) entre catégories (k) n'est pas homogène et qu'elle peut être estimée par la répartition moyenne réalisée par les annotateurs. L'accord attendu (A_e^π) est donc calculé de la façon suivante, n_k étant le nombre d'affectations à k pour les deux annotateurs :

$$A_e^\pi = \sum_{k \in K} \left(\frac{n_k}{2i}\right)^2$$

π se calcule ensuite comme suit :

$$\pi = \frac{A_o - A_e^\pi}{1 - A_e^\pi}$$

6.2.2.3 κ *de Cohen*

Le coefficient κ [Cohen, 1960] suppose dans sa modélisation du hasard que la répartition des éléments entre catégories peut être différente pour chaque annotateur. Dans ce cas, la probabilité pour qu'un élément (i) soit assigné dans une catégorie (k) est le produit de la probabilité que chaque annotateur l'assigne dans cette catégorie. L'accord attendu (A_e^κ) est donc calculé de la façon suivante, n_{c1k} étant le nombre d'affectations à k pour l'annotateur 1 :

$$A_e^\kappa = \sum_{k \in K} \frac{n_{c1k}}{i} . \frac{n_{c2k}}{i}$$

κ se calcule ensuite comme suit :

$$\kappa = \frac{A_o - A_e^\kappa}{1 - A_e^\kappa}$$

Il faut noter que $\pi \leqslant \kappa$ et que s'ils sont proches (ce qui est souvent le cas) [Di Eugenio and Glass, 2004], cela signifie qu'il y a peu de biais entre les annotateurs. Il est donc intéressant de calculer ces deux coefficients.

6.2.3 *Généralisation et adaptation des mesures d'accords*

Les accords que nous avons présentés sont valables pour deux annotateurs annotant les mêmes instances et pour lesquelles la distance entre chaque classe possible a la même valeur.

6.2.3.1 *Généralisations classiques des accords inter-annotateurs*

Tout d'abord, les distances entre les trois classes possible lors de l'annotation ne sont pas les mêmes. En effet, il est plus grave d'avoir un désaccord entre positif et négatif qu'entre positif et neutre. De plus, dans notre expérience, si un annotateur ne peut pas se prononcer, il doit également cocher la case "neutre/ne sais pas". Celle-ci est donc encore moins éloignée des deux autres.
Afin de prendre en compte ces différentes distances entre les classes, il existe des indices d'accords pondérés tels que l'indice α [Krippendorff, 1980, 2004]. Les poids des désaccords impliquant deux classes différentes sont rassemblés dans ce qu'on appelle une matrice de pondération. La principale difficulté pour ce type d'accord est de définir une matrice de pondération adaptée à la tâche considérée [Fort, 2012].

Il existe également des généralisations des accords classiques afin de prendre en compte plusieurs annotateurs. Par exemple, le multi-π ou κ de Fleiss, est une généralisation de π de Scott. Il considère, pour chaque élément à annoter, le nombre de paires d'annotateurs en accord par rapport à toutes les paires de jugements possibles.
Il est également possible de généraliser le κ de Cohen à plusieurs annotateurs en suivant le même principe. Pour plus d'information sur les calculs techniques de ces indicateurs, le lecteur pourra se reporter à Fort [2012] Cependant, dans l'annotation à grande échelle que nous avons menée par la suite (voir section 6.3), le nombre de phrases annotées était laissé au bon vouloir des participants. Aussi, tous les participants n'ont pas annoté le même nombre de phrases. Pour ceux ayant annoté de grandes quantités, il est tout à fait possible de calculer une distribution des classes qui leur sont propres. Par contre, il est impossible d'estimer une telle distribution pour les participants ayant annoté dix phrases, qui représentent la majorité. Aussi avons nous choisi d'adapter le score multi-π et non le multi-κ.

6.2.3.2 *Accord inter-annotateurs personnalisé*

Le coefficient multi-π considère que chaque annotateur annote toutes les instances. Or, ce n'est pas le cas dans notre expérience. Nous avons donc adapté sa formule afin d'en tenir compte.
L'accord observé multi-annotateurs se calcule donc par paires d'annotations de la façon suivante :

$$A_o^{\pi\text{-sm}} = \frac{1}{i}\sum_{i\in I} agr_i = \frac{1}{i}\sum_{i\in I}\frac{1}{c_i(c_i-1)}\sum_{k\in K} n_{ik}(n_{ik}-1)$$

Pour l'accord attendu, cela donne :

$$A_e^{\pi\text{-sm}} = \sum_{k\in K}\left(\frac{n_k}{nbAnnot}\right)^2$$

De plus, nous avons introduit des pondérations comme pour le coefficient α. En effet, il est moins problématique d'avoir un désaccord entre positif et neutre qu'entre positif et négatif.
L'accord attendu que nous avons finalement utilisé se calcule donc comme suit :

$$A_e^{\pi\text{-smp}} = \sum_{k \in K} \sum_{k' \in K} pond(k, k') \frac{n_k}{nbAnnot} \frac{n_{k'}}{nbAnnot}$$

Et pour l'accord observé :

$$A_o^{\pi\text{-smp}} = \frac{1}{i} \sum_{i \in I} \frac{1}{c_i(c_i - 1)} * \left(\sum_{k \in K} pond(k, k) * n_{ik}(n_{ik} - 1) + \sum_{k \in K} \sum_{k' \neq k} pond(k, k') * n_{ik} n_{ik'} \right)$$

Nous calculons finalement notre accord selon la formule générale classique :

$$\pi\text{-sparse-multi-pond} = \pi\text{-smp} = \frac{A_o^{\pi\text{-smp}} - A_e^{\pi\text{-smp}}}{1 - A_e^{\pi\text{-smp}}}$$

Nous avons donc défini notre propre accord inter-annotateur sur le modèle de ceux existant afin de prendre en compte le fait que les annotateurs peuvent ne pas tous annoter les mêmes phrases ni même le même nombre de phrases et également d'introduire des pondérations entre les classes.

Nous avons choisi une matrice de pondération classique : 1 si il y a accord, 0 pour un désaccord entre les classes positif et négatif et 0,5 pour un désaccord entre la classe neutre et l'une des deux autres.
Comme les annotateurs avaient pour consigne de cocher la case neutre en présence de phrases neutres mais également en présence de phrases ne permettant pas de décider, la classe neutre n'est en réalité pas exactement au milieu des deux autres. Cependant, comme il ne nous est pas possible de quantifier la part des phrases non décidables, nous avons décidé de conserver la pondération de 0,5 entre la classe neutre et les deux autres (3 classes - pond 0,5). Il s'agit d'un choix conservatif, les accords obtenus étant ainsi légèrement sous-estimés.

Nous présentons également les scores obtenus en considérant que les trois classes sont à égale distance à des fins de comparaison (3 classe - sans pond). Nous présentons aussi le score obtenu en ne prenant en compte que les paires de jugements pour lesquelles les deux annotateurs n'ont pas choisi la classe neutre (2 classes). Ce score indique donc si, quand les annotateurs sont d'accord pour considérer une phrase comme polaire, ils lui affectent la même polarité.

6.2.3.3 *Interprétation de la valeur des accords*

L'interprétation de la valeur des accords inter-annotateurs obtenus n'est pas consensuelle [Artstein and Poesio, 2008; Fort, 2012]. Des exemples d'interprétation du coefficient κ issus de la littérature sont présentés à la figure 10.

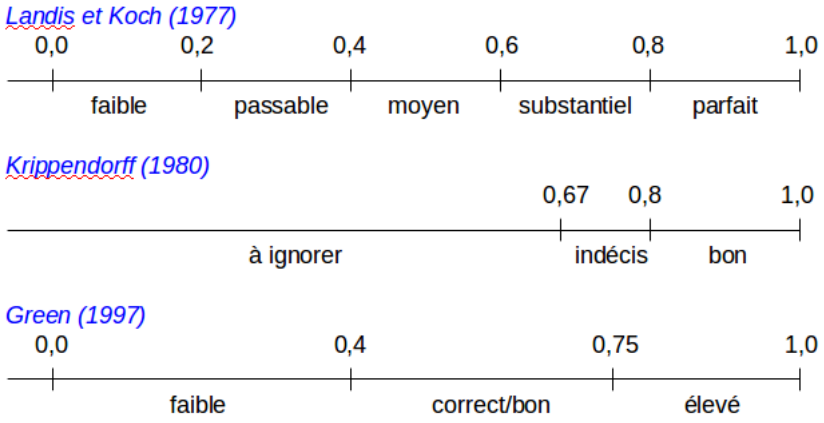


FIGURE 10: Différentes échelles d'interprétation du coefficient κ. D'après Fort [2012]

6.2.4 *Analyse de l'accord inter-annotateurs global*

Nous avons calculé l'accord observé sur ces données ainsi que le κ de Fleiss pour les cinq annotateurs ensembles (tableau 11).

		Annotateurs considérés			
		A,B,C,D,E	B,C,D,E	A,B,C,D	B,C,D
Accord observé	3 classes sans pond	74,1 %	71,7 %	79,3 %	78,2 %
	3 classes pond 0,5	84,6 %	83,1 %	87,6 %	86,4 %
	2 classes	89,1 %	88,2 %	90,8 %	89,7 %
κ de Fleiss	3 classes sans pond	57,4 %	53,8 %	64,7 %	62,7 %
	3 classes pond 0,5	66,1 %	62,8 %	72,0 %	69,5 %
	2 classes	76,1 %	74,5 %	78,9 %	76,7 %

TABLE 11: Accords calculés sur différents groupes d'annotateurs lorsque tous les choix sont pris en compte (3 classes) avec ou sans pondération ou bien seulement les choix positifs et négatifs (2 classes). Haut : Accord observé (avec plusieurs annotateurs). Bas : κ de Fleiss (avec plusieurs annotateurs).

Nous avons également calculé ces scores en excluant l'annotateur ayant une connaissance du traitement automatique des langues (annotateur A) et, comme on peut le constater, l'accord baisse alors légèrement tout en restant correct. L'analyse des accords entre les annotateurs deux à deux présentée à la section suivante (6.2.5) montre que l'annotateur E présente un fort biais par rapport au autres annotateurs, les résultats globaux sont donc également présentés sans sa contribution.
Sans ce dernier, nos scores d'accords pondérés sont, selon les interprétations, substantiels, indécis ou bon, ce qui montre que notre protocole d'annotation est efficient pour annoter la polarité au niveau des phrases. Lorsque l'on ne considère que les cas où les annotateurs sont d'accord pour assigner une polarité à la phrase, les résultats sont encore plus convergents.

6.2.5 *Analyse des accords inter-annotateurs deux à deux*

Nous avons également calculé le π de Scott et le κ de Cohen pour chaque paire d'annotateurs pour trois (tableau 12) et deux classes (tableau 13).

π de Scott

	A	B	C	D	E
A	.	65.9%	63.9%	69.8%	51.8%
B		.	66.2%	60.7%	44.3%
C			.	61.1%	41.6%
D				.	49.1%
E					.

κ de Cohen

	A	B	C	D	E
A	.	66.1%	63.9%	69.9%	52.7%
B		.	66.5%	60.8%	46.0%
C			.	61.2%	42.6%
D				.	49.7%
E					.

TABLE 12: Haut : π de Scott entre deux annotateurs pour trois classes. Bas : κ de Cohen entre deux annotateurs pour trois classes.

Comme évoqué précédemment, ces chiffres montrent bien que l'annotateur E est une exception. Son taux d'accord avec les autres annotateurs est très faible lorsqu'on considère les trois classes. Cependant, ce score devient tout à fait correct lorsqu'on ne considère que les phrases pour lesquelles les annotateurs ont fait un choix autre que neutre. C'est pourquoi nous avons comparé la répartition du choix des classes pour chaque annotateur (tableau 14).

π de Scott

	A	B	C	D	E
A	.	86.3%	95.5%	91.9%	82.4%
B		.	82.0%	82.8%	75.1%
C			.	87.2%	76.4%
D				.	91.1%
E					.

κ de Cohen

	A	B	C	D	E
A	.	86.3%	95.5%	91.9%	83.0%
B		.	82.0%	82.8%	75.8%
C			.	87.2%	77.3%
D				.	91.3%
E					.

TABLE 13: Haut : π de Scott entre deux annotateurs pour deux classes. Bas : κ de Cohen entre deux annotateurs pour deux classes.

	A	B	C	D	**E**
nb pos	25,3 %	29,3 %	23,3 %	29,3 %	**34,0 %**
nb neut	18,7 %	10,7 %	21,3 %	18,0 %	**28,0 %**
nb neg	56,0 %	60,0 %	55,4 %	52,7 %	**38,0 %**

TABLE 14: Répartition des choix de classe des annotateurs en pourcentage.

Comme on peut le constater, l'annotateur E a sélectionné très peu de fois la classe négative par rapport aux autres annotateurs. Nous avons regardé plus en détails les phrases pour lesquelles tous les annotateurs hormis lui avaient choisi la classe négative dont voici quelques exemples :

- *Too bad its not worth the return shipping charges.*
- *But I'm just thankful Amazon has a customer-friendly return policy.*
- *I purchased one in March 2006 and returned it September 2006, as I had purchased replacement insurance and was able to return it because it broke.*
- *I owned this product for about one week, very nice little piece which I can take to work or jogging, but today when I am copying some files to it from my computer, it stopped working, even reset didn't work, so I will return it for a refund and try different products.*
- *However, Amazon.com was very quick and gracious in the return.*
- *Item was shipped with a short in the electronics and had to return.*
- *Two weeks after putting in a fresh set of batteries and not using the keyboard, you'll return to find them dead.*
- *I was able to return the D2, but the GEX-P10XMT and the IB100II are not returnable.*

– *I was informed I would have to return the entire unit.*

Nous avons discuté avec cet annotateur et lui avons demandé de confirmer et d'expliquer son choix pour ces phrases. Il a cette fois choisi la classe négative pour quelques unes mais a de nouveau choisi la classe neutre pour une grande majorité. En effet, pour lui ces phrases n'apportent pas d'information spécifiquement sur le produit discuté mais sur des aspects annexes de la vente. Or, nous nous intéressons à l'avis général de la critique, nous voulons que ces aspects annexes soient pris en compte. Il a donc été nécessaire de réécrire les instructions d'annotation afin d'insister sur le fait que nous sommes intéressée par l'humeur générale de la critique, pas seulement par un avis factuel sur l'objet de la critique.

6.2.6 *Validation des hypothèses et lancement de l'annotation à grande échelle*

En utilisant les étiquettes de polarité au niveau des textes, nous estimons la positivité de *I loved* à 9,1 % dans le domaine *electronics* et à 75,5 % dans le domaine *books*. Pour *return*, la positivité est de 5,5 % dans *electronics* et 63,3 % dans *books*.
Nous définissons un score de positivité lié à l'annotation manuelle des phrases comme la somme pondérée des choix effectués :

$$P_{humain} = 1 * N_p + 0.5 * N_x + 0 * N_n$$

avec respectivement N_p, N_x et N_n le nombre de fois où une phrase contenant le marqueur étudié a été jugée positive, neutre ou négative.

		score texte	score phrastique
I loved	books	75,5 %	85.8 %
	elec	9,1 %	59.1 %
		score texte	score phrastique
return	books	63,3 %	41.7 %
	elec	5.5 %	13 %

TABLE 15: Comparaison du score de positivité calculé à l'aide des étiquettes au niveau du texte par rapport à celui calculé à l'aide des annotations au niveau des phrases. Haut : *I loved*. Bas : *return*.

Comme l'on peut le voir dans le tableau 15, les scores de positivité calculés automatiquement au niveau du texte sont à chaque fois sous-estimés par rapport à ceux phrastiques, calculés à partir des annotations humaines, sauf dans le cas de *return* dans le domaine *books*. De plus, les scores très négatifs obtenus pour le marqueur *I loved* sur le corpus *electronics* en utilisant les étiquettes au niveau des textes est ramené vers le neutre avec les jugements au niveau des phrases.

Paire de corpus	Marqueurs étudiés
DVDs-Electronics	i loved, one day, them but, today, returning, region, broken, b, case and, draw, theater, expected, found the, value
Books-Kitchen	loved this, loved it, no matter, pain, return, or other, easier, comfortable, the house, the cover

TABLE 16: Liste des marqueurs choisis. Ils présentent une différence de positivité d'au moins 0,5 entre les deux domaines auxquels ils sont liés.

Néanmoins, les écarts de positivité entre les deux corpus, bien qu'atténués, demeurent bien présents. Ainsi, un jugement humain au niveau des phrases conserve aux marqueurs *I loved* et *return* leur statut de marqueur multi-polaire. Cette pré-étude valide donc notre méthode d'évaluation des marqueurs multi-polaires.

6.3 EXPÉRIENCE MULTI-ANNOTATEURS

6.3.1 *Présentation des mots choisis*

Nous avons décidé d'étudier les mots ou bigrammes sélectionnés par le classifieur boostexter comme classifieurs faible, c'est à dire utiles à la classification des textes, et qui se trouvaient être des marqueurs multi-polaires forts. C'est à dire que la différence de score de positivité calculé par l'ordinateur en fonction de leurs apparitions dans des critiques positives et négatives dans le domaine source et dans le domaine cible est de plus de 0,5. Rappelons que le score de positivité va de 0 (complètement négatif) à 1 (complètement positif). Une différence de 0,5 peut donc faire passer un mot de complètement neutre à complètement valué.
Nous avons choisi de nous intéresser aux paires de domaines *DVDs-Electronics* et *Books-Kitchen* étant donné que les domaines *DVDs* et *Books* se comportent de manière proche ainsi que *Electronics* et *Kitchen*. La liste des marqueurs étudiés se trouve au tableau 16. Ces marqueurs se retrouvent au final dans 1160 phrases de nos corpus.

6.3.2 *Accords inter-annotateurs globaux multi-annotateurs*

En plus de nous même, qui avons annoté l'intégralité des phrases, 51 annotateurs ont participé à l'expérience. Comme la durée de l'expérience était libre, les participations varient de 10 à 668 annotations par annotateur.

	3 classes - sans pond	3 classes - pond 0,5	2 classes
$A_O^{\pi\text{-smp}}$	78,6 %	87,5 %	92,5 %
π-smp	67,1 %	73,7 %	84,9 %

TABLE 17

En définitive, nous obtenons un accord inter-annotateur global pondéré de 73,7 % pour trois classes et de 84,9 % pour deux classes. Ces scores sont cohérents avec les mesures effectuées lors de la pré-étude, et même meilleurs, ce qui montre donc que le passage à l'échelle a été réussi.

6.3.3 *Score de positivité au niveau des phrases*

Pour chaque marqueur étudié, nous regardons les annotations effectuées par domaine. Comme expliqué à la partie 6.2.6, le score de positivité phrastique est la somme pondérée des choix effectués par les annotateurs.
Nous avons testé ce score en prenant en compte tous les votes, c'est à dire de 3 à 7 votes selon les phrases, ou bien en prenant le vote majoritaire pour chaque phrase où apparaît le marqueur étudié. Comme les scores calculés ainsi ne diffèrent pas beaucoup, nous avons conservé la première façon de calculer le score afin d'avoir plus de poids statistique. De même, nous avons observé les scores obtenus avec ou sans nos propres annotations. Là encore, les différences observées étant faibles, nous avons conservé toutes les annotations dans l'analyse.

6.4 ANALYSE DES POSITIVITÉS DES MARQUEURS AU NIVEAU DES PHRASES

Pour chaque marqueur multi-polaire, nous avons comparé le score de positivité calculé à l'aide des étiquettes au niveau du texte (référence) avec celui calculé à partir des annotations des phrases (score humain) et ce, pour chacun des deux domaines. Les figures 11 et 12 montrent les écarts de scores de positivité entre les deux domaines source et cible pour chacune des méthodes de calcul. Les barres du graphique représentent donc la différence de positivité d'un marqueur multi-polaire entre les deux domaines. Dans l'idéal, les barres bleu foncé (score humain) devraient être semblables aux barres bleu claire rayées (référence). En pratique, les scores calculés à l'aide des annotations des phrases sont très souvent moins tranchés, c'est à dire que la différence de positivité entre les deux domaines est réduite. La plupart du temps, la positivité est conservée dans un domaine tout

en étant plus neutre dans le deuxième. Cela est dû à plusieurs phénomènes que nous étudions plus en détail ci-dessous.
En pratique, un marqueur est considéré comme validé comme marqueur multi-polaire si sa barre bleu foncé demeure conséquente. C'est à dire si la différence de positivité entre les deux domaines demeure, même si celle-ci est réduite.

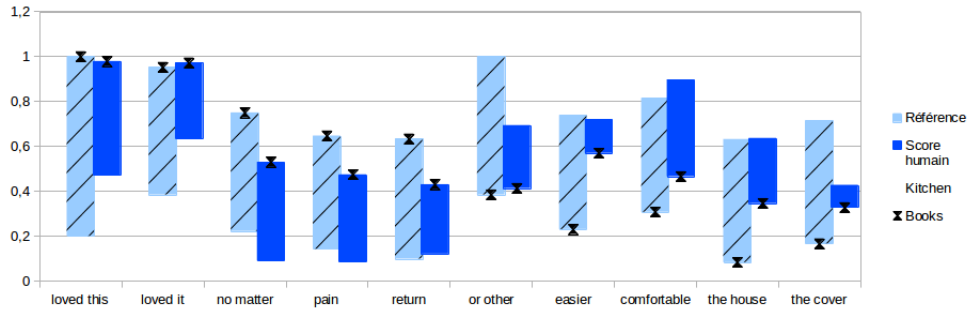


FIGURE 11: Écart de positivité selon le score calculé au niveau des textes sur le corpus et le score phrastique calculé grâce aux annotations entre les domaines *Books* et *Kitchen*.

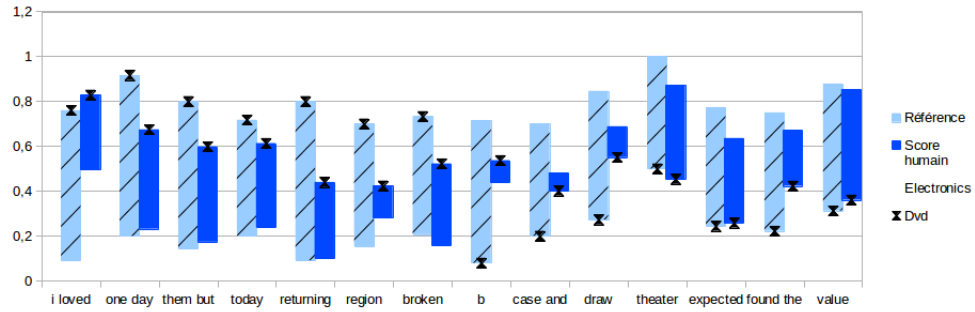


FIGURE 12: Écart de positivité selon le score calculé au niveau des textes sur le corpus et le score phrastique calculé grâce aux annotations entre les domaines *DVDs* et *Electronics*.

6.4.1 *Différences dans les écarts de positivité*

On remarque tout d'abord, qu'il est très rare qu'un score phrastique soit plus polaire (proche des extrêmes 0 et 1 par rapport au neutre 0,5) qu'un score calculé automatiquement au niveau du texte.
Certains mots semblent être des biais de corpus étant donné que les scores phrastiques sur les deux domaines sont très proches. C'est par exemple le cas de la lettre *b* pour laquelle on observe une diminution drastique de l'écart entre les positivités sur les deux domaines ainsi même qu'une inversion.
Pour la plupart des marqueurs étudiés, l'écart de positivité est réduit.

Cela peut s'expliquer par l'attraction de la classe "neutre/ne sais pas". Il y a néanmoins souvent un domaine pour lequel le score de positivité est conservé et le deuxième pour lequel il y a une réduction vers le neutre.

Nous rappelons ici que les marqueurs multi-polaires potentiels ont entre autres été sélectionnés à l'aide d'un seuil d'apparition minimal assurant pour chaque domaine une présence dans les critiques positives et négatives permettant une analyse statistique significative à l'aide du test du χ^2. Il y avait lors de cette étape de sélection deux classes : positif et négatif. Nous regardons à présent la répartition des marqueurs dans trois classes : positif, négatif et neutre/ne sais pas. Ainsi, certains mots qui passaient le critère avec deux classes ne le passent plus avec trois classes. Il s'agit de *or other*, *one day*, *them but*, *b*, *case and* et *draw*. Nous les avons donc exclus des analyses. Pour tirer des informations des phrases ayant été annotées contenant ces marqueurs, il faudrait utiliser des méthodes statistiques non paramétriques, qui permettent de traiter des faibles échantillons. Cette analyse n'a pas été conduite dans cette thèse.

6.4.2 *Coefficient d'accord S*

Il est important de pouvoir juger de l'accord des différents annotateurs sur l'annotation des phrases pour chaque domaine et pour chaque marqueur. Ainsi, il est possible d'avoir une information sur la clarté des phrases et sur la difficulté de la tâche d'annotation pour chaque marqueur.

Comme nous l'avons expliqué à la section 6.2.4, l'accord observé seul n'est pas pertinent car il ne prend pas en compte l'effet du hasard. Dans le cas qui nous préoccupe, le κ de Cohen n'est pas non plus indiqué. En effet, comme les phrases à annoter étaient présentées de manière aléatoire aux différents annotateurs et que la durée de participation était libre, ce ne sont pas les mêmes annotateurs qui ont annoté toutes les phrases contenant un même marqueur multi-polaire. Il est donc impossible de calculer une distribution des classes propre à chaque annotateur pour un marqueur particulier.
Le coefficient classique qui semble, dans un premier temps, le plus indiqué ici est le π de Scott qui modélise le hasard de manière non-équirépartie en estimant la distribution des différentes classes par la répartition moyenne réalisée par les annotateurs. Ce coefficient a été créé afin de contrebalancer l'effet d'une classe prépondérante. Or, lorsqu'un marqueur multi-polaire est très tranché, nous sommes précisément dans une situation de classe prépondérante que nous ne voulons surtout pas effacer. C'est pourquoi nous avons jugé que dans notre cas, le coefficient le plus adapté était le coefficient S.

Le coefficient S est l'accord observé corrigé du hasard modélisé comme équiréparti entre les classes [Bennett et al., 1954]. L'accord attendu (A_e^S) est donc calculé de la façon suivante, n_k étant le nombre d'affectations à k pour les deux annotateurs :

$$A_e^\pi = \sum_{k \in K} \left(\frac{n_k}{card(K)} \right)^2$$

S se calcule ensuite comme suit :

$$S = \frac{A_o - A_e^S}{1 - A_e^S}$$

Le coefficient S sert ici surtout à comparer l'accord des annotateurs entre deux mots. Il ne donne pas de valeur d'accord absolu.

Nous présentons les valeurs du coefficient S à la section 7.2, en même temps que les répartitions des marqueurs dans des phrases positives, négatives ou neutre selon les votes des annotateurs.
Certains domaines sont particulièrement difficiles pour certains marqueurs (en enlevant ceux avec trop peu d'annotation) : easier-books, loved it-kitchen, loved this-kitchen, no matter-books, i loved-elec, region-elec et today-dvd. D'autres sont particulièrement accordés : comfortable-kitchen, loved it-books, loved this-books, pain-kitchen, i loved-dvd, returning-elec et dvd, theater-elec, today-elec et value-elec.

6.5 CONCLUSION

Dans cette partie, nous avons présenté l'expérience d'annotation que nous avons mise en place afin de vérifier que les marqueurs multi-polaires sélectionnés au niveau du texte sont bien porteur de l'information polaire. Pour ce faire, nous avons souhaité vérifier que la restriction aux phrases contenant ces marqueurs multi-polaires permettait toujours de les identifier comme tels. Notre annotation, réalisée grâce à ciquante et un annotateurs volontaires, a montré que pour la plupart des marqueurs sélectionnés automatiquement l'écart de positivité est réduit mais se conserve bel et bien.
La partie suivante rend compte de l'observation des différents comportements des marqueurs multi-polaires et présente la classification des différents types de changement de polarité que nous avons établie.

7

CLASSIFICATION DES MARQUEURS MULTI-POLAIRES

Dans le chapitre précédent, nous avons décrit la mise en place de notre expérience d'annotation. Nous allons à présent exploiter les données rassemblées et observer le comportement des marqueurs multi-polaires en situation dans le corpus afin d'en déduire une classification typologique.
Nous commençons tout d'abord par étudier le comportement des marqueurs dans un seul domaine. Nous observons ensuite les différences de comportement des marqueurs selon les domaines et en quelle mesure ces comportements sont distinguables. Enfin, nous établirons une classification des explications de ces changements de comportement.

7.1 COMPORTEMENT DES MARQUEURS DANS UN DOMAINE

Un marqueur a un comportement dans chaque domaine qui induit une répartition en phrases positives, négatives et neutres. Par définition, puisqu'il s'agit d'un marqueur multi-polaire, ce comportement va être différent d'un domaine à l'autre.
Nous présentons ci-dessous les types de comportement dans un domaine particulier. Un marqueur peut être un représentant des différentes parties classiques de l'expression de l'opinion : la cible de l'opinion, l'expression de l'opinion et la raison de l'opinion. Par exemple, dans la phrase "Je trouve que le couvercle est lourd.", "Je trouve" indique l'expression de l'opinion, "le couvercle" est l'objet sur lequel porte l'opinion et "lourd" est la raison de l'opinion.
Pour un marqueur donné, chaque comportement se traduit dans un corpus par l'apparition de ses occurrences dans des phrases positives, négatives ou neutres en proportions différentes. Nous définissons donc un profil de répartition des mots dans les différents types de phrases que nous représentons comme à la figure 13.

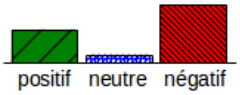


FIGURE 13: Exemple de profil de répartition des occurrences d'un marqueur dans les phrases positives, négatives et neutres. Ici, le marqueur apparaît beaucoup dans des phrases négatives, un peu dans des phrases positives et presque pas dans des phrases neutres.

7.1.1 *Contexte : parler d'autre chose*

Le premier comportement observé est l'apparition des marqueurs dans des phrases de contexte, extérieures à l'expression d'une opinion. Ce comportement est spécifique au corpus considéré et se traduit par une répartition majoritaire sur des phrases neutres vis à vis de l'objet de la critique.
Pour les corpus *Books* et *DVDs*, un mot peut faire partie de la **description de l'intrigue** sans qu'un jugement soit porté sur celle-ci. Cela est propre au fait que les livres et les films racontent des histoires. Nous séparons ce comportement de la description objective d'un objet car dans une description d'histoire, des mots *a priori* très fortement polaires peuvent être utilisés dans un contexte en réalité neutre.

– return/book - An elf appearing to Harry Potter warned him not to return for a second year to the school of sorcery.

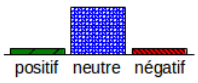
Catégorie : Contexte	Profil
- Description de l'intrigue - Description du contexte - Parler d'autre chose - Parler d'un autre objet similaire	

TABLE 18: Profils de répartition dans les phrases positives, négatives et neutres des marqueurs utilisés dans des phrases de contextes.

Dans les corpus *Electronics* et *Kitchen*, les critiques parlent d'objets manufacturés produits en plusieurs exemplaires. Chaque exemplaire est censé être identique mais ça n'est pas toujours le cas et les conditions d'utilisation peuvent changer d'un objet à l'autre. Aussi, les phrases parlant d'un **objet similaire mais différent** ne donne pas à proprement parlé d'indication sur la satisfaction de la personne vis à vis de son propre exemplaire. Les deux objets peuvent fonctionner de manière semblable ou différente. Il est rare qu'au niveau de la phrase l'information soit suffisante pour le déterminer, aussi, ces marqueurs apparaissent dans des phrases neutres vis-à-vis de l'opinion globale du texte. Le profil de répartition associé à ces comportements est représenté au tableau 18.

– loved it/kitchen - My father-in-law had the pan at his house and we just loved it.

7.1.2 *Opinion : expression d'un jugement*

Certains marqueurs multi-polaires vont indiquer qu'un **jugement non neutre** est exprimé. Les phrases les contenant vont être soit positives, soit négatives mais jamais neutres.
Le marqueur peut indiquer un **jugement toujours positif**, un **jugement toujours négatif** ou bien un **jugement non exclusif**. Dans ce dernier cas, en observant le corpus, soit le marqueur est équitablement réparti entre des phrases positives et des phrases négatives, on le dit alors **sans préférence**, soit il apparaît plus souvent dans une catégorie de phrases, on le dit alors **avec préférence**. Cette préférence observée peut selon les cas être expliquée par exemple par l'usage d'une formulation ou d'une expression spécifique, par des présupposés culturels ou encore un simple biais de corpus. L'explication peut varier mais le type de comportement au niveau du domaine est le même.

– i loved/dvd - I loved the Virgin Suicide's character of Lux.
– found the/dvd - I found the double-sided disc annoying.

<table>
<tr><th colspan="3">Catégorie : Opinion</th><th>Profil</th></tr>
<tr><td rowspan="4">- Expression d'un jugement actuel

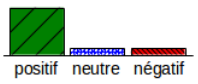
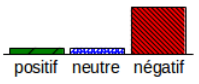
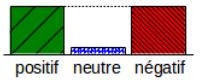
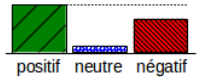
- Expression d'un changement de jugement</td><td colspan="2">positif / glissement vers le positif</td><td> 
</td></tr>
<tr><td colspan="2">négatif / glissement vers le négatif</td><td> 
</td></tr>
<tr><td rowspan="2">sans indication</td><td>répartition équitable</td><td> 
</td></tr>
<tr><td>répartition préférentielle</td><td> 
</td></tr>
</table>

TABLE 19: Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une expression d'opinion.

Sur le même modèle, un marqueur peut exprimer un **changement de jugement**, comme présenté à la figure 19. Nous retrouvons les sous-catégories équivalentes au premier cas : **changement vers le négatif**, **changement vers le positif**, **changement non précisé avec ou sans préférence observée**.

– loved it/kitchen - I loved it for about 2 months.

7.1.3 *Raison : qualité ou défaut*

Un marqueur peut désigner une **propriété explicite** de l'objet, une **propriété induite** ou bien un **effet provoqué** par l'objet que ce soit une action de l'auteur de la critique ou bien un ressenti émotionnel ou physique. Ces propriétés ou effets peuvent être soit **positifs**, le marqueur apparaît alors principalement dans des phrases positives, soit **négatifs**, le marqueur apparaît alors principalement dans des phrases négatives, soit **normales**, le marqueur apparaît alors dans tous les types de phrases. Ces comportements sont présentés dans le tableau 20

– comfortable/kitchen - The mugs have a curved top and nice handles that are comfortable to hold.
– expected/dvd - This film was not what I expected.
– no matter/books - you're sure to find something interesting in it no matter what page you choose.
– broken/dvd - The DVD is broken down into sections and thoroughly explains each topic.

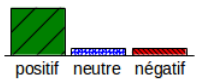
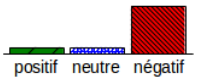
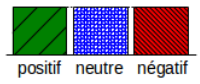
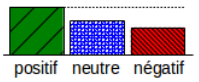
Catégorie : Raison			Profil
- Propriété explicite - Propriété induite - Effet provoqué - Action provoquée	positif		
	négatif		
	normal	répartition équitable	
		répartition préférentielle	

TABLE 20: Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une propriété digne d'opinion.

– return/book - You will return for second and third helpings.
– returning/dvd I watched it over 6 times before returning it back.

– pain/kitchen - It's a pain to clean out the nozzle and the filter.

7.1.4 *Cible : élément de l'objet*

Un marqueur peut parfois désigner un **élément de l'objet** dont on parle dans la critique, **l'objet lui-même** ou bien un **mot générique du domaine** concerné. Il y a différents types d'éléments possibles. Il peut s'agir d'un élément comme un autre dont on parle lors de la description et sur lequel **donner un avis est optionnel**. Il peut alors apparaître dans des phrases positives, négatives ou neutres, avec ou sans préférence constatée.
Certains éléments de l'objet **implique obligatoirement de donner son avis**. Ces mots apparaissent dans des phrases positives et négatives, avec ou sans préférence constatée (tableau 21).

– the cover/books - I read this book because the cover and title sounded good.
– the cover/kitchen - To put the cover on, you need to stretch it to the point that several keys at the edge are forced down.

– the house/kitchen - Great for light work in the car or around the house.
– theater/electronics - My Bose home theater system could not find and play certain stations that this player can pick and play very clearly.

7.1.5 *Possibilité de préférence*

Lorsque le comportement linguistique du marqueur induirait une répartition équitable entre plusieurs classes (positif et négatif ou positif, négatif et neutre) mais qu'on observe une différence significative dans le corpus, on parle de préférence. En observant le corpus, nous avons observé plusieurs causes de préférences. Elles sont le reflet d'un cadre de pensées commun, d'ordre culturel, moral ou politique, entre le lecteur et l'auteur de la critique. On parle parfois de stéréotypes culturels [Vernier, 2011] . Il peut s'agir d'une tournure de phrase habituelle, d'un objet ou d'une partie d'un objet dont la présence est bonne en soi sans besoin de préciser ou bien d'une caractéristique tellement évidente et nécessaire qu'on ne songe à la mentionner que lorsqu'elle n'est pas assurée.

– region/electronics - The player didn't play DVDs from Germany or from the UK, or from any other country, but Region code 1 DVDs.

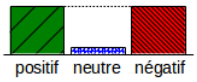
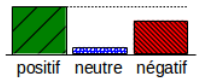
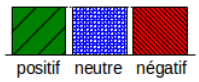
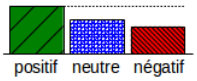
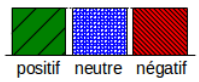
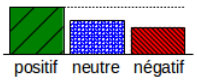
Catégorie : Cible			Profil
- Partie de l'objet	avis obligatoire	répartition équitable	
		répartition préférentielle	
	avis optionnel	répartition équitable	
		répartition préférentielle	
- Objet lui-même - Terme générique	normal	répartition équitable	
		répartition préférentielle	

TABLE 21: Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une cible d'opinion.

- theater/electronics - Home theater systems usually do not include optical/digital coax cables, so do not forget to order one together with your home theater system.
- value/electronics - This phone works great, looks sleek and is a great value for the price.

7.2 PAIRES DE COMPORTEMENTS DISTINGUABLES ENTRE DEUX DOMAINES

Un mot ou bigramme est réellement un marqueur multi-polaire si dans les deux domaines, il adopte un comportement différent distinguable. En effet, certains comportements, bien que différents, ne vont pas être distinguables. Cela peut par exemple être le cas si un mot est une propriété normale dans un domaine et l'expression d'un jugement non exclusif sans préférence dans le second domaine. Ce mot aura une positivité autour de 0,5 dans les deux domaines et ne sera

donc pas repéré. Il est à noter qu'il ne gênera pas non plus le transfert d'information d'un domaine à un autre.

Une difficulté survient lorsqu'un mot a plusieurs comportements à l'intérieur d'un même domaine. Les différents comportements peuvent se compenser et le score final sera encore une fois plutôt neutre. C'est souvent le cas dans les corpus *DVDs* et *Books* ou un mot, en plus d'un comportement particulier, va également être utilisé dans une phrase descriptive de l'intrigue.
Au final, un mot ou bigramme est effectivement détecté comme marqueur multi-polaire si dans au moins un des domaines, il a un comportement ou un mélange de comportements entraînant une polarité claire. Si le mot a une polarité claire dans les deux domaines, il faut que ces polarités soient opposées.

Nous présentons dans les sections suivantes certaines associations de comportements distinguables qui sont plus fréquentes que les autres. Pour chacune, nous présentons les marqueurs multi-polaires qui fonctionnent majoritairement selon ce schéma. Pour certains marqueurs, le comportement global distingué est en réalité dû à une superposition de plusieurs comportements. Nous proposons également des explications de ces différences de comportement, qui peuvent être variées selon les marqueurs. Un récapitulatif de ces explications est proposé dans la section 7.3.

7.2.1 *Polaire versus neutre : description prédominante*

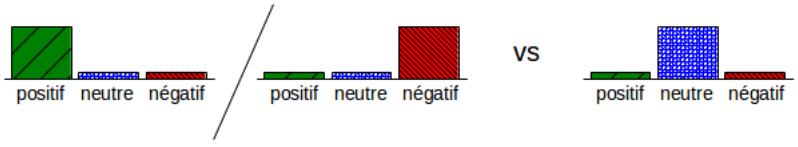


FIGURE 14: Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil uniquement neutre.

Certains mots restent principalement polaires dans un domaine et très neutres dans l'autre. Les profils de comportements purs associés sont présentés à la figure 14. Dans le cas réel, il s'agit des profils de comportement prédominants auxquels s'ajoutent d'autres profils de moindre importance.
Les marqueurs multi-polaires *return* ou *pain* se comportent de cette manière. D'autres mots ont parfois très peu d'annotations pour un domaine car ils apparaissaient peu de fois. Aussi, les résultats sont moins significatifs mais vont tout de même dans le même sens. C'est par exemple le cas de *region* ou bien *returning* qui apparaît peu dans

le corpus *DVDs* (tableau 22).

	nb pos	nb neut	nb neg	score S
return - books	12,1 %	60,3 %	27,6 %	0,69
return - kitchen	3,5 %	7,1 %	89,4 %	0,72
pain - books	10,7 %	71,4 %	17,9 %	0,62
pain - kitchen	4,3 %	0,0 %	95,7 %	0,88
returning - dvd	9,1 %	72,7 %	18,2 %	0,87
returning - elec	4,8 %	2,4 %	92,8	0,87
region - dvd	0,0 %	75,0 %	25,0 %	0,65
region - elec	15,4 %	15,4 %	69,2 %	0,59

TABLE 22: return, returning, pain et region

On remarque qu'à chaque fois la neutralité se trouve dans les corpus *Books* et *DVDs*. En effet, ces corpus contiennent en général des descriptions de l'intrigue des films ou livres. Lorsque les marqueurs apparaissent dans un contexte descriptif, la phrase est en général neutre, que le sens du marqueur soit polarisé ou non.

RETURN ET RETURNING Lorsque l'on parle d'un objet électroménager, employer le mot *return* fait presqu'exclusivement référence au renvoi de l'objet pour cause de mauvais fonctionnement. Il est bien sûr possible que l'objet soit remplacé rapidement et que l'impression générale reste bonne mais la plupart du temps c'est quand même le signe d'un jugement négatif à propos de l'objet incriminé.
Pour les livres, c'est le neutre qui l'emporte, avec beaucoup de phrases descriptives. Il y a quelques emplois spécifiques au domaine des livres, qui indique une envie de relecture qui est un signe d'appréciation. Cependant, cet effet est globalement noyé par la neutralité de la description.

Le marqueur *returning* fonctionne comme *return* avec quelques phrases neutres supplémentaires mentionnant la location de DVDs.

PAIN Dans les critiques de livres, lorsqu'apparaît le mot *pain*, il s'agit en général de la description de l'intrigue d'un roman. Et bien sûr, la fait que les protagonistes souffrent de grandes douleurs morales ou physiques ne donne aucune indication sur la qualité du roman.
A l'inverse, la douleur physique ou bien la difficulté ressentie lors de la manipulation d'ustensiles de cuisines mal adaptés est un défaut

bien souvent rédhibitoire.

REGION Le cas de *region* est légèrement différent. Bien que le faible nombre d'annotations ne permette pas de faire de généralité, le mot *region* est en général employé dans le domaine *electronics* pour parler du *region code*. Il se trouve que lorsque l'on pense évidentes certaines caractéristiques, on ne pense pas à les mentionner lorsqu'elles sont effectivement présentes. On ne les mentionne que lorsqu'il y a un problème avec elles. Nous avions également le mot *english* qui ressortait en négatif pour le domaine *electronics*. En effet, personne ne pense à préciser que l'employé du service après vente parle bien anglais lorsque c'est le cas.

7.2.2 *Polaire versus neutre : mélange de trois classes*

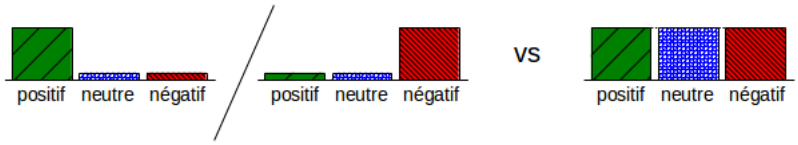


FIGURE 15: Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil mélangeant les trois classes.

Pour certains marqueurs, un domaine est fortement polaire alors que l'autre est neutre en raison d'un mélange entre les trois classes. Ce type de profils est schématisé à la figure 15. Le tableau 23 présente quatre exemples de marqueurs avec un tel comportement.

	nb pos	nb neut	nb neg	score S
comfortable - books	25,0 %	41,7 %	33,3 %	0,67
comfortable - kitchen	86,2 %	3,4 %	10,4 %	0,89
easier - books	42,8 %	28,6 %	28,6 %	0,56
easier - kitchen	70,2 %	6,4 %	23,4 %	0,70
no matter - books	22,2 %	50,0 %	27,8 %	0,57
no matter - kitchen	0,0 %	0,0 %	100,0%	0,68
broken - dvd	28,6 %	42,8 %	28,6 %	0,67
broken - elec	6,9 %	13,8 %	79,3 %	0,78

TABLE 23: comfortable, easier, no matter et broken

COMFORTABLE Lorsque l'on parle des appareils électroménagers, le mot *comfortable* est non ambigu et est clairement une caractéristique physique enviable pour les objets dont on parle. Quelques rares phrases emploient ce mot de concert avec des négations mais habituellement c'est le mot *uncomfortable* qui est employé pour parler d'un objet qui n'est pas confortable.
Dans les critiques de livres par contre, la situation est très mélangée. Le mot *comfortable* peut être employé dans la description de l'intrigue, pour parler de l'objet livre en lui-même qui peut être lourd à porter, pour signifier un accord ou un désaccord avec les idées développées dans un essai ou encore pour vanter les vertus d'un livre de développement personnel. Il semble qu'ici, à part la critique sur le format du livre, chaque emploi du mot *comfortable* est plus précisément associé à un certain type de livre : romans, essais, conseils au développement personnel. Si les critiques de livres étaient séparées selon le type de livre, il est vraisemblable que le mot *comfortable* soit également détecté comme changeant de polarité entre ces domaines plus restreints. Comme nous nous sommes ici intéressée au domaine "livres" en général, les effets de ces différents emplois sont mélangés et les phrases sont équiréparties entre les trois classes.

EASIER De la même façon, l'emploi de *easier* dans les critiques de livres recouvre un mélange de plusieurs réalités. Le mot *easier* indique une comparaison, reste à savoir de quoi on parle. On peut critiquer l'intrigue, la difficulté de compréhension d'un écrit, les idées développées dans un essai ou encore faire référence à l'évolution des capacités du lecteur lorsqu'il s'agit de livres présentant un apprentissage. Au final, le mot *easier* se retrouve de manière équitable dans des phrases positives, négatives ou neutres dans les critiques de livres. Le score de positivité calculé à partir des phrases est donc neutre alors que celui calculé à partir des textes était très négatif. Cela peut avoir pour raison les phrases de descriptions qui sont neutres dans des critiques forcément positives ou négatives. On constate également que beaucoup des phrases positives contenant le mot *easier* portent sur la facilité de compréhension. Il est possible que la facilité de lecture ne soit pas le critère majoritaire lorsque l'on juge un livre et que des livres inintéressants soient faciles à lire, introduisant ainsi une polarité différente lorsque l'on juge au niveau du texte ou au niveau de la phrase. Enfin, le score S pour le mot *easier* dans le domaine des livres est faible (0,56), montrant que les annotateurs ont eu parfois des difficultés à juger, ce qui peut expliquer l'importance de la classe neutre/ne sais pas.
Dans le domaine des appareils électroménagers, l'emploi du terme *easier* implique bien plus clairement une comparaison de l'objet de la critique avec un autre objet équivalent, la difficulté étant de savoir

quel objet est "plus facile" que l'autre. Il y a très peu de phrases jugées neutres et trois fois plus de phrases positives que de phrases négatives contenant ce mot. Il est difficile de dire si le sens préférentiel de comparaison observé ici est un phénomène que l'on retrouverait dans tous les corpus d'appareils électroménager ou bien un biais de ce corpus particulier.
Au vu des annotations au niveau des phrases, on ne peut pas retenir *easier* comme un marqueur multi-polaire. Dans le domaine *kitchen*, le marqueur exprime bien une opinion comparative présentant une préférence vers la classe positive. Par contre, le comportement de *easier* dans le corpus *books* que l'on observait au niveau des textes n'est pas conservé au niveau des phrases.

NO MATTER Le marqueur *no matter* se révèle être très intéressant. Comme souvent, le corpus de critiques de livres contient une bonne part de phrases relatives à la description de l'histoire et qui n'apportent donc pas d'information sur l'opinion de l'auteur de la critique. Par contre, les mots *no matter* sont également employés pour exprimer l'universalité. Cette dernière peut être positive ou négative. A l'inverse, le comportement de *no matter* dans le corpus *kitchen* est très tranché. L'intégralité des phrases le contenant est jugée en majorité négative. Ces mots sont en effet employés pour évoquer l'idée de non adaptabilité, qui est une caractéristique négative.

BROKEN Lorsqu'un appareil électronique est cassé, c'est plutôt mauvais signe pour la satisfaction client, que le dommage soit constaté dès l'ouverture du colis ou bien survienne quelques temps après l'achat. Par contre, le mot *broken* peut exprimer différentes choses dans des critiques de films. Il peut faire référence au contexte, au sentiments éprouvés en regardant le film ou encore à l'organisation de l'objet dvd en chapitres.

7.2.3 *Polaire versus neutre : mélange de deux classes*

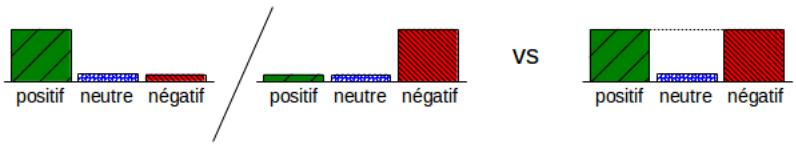


FIGURE 16: Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil mélangeant les classes positives et négatives.

Certains mots sont très polaires dans un domaine et ressortent comme neutre dans l'autre car ils sont en réalité presque équirépartis entre des phrases positives et négatives, comme l'on peut le constater au tableau 24. Il peut s'agir de l'expression d'un jugement sans préférence comme dans le cas de *found the* ou bien d'un mélange d'expression d'un jugement positif avec l'expression d'un changement de jugement vers le négatif comme dans le cas de *i loved*. Les profils de comportements purs associés sont schématisés à la figure 16.

	nb pos	nb neut	nb neg	score S
loved it - books	95,2 %	4,8 %	0,0 %	0,91
loved it - kitchen	53,8 %	7,7 %	38,5 %	0,57
loved this - books	100,0 %	0,0 %	0,0 %	0,91
loved this - kitchen	42,9 %	0,0 %	57,1 %	0,45
i loved - dvd	76,8 %	10,7 %	12,5 %	0,88
i loved - elec	41,7 %	0,0 %	58,3 %	0,34
found the - dvd	40,0 %	8,0 %	52,0 %	0,75
found the - elec	65,0 %	20,0 %	15 %	0,61
value - dvd	34,5 %	3,4 %	62,1 %	0,66
value - elec	85,7 %	1,4 %	12,9 %	0,84

TABLE 24: loved it, loved this, i loved et found the

I LOVED, LOVED IT ET LOVED THIS *loved it*, *loved this* et *i loved* fonctionnent tous de la même façon. Bien qu'ils apparaissent moins sur les corpus *Kitchen* et *Electronics* que *Books* et *DVDs*, comme ils sont tous les trois formés sur *loved*, leur rassemblement semble légitime et donne à nos observations une plus grande significativité.
Il s'agit ici de l'exemple type du fait que le rapport temporel à l'objet peut influer sur les marqueurs qui sont des verbes polaires au passé. Il est en effet normal de parler d'un livre ou d'un film au passé car il faut les avoir lus/vus au préalable avant d'en parler. Alors que parler au passé d'un objet dont on est censé se servir quotidiennement est souvent le signe que notre avis à changé. Les phrases positives au passé employant *loved* pour parler d'objets du quotidien ne font en fait souvent pas référence à l'objet spécifiquement acheté par l'auteur de la critique mais à un objet semblable possédé par le passé ou bien appartenant à un proche. C'est la bonne expérience vécue dans le passé sur un autre objet qui a motivé l'achat de l'objet dont on se sert au quotidien. Dans certains cas, cet achat sera suivi de déception bien que ce soit rarement visible au niveau de la phrase.

Il est à noter que les scores calculés à partir des textes les classaient

tous les trois comme très négatifs sur les corpus *kitchen* et *electronics*. Or, les avis des annotateurs sont très partagés entre positif et négatif et ce sont les cas pour lesquels il y a le plus de désaccords. Nous pensons que ces phrases pour ces domaines étaient particulièrement difficiles à évaluer, ce qui se traduit par un score S très faible (de 0,57 à 0,34 selon les cas). Cette incertitude vient en grande partie du fait qu'il est difficile, bien que nous ayons fourni le nom de l'objet général de la critique, de toujours savoir si on parle de l'objet général ou bien d'un concurrent lorsque l'on dispose d'une seule phrase de la critique. A l'inverse, les phrases parlant de livres ou de films sont non ambiguës, avec un score S allant de 0,88 à 0,91.

FOUND THE Dans les critiques de films, l'expression *found the* indique la présence d'une opinion, qu'elle soit positive ou négative. Les phrases contenant ce marqueur sont donc équitablement réparties entre les classes positives et négatives, rares étant celles classées neutres.
Dans les critiques d'appareils électroniques, on trouve également ce comportement de marqueur général d'opinion. Cependant, un autre emploi, spécifique à ce domaine, est plus répandu. Aujourd'hui, nous utilisons beaucoup d'appareils électroniques pouvant se connecter les uns aux autres, par câbles ou réseaux sans fils. Il est ainsi très important que les différents appareils électroniques "trouvent le réseau" ou "se trouvent entre eux" facilement. Il s'agit d'une caractéristique positive très importante dans ce domaine. Les phrases le mentionnant sont donc positives. Il ne s'agit pas ici d'une caractéristique tellement évidente que l'on ne pense à la mentionner que lorsqu'elle manque. Cela devrait sans doute l'être mais les problèmes d'interconnexions sont encore bien trop fréquents de nos jours.

VALUE Dans les critiques de films, le mot *value* peut faire référence aux valeurs morales véhiculées par le film ou bien à la valeur artistique dudit film. Ce sont donc à chaque fois des objets sur lesquels l'auteur a une opinion et c'est pourquoi le mot *value* apparaît presqu'exclusivement dans des phrases exprimant une opinion. Cette dernière peut par contre être soit positive soit négative.
Dans le domaine des appareils électroniques par contre, le terme *value* est presque toujours mis en relation avec le prix de l'objet : *"Very good value for the price"*. Ces phrases sont en général positives. Il existe bien évidemment des objets qui sont de piètre qualité compte tenu de leur prix mais dans ce cas les expressions utilisées ne font plus apparaître le mot *value* et se focalisent sur l'argent perdu : *"Don't waste your money."*

7.2.4 *Positif versus négatif : présence de préférences*

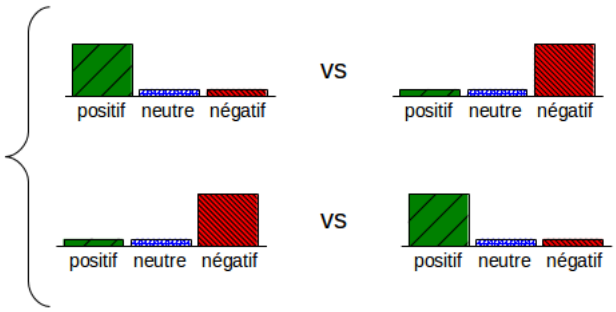


FIGURE 17: Comportement globalement positif par opposition à un comportement globalement négatif.

Les mots passant de majorité positive à majorité négative sont plus rares mais très intéressants. Ils sont en général le fruit de présence de préférences différentes dans les deux corpus. Les profils de comportements purs associés sont représentés à la figure 17. Bien évidement, en condition réelle, d'autres profils de comportement de moindre importance se mêlent à ceux-ci. Nous présentons deux exemples de tels marqueurs dans le tableau 25.

	nb pos	nb neut	nb neg	score S
expected - dvd	13,0 %	19,6 %	67,4 %	0,76
expected - elec	62,0 %	6,0 %	32,0 %	0,73
today - dvd	47,8 %	28,3 %	23,9 %	0,57
today - elec	20,8 %	8,3 %	70,9 %	0,81

TABLE 25: expected et today

EXPECTED *expected* est un très bon exemple de qualité relative. La présence de ce mot indique la plupart du temps que l'on compare l'objet dont on parle à ce qu'on attendait de lui auparavant. Il peut s'agir d'attente déçue ou confirmée. Le terme *expected* est donc par nature polaire, on ne le trouve pratiquement pas dans des phrases neutres. Il est par contre possible que la prédominance d'attentes confirmées et donc d'opinions positives par rapport aux opinions négatives soit un biais du corpus étudié. Il semble que dans le corpus *electronics* il y ait plus d'attente confirmée que déçue. En effet, c'est bien qu'un appareil photo que l'on a commandé sur internet soit tel que l'on s'y attendait.
Lorsque l'on parle des films, on retrouve cette notion d'attentes déçues ou confirmées. Par contre, un autre phénomène très présent se rajoute : si le scénario d'un film est trop convenu, sans surprise, le

film devient lassant et de peu d'intérêt. Ainsi, le marqueur *expected* prend une valeur très négative dans le corpus *DVDs*.

TODAY Lorsque l'on parle de films, le mot *today* peut bien sûr faire partie de la description de l'intrigue et ainsi apparaître dans des phrases neutres. Néanmoins, sa présence peut également indiquer une comparaison entre le film dont on parle et un autre film référent, plus ancien ou au contraire actuel. Ces comparaisons peuvent être indifféremment positives ou négatives pour le film sujet de la critique, bien que l'on observe ici une préférence pour la classe positive.
Dans le corpus dédié aux appareils électroniques, la situation est cette fois drastiquement différente. Quelques rares phrases parlent de l'évolution du prix de l'objet et non de l'objet lui-même, d'autres, positives, conseillent d'acheter ce produit aujourd'hui même. Cependant, dans la plupart des phrases, la présence du mot *today* indique une évolution dans le fonctionnement de l'appareil. Et comme il est peu courant que le fonctionnement d'un appareil électronique s'améliore tout seul, ces phrases sont négatives. Comme *I loved*, *today* a à voir avec un rapport temporel à l'utilisation d'un objet.

7.2.5 *Autres comportements distinguables*

Enfin, les profils de comportement de certains marqueurs sont moins facilement classables. Pour certains d'entre eux, les préférences observées n'ont pu trouver d'explications et sont donc vraisemblablement le fruit de biais de corpus.

	nb pos	nb neut	nb neg	score S
the cover - books	7,4 %	44,4 %	48,2 %	0,66
the cover - kitchen	25,0 %	25,0 %	50,0 %	0,68
the house - books	0,0 %	70,0 %	30,0 %	0,74
the house - kitchen	52,6 %	10,5 %	36,9 %	0,64
theater - dvd	22,9 %	37,1 %	40,0 %	0,73
theater - elec	77,8 %	16,7 %	5,5 %	0,84

TABLE 26

THE COVER Dans la cuisine, *the cover* est juste une partie comme une autre des objets dont on parle. Ils peuvent être bien faits ou mal faits, mais la simple présence d'un couvercle n'est pas suffisante pour déclencher un avis positif ou négatif. Une préférence pour la classe négative est néanmoins observée, qui dans ce cas précis est vraisem-

blablement le fait d'un biais de corpus.
Pour les livres, ce mot désigne la couverture. On en trouve des descriptions, qui peuvent être positives, négatives ou neutres. A cela s'ajoute le fait que l'on choisi souvent en livre en fonction de sa couverture qui est censée représenter fidèlement le livre et porte donc nos attentes. Et comme beaucoup de choses que l'on s'estime en droit d'attendre naturellement, on n'en fait mention que si l'on en est déçu.

THE HOUSE Pour les livres, la maison est un élément de décor neutre. Lorsque l'on parle d'équipements ménager, le terme "la maison" désigne le lieu de vie de l'auteur, qui n'a lui non plus aucune raison de porter une charge polaire. L'observation des phrases contenant le terme *the house* montre bien que la charge polaire n'est pas portée par le mot en lui-même. Nous sommes ici en présence d'un biais de corpus.

THEATER De manière similaire, si le mot *theater* apparaît dans une critique de film, c'est en général l'occasion pour l'auteur de préciser qu'il a vu ce film pour la première fois au cinéma mais cela n'a pas d'influence dans les phrases que nous avons observées sur l'opinion de l'auteur sur le film en question. Dans le domaine des appareils électroniques, le *home theater*, ou home cinéma en français, est juste un appareil comme un autre qui peut ou non fonctionner. Le biais vers la classe positive dans notre corpus peut sans doute s'expliquer par le fait qu'en 2007, date à laquelle le corpus a été constitué, les home cinémas ne sont pas très répandus dans les foyers. Aussi, même si l'appareil peut avoir ses défauts, rien que le fait d'en avoir un est quelque chose de positif. Il s'agit ici d'une influence culturelle qui s'effacera éventuellement avec le temps.

7.3 CATÉGORIES DE JUSTIFICATION DE CHANGEMENT DE COMPORTEMENT

Comme nous l'avons vu, un mot ou bigramme est retenu comme marqueur multi-polaire s'il présente un comportement, ou un ensemble de comportements, observable différent d'un domaine à un autre. Nous avons décrit précédemment les couples de comportements observés dans notre corpus et donné des explications possibles à ces changements de comportement d'un domaine à un autre. Nous proposons ici un récapitulatif ordonné des différentes raisons que nous avons rencontrées. Bien évidemment, comme un marqueur peut avoir plusieurs comportements dans un même domaine, il peut se retrouver dans plusieurs catégories.

Les différentes catégories de changement de comportement sont donc les suivantes :

- Description contextuelle
- Changement de sens
- Changement d'objet
- Changement d'utilisation
- Biais de corpus

7.3.1 *Description contextuelle*

Comme nous l'avons vu, un mot potentiellement polaire peut être utilisé de manière neutre dans la description d'une intrigue, d'une contextualisation d'un achat ou bien d'un autre objet. Toute combinaison avec un emploi polaire dans l'autre domaine donnera un marqueur multi-polaire.
Dans notre annotation, les marqueurs rentrant parfois dans cette catégorie sont : return, pain, returning, region, comfortable, no matter, broken, loved it. Des exemples de phrases sont présentés au tableau 27.

7.3.2 *Changement de sens*

Dans deux domaines différents, un mot peut changer franchement de sens et ainsi possiblement également de polarité.
Les marqueurs rentrant parfois dans cette catégorie sont : return, found the, theater, the cover, value. Des exemples de phrases sont présentés au tableau 28.

7.3.3 *Changement d'objet*

Un qualificatif peut s'appliquer à un objet différent, un mot faire référence à un objet différent. Aussi, les qualités et défauts attendus, les connotations associées, ne seront pas les mêmes. Si elles sont de polarité opposée ou tout simplement neutre et polaire, il s'agira d'un marqueur multi-polaire.
Les marqueurs rentrant parfois dans cette catégorie sont : expected, broken, no matter, comfortable. Des exemples de phrases sont présentés au tableau 29.

7.3.4 *Changement d'utilisation*

Tous les objets ne sont pas utilisés de la même manière. Aussi un mot faisant référence à un certain comportement pourra être positif, normal ou négatif selon les cas. Nous avons observé des exemples

return	Books	An elf appearing to Harry Potter warned him not to return for a second year to the school of sorcery.
	Kitchen	Go ahead and return it to the manufacturer for a refund.
pain	Books	The villain of "Prayers for Rain" is sinister, smug, brilliant, and sadistic, inflicting pain and death because he can.
	Kitchen	It's a pain to clean out the nozzle and the filter.
returning	DVDs	She deals with the returning of a former boyfriend as the season comes to a close.
	Electronics	I ended up returning the item 2 days after i bought it.
region	DVDs	Rebel troops are nearing the region where they live.
	Electronics	If you are looking for a cheap, very good, and region free DVD player this one is the right choice.
comfortable	Books	She's comfortable with the idea of laying down roots but he's not.
	Kitchen	Beautiful design, comfortable grips and well made.
no matter	Books	The Shield Zone is tighter than ever, denser than a black hole as no matter escapes this closely knit family to the chagrin of Evan.
	Kitchen	We think no matter which one you choose, the temperature is the same.
broken	DVDs	A wooden chair is broken in the ring.
	Electronics	The main unit looks broken on the backside.
loved it	Books	It is a beautiful book and the children absolutely loved it.
	Kitchen	My father-in-law had the pan at his house and we just loved it.

TABLE 27: Exemple de changement de polarité dû à l'appartenance d'une des phrases à une description de contexte.

d'utilisation directe mais également des exemples d'utilisation temporelle différente comme pour les marqueurs du type *loved* ou *today*.

return	Books	You will return for second and third helpings.
	Kitchen	Go ahead and return it to the manufacturer for a refund.
found the	DVDs	While online we found the Train Your Dog DVD.
	Electronics	The software found the Gamepad immediately.
theater	DVDs	I saw it in the theater and the people all around me were all complaining how boring it was.
	Electronics	My Bose home theater system could not find and play certain stations that this player can pick and play very clearly.
the cover	Books	From the cover and the blurbs I expected actual case histories.
	Kitchen	I put the cover on, close the vent, and can ensure proper doneness with juicy flavor.
value	DVDs	Low production values and poor acting.
	Electronics	This 2Gb card is the best value for money.

TABLE 28: Exemple de changement de polarité dû à un changement de sens du marqueur multi-polaire.

Les marqueurs rentrant parfois dans cette catégorie sont : returning, loved it, loved this, i loved, today. Des exemples de phrases sont présentés au tableau 30.

7.3.5 *Biais de corpus*

Enfin, il y a évidemment toujours des biais de corpus. C'est à dire que certaines répartitions préférentielles de phrases dans certaines classes sont dues au hasard de la constitution du corpus. Ces préférences n'ont dans ce cas pas d'explication et ne sont pas généralisables à d'autres textes.
Les marqueurs rentrant parfois dans cette catégorie sont : the house, easier. Nous n'avons pu observé aucune raison justifiant le biais observé au niveau du texte.

expected	DVDs	The menace of the character has gone, his appearances are too numerous and aren't scary, and the moments when he pops up to stick his hook through someone's back are completely expected.
	Electronics	As advertised and as expected.
broken	DVDs	There are hearts broken in ways you wouldn't have imagined.
	Electronics	The main unit looks broken on the backside.
no matter	Book	No matter your religion, ethnicity or basic beliefs, everyone should read this book.
	Kitchen	We think no matter which one you choose, the temperature is the same.
comfortable	Book	A couple of his plays in the book I'm not comfortable with but I haven't tried them.
	Kitchen	The butter knives are a good weight making them comfortable to use.

TABLE 29: Exemple de changement de polarité dû au fait que le marqueur multi-polaire caractérise un objet d'un type différent.

7.4 CONCLUSION

Les marqueurs multi-polaires tels que nous les avons définis rendent compte des comportements globaux au niveau du domaine. Il est très rare qu'un marqueur ait un unique usage dans un domaine. Par contre, il est fréquent qu'un usage spécifique à un domaine se rajoute par dessus les usages généraux. Si cet usage particulier est suffisamment fréquent, il influera sur les statistiques globales et le mot sera donc ainsi détecté en tant que marqueur multi-polaire par notre méthode.

Nous avons identifié les différents objets pouvant être des marqueurs ainsi que leurs comportements observés dans le corpus. S'ils n'apparaissent pas dans une phrase de contexte, les marqueurs multi-polaires peuvent exprimer directement une opinion, expliciter une opinion ou bien désigner l'objet sur lequel porte l'opinion. Nous avons ensuite établi une typologie des raisons de changement de polarité des marqueurs dans les différents domaines. Les trois causes principales sont un changement de sens, un changement d'objet ou bien un changement d'utilisation.

Dans la partie suivante, nous nous intéressons à l'effet que ces mar-

returning	DVDs	I watched it over 6 times before returning it back.
	Electronics	I ended up returning the item 2 days after i bought it.
loved it	Books	It is a beautiful book and the children absolutely loved it.
	Kitchen	When I first got this item I loved it I used it 4 times and it stopped working.
loved this	Books	I loved this sparse tale about letting go.
	Kitchen	As other reviewers, we loved this blender until it broke.
i loved	DVDs	I loved every aspect of this film.
	Electronics	I loved this player until the day I really, really needed to use it for recording.
today	DVDs	The film is still pretty fast, but would be considered slow by today's standards.
	Electronics	Today, I turn it on, it is broken !

TABLE 30: Exemple de changement de polarité dû au fait que le marqueur multi-polaire caractérise un objet qui s'utilise de manière différente.

queurs multi-polaires ont sur les performances des classifieurs automatiques d'opinion au niveau du texte.

Troisième partie

UTILISATION PRATIQUE DES MARQUEURS MULTI-POLAIRES

Dans cette partie, nous nous intéressons à l'influence pratique des marqueurs multi-polaires sur les classifieurs automatiques d'opinion au niveau du texte lorsque plusieurs domaines sont en jeu. Dans un premier temps, nous nous intéressons au problème classique d'adaptation au domaine lorsque le classifieur est entraîné sur un domaine et utilisé pour classifier des textes issus d'un autre domaine. Dans un second temps, nous verrons comment utiliser le concept des marqueurs multi-polaires lorsque l'on souhaite entraîner et utiliser un classifieur sur des corpus comportant plusieurs domaines communs. Enfin, nous explorons la problématique liée à l'absence d'étiquette de domaine dans les corpus d'entraînement et de test.

8

ADAPTATION D'UN DOMAINE À UN AUTRE

Dans les parties précédentes, nous nous sommes intéressée à la définition et à la caractérisation des mots multi-polaires. Nous présentons dans cette partie une étude sur leur influence pratique sur les classifieurs d'opinion. La première tâche que nous considérerons est celle d'adaptation d'un domaine à un autre. Il s'agit d'entraîner un classifieur sur un domaine source et de l'utiliser sur un domaine cible. Nous disposons donc d'un corpus du domaine source dont au moins une partie est annotée au niveau du texte en positif ou négatif. Nous disposons également d'un corpus dans le domaine cible. Comme nous l'avons vu au chapitre 5.3, l'objectif à terme est de pouvoir utiliser un corpus cible entièrement dépourvu d'annotation manuelle. Les résultats qui suivent sont obtenus avec des marqueurs multi-polaires calculés à partir d'une petite sous-partie du corpus cible annotée au niveau du texte.

8.1 FACILITER L'ADAPTATION D'UN DOMAINE À UN AUTRE

Lorsqu'on utilise des algorithmes d'apprentissage, on présuppose généralement que les données d'entraînement ont la même distribution que les données de test. En pratique, cela n'est pas le cas. On ne peut bien sûr pas espérer obtenir de bons résultats si les distributions des données sources et cibles diffèrent de manière trop importante. Cependant, si elles ne sont que légèrement différentes, l'apprentissage peut être efficace.

Comme nous l'avons vu au chapitre 2, il existe trois types d'approches afin de faciliter l'adaptation d'un domaine à un autre. La première est justement de réduire cet écart de distribution entre le domaine source et le domaine cible en proposant un nouveau système de représentation pour les textes [Daumé, 2007; Blitzer et al., 2006; Pan et al., 2010]. Il est également possible d'injecter de la connaissance dans les *a priori* des modèles d'apprentissages discriminatifs [Finkel and Manning, 2009; Chelba and Acero, 2006; Chan and Ng, 2006]. Enfin, l'adaptation plus précise au domaine cible peut être réalisée en attribuant des poids aux exemples [Bickel et al., 2007] ou aux traits [Satpal and Sarawagi, 2007].

La méthode que nous présentons dans ce chapitre est apparentée à la première et à la troisième catégorie puisque, une fois les marqueurs multi-polaires détectés, nous modifions la représentation des

textes, soit en supprimant certains traits, ce qui revient à leur donner un poids nul, soit en créant de nouveaux traits propres à chaque domaine. Nous exposont tout d'abord le déroulement de cette méthode d'intégration des marqueurs multi-polaires au transfert d'apprentissage avant de présenter les résultats obtenus sur des corpus en anglais et en français.

8.2 INTÉGRATION DES MARQUEURS MULTI-POLAIRES AU TRANSFERT D'APPRENTISSAGE

Nous étudions dans ce chapitre l'influence des marqueurs multi-polaires sur la classification de l'opinion lorsque le classifieur a été entraîné sur un corpus appartenant à un domaine différent de celui sur lequel il est utilisé. Nous supposons donc que nous disposons d'un corpus d'entraînement dont une partie des texte est annotée globalement en positif/négatif.

8.2.1 *Méthode d'intégration des marqueurs multi-polaires*

Après avoir détecté les marqueurs multi-polaires entre le domaine source et le domaine cible, nous nous en servons pour modifier les corpus source et cible afin de réduire leurs différences. Ainsi, un classifieur entraîné sur le corpus source modifié sera plus performant sur le corpus cible modifié que si aucune modification n'avait été effectuée. Les différentes étapes de notre méthode sont les suivantes :

DÉTECTION DES MARQUEURS MULTI-POLAIRES Pour extraire les marqueurs multi-polaires, nous utilisons les sous-parties annotées des domaines source et cible et réalisons le test du χ^2 comme décrit à la section 5.2. Cette détection utilise donc actuellement des données annotées du corpus cible, afin de démontrer l'utilité de ces marqueurs, mais l'objectif, à terme, est de réaliser cette détection de façon automatique sans utiliser d'annotation dans le domaine cible.

PARTICULARISATION DES CORPUS L'information apportée par la détection des marqueurs multi-polaires est intégrée aux corpus avant l'entraînement des classifieurs. Le corpus source comme le corpus cible sont modifiés. Deux modifications différentes, dont la description plus précise se trouve dans la section 9.5, sont testées : distinction des marqueurs entre source et cible et suppression des marqueurs.

ENTRAÎNEMENT DES CLASSIFIEURS SUR LE CORPUS SOURCE MODIFIÉ Pour la classification automatique des textes en opinion positive/négative, nous avons utilisé un algorithme de boosting : *AdaBoost* dans son implémentation *BoosTexter* [Freund et al., 1996;

Schapire and Singer, 2000] et également des séparateurs à vaste marge implémentés dans *SVMlight*.

CLASSIFICATION DU CORPUS CIBLE MODIFIÉ Pour le test, nous utilisons la totalité des textes disponibles du domaine cible (la petite sous-partie ayant servie à la détection des marqueurs multi-polaires ainsi que tous les textes de test supplémentaires).

8.2.2 *Détection des marqueurs multi-polaires*

La détection des marqueurs multipolaires entre le domaine source et le domaine cible s'effectue sur les sous-partie des corpus source et cible ayant des annotations à l'aide d'un test du χ^2 comme présenté au chapitre 5. Pour être candidat, un mot ou bigramme doit apparaître au moins 10 fois dans chacun des corpus. Au final, il est retenu si sa différence de positivité entre le domaine source et le domaine cible est d'au moins 0,1 avec une p-valeur lors du test du χ^2 de 0,005. Ces paramètres ont été sélectionnés empiriquement. Des études sur l'influence de ces paramètres sont présentées au chapitre 10.

8.2.3 *Particularisation des corpus d'apprentissage et test*

Une fois les marqueurs multi-polaires détectés, nous les utilisons afin de réduire les différences de distribution entre le domaine source et le domaine cible. Nous avons testé deux particularisations simples décrites ci-dessous :

SUPPRESSION DES MARQUEURS Dans cette particularisation, les mots ou bigrammes détectés comme marqueurs multi-polaires entre le domaine source et le domaine cible sont simplement supprimés des sacs de mots. Cela évite ainsi au classifieur d'apprendre des règles à propos de ces mots, qui amèneront nécessairement des erreurs de classement dans le domaine cible.

DIFFÉRENCIATION DES MARQUEURS Dans cette version, un mot ou bigramme *xxx* détecté comme marqueur multi-polaire n'est pas supprimé mais remplacé par "*xxx*-source" dans le corpus source et "*xxx*-cible" dans le domaine cible. Ainsi, les classifieurs les considèrent comme deux mots différents et ne propagent pas d'erreur. Par contre, cela permet au classifieur de s'entraîner avec "*mot*-source" et ainsi de ne pas donner abusivement du poids à d'autres mots potentiellement moins importants dans le domaine source.

8.3 APPORT DES MARQUEURS MULTI-POLAIRES À L'ADAPTATION AU DOMAINE

Pour tester l'apport de la détection des marqueurs multi-polaires sur la classification de l'opinion au niveau du texte, nous avons utilisé les mêmes corpus qu'à la partie 5.2.1.1. Pour l'anglais, nous disposons donc du corpus *Multi-Domain Sentiment Dataset* (MDSD) réparti en quatre domaines : *DVDs*, *kitchen*, *electronics* et *books*. Chaque domaine contient 1000 critiques positives et 1000 critiques négatives pour l'entraînement ainsi qu'un certain nombre d'autres critiques pour le test. Pour le français, nous utilisons ici un extrait du corpus issu du Défi Fouille de Textes (DEFT) 2007 contenant deux domaines : les jeux vidéos et les films. Chacun contient 420 critiques positives et 420 critiques négatives ainsi que d'autres critiques qui sont considérées sans étiquette.

8.3.1 *Résultats sur les corpus entiers*

Les figures 18 et 19 présentent les résultats obtenus en exactitude (*accuracy*) respectivement pour le français et l'anglais. Ces résultats montrent que notre méthode donne de bons résultats sur le corpus *DEFT*. Sur le corpus *MDSD*, les résultats sont plus mitigés, avec des améliorations statistiquement significatives pour la moitié des paires testées et deux cas de détérioration. Il est cependant intéressant de noter que les meilleures améliorations sont le plus souvent observées pour les paires de corpus pour lesquelles le transfert est le plus difficile (ceux dont l'exactitude du classifieur sans modification est déjà faible).

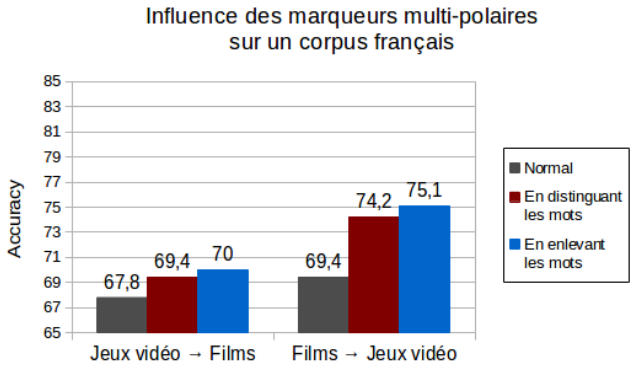


FIGURE 18: *Accuracy* pour un classifieur BoosTexter entraîné sur un domaine source et testé sur un domaine cible en français (DEFT).

52% de tous les traits sélectionnés par BoosTexter sont des bigrammes. Si parmi ces traits on ne considère que ceux détectés comme mar-

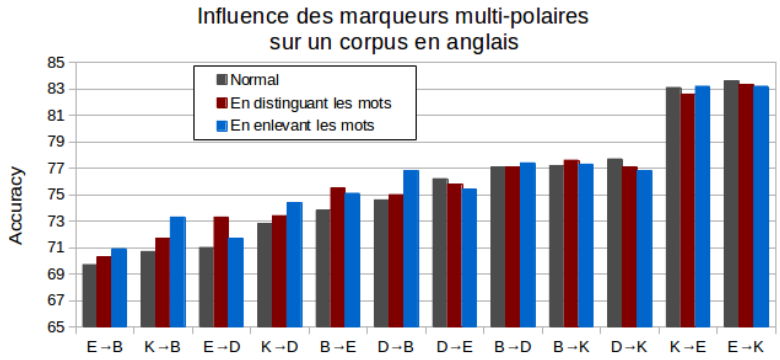


FIGURE 19: *Accuracy* pour un classifieur BoosTexter entraîné sur un domaine source et testé sur un domaine cible en anglais (MDSD) ; D : DVDs, B : books, E : electronics, K : kitchen.

queurs multi-polaires, alors seul 42% sont des bigrammes. Ainsi, en proportion, les marqueurs multi-polaires détectés sont plus souvent des unigrammes même si la part de bigrammes reste importante. Ce résultat est compréhensible puisque plus un n-gramme est long, moins il y a possibilité d'ambiguïté.

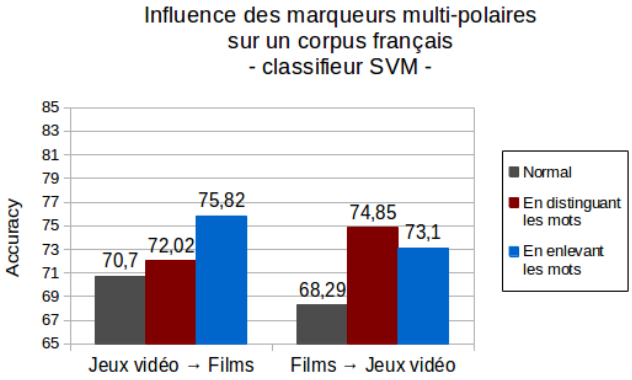


FIGURE 20: *Accuracy* pour un classifieur SVM entraîné sur un domaine source et testé sur un domaine cible en français (DEFT).

Nous avons également effectué les mêmes expériences avec un séparateur à vaste marge en utilisant l'implémentation de SvmLight (figures 20 et 21). Les comportements des deux classifieurs sont similaires. Néanmoins, les méthodes utilisant des SVM ne se comportent pas très bien lorsque la dimension des objets de test et d'apprentissage sont très différents. Or, en passant d'un domaine à l'autre la taille des textes peut fortement varier. C'est pourquoi nous avons décidé de poursuivre nos expériences avec un classifieur entraîné à partir d'un algorithme de boosting.

Ça serait bien d'avoir une citation technique.

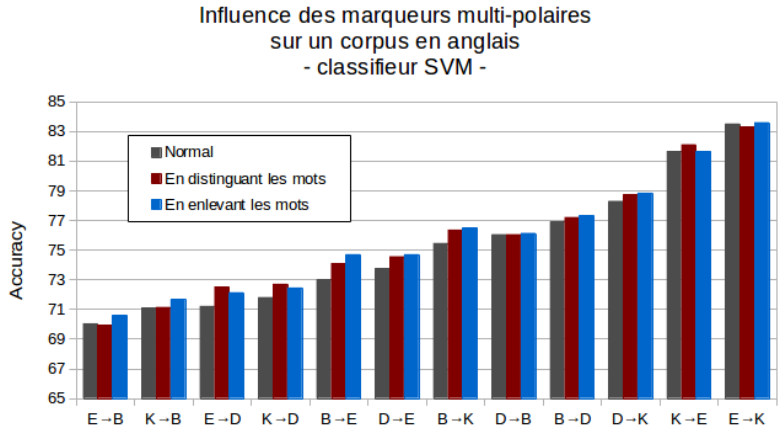


FIGURE 21: *Accuracy* pour un classifieur SVM entraîné sur un domaine source et testé sur un domaine cible en anglais (MDSD) ; D : DVDs, B : books, E : electronics, K : kitchen.

8.3.2 *Discussion des résultats*

Nos deux corpus de tests sont de nature légèrement différente. Les critiques du corpus anglais sont des retours écrits par des utilisateurs volontaires alors que celles du corpus français sont écrites par des critiques de métier et sont donc plus longues avec un vocabulaire plus varié. De plus, les formes fléchies du français sont plus nombreuses, ce qui peut avoir un impact sur la détection des marqueurs multi-polaires. En effet, bien que dans son fonctionnement théorique notre méthode, étant purement statistique, ne fasse pas intervenir des objets spécifiques à une langue, elle s'appuie néanmoins sur la segmentation en mots et les formes fléchies.
Les comportements observés sur les deux corpus vont tout de même dans le même sens. Les améliorations sur le corpus français sont en moyennes bien plus élevées que celles observées sur le corpus anglais mais le transfert sur le corpus français est initialement plus difficile. En effet, l'*accuracy* des classifieurs français sans modification est au niveau des classifieurs anglais les moins performants (entraînement sur le corpus *Books* et test sur le corpus *Kitchen*). Comme nous l'avons déjà fait remarquer, il est normal que si le transfert initial est déjà performant, le potentiel d'amélioration de la détection des marqueurs multi-polaires soit faible. Il y a en effet moins d'erreurs à corriger. Ainsi, sans que cette relation soit linéaire, si un transfert a une faible *accuracy* initiale, alors on peut espérer que la détection des marqueurs multi-polaire lui soit d'autant plus profitable.

Il est également intéressant de noter que certains sens d'adaptation donne de meilleurs résultats que d'autres. En effet, bien que pour

une même paire de domaines les marqueurs changeant de polarité soient les mêmes dans un sens ou dans l'autre, ces marqueurs ne sont pas forcément sélectionnés comme traits discriminant dans les deux sens. Il est beaucoup plus fréquent que des marqueurs polaires deviennent neutres plutôt qu'ils passent de positif à négatif. Ainsi, si un marqueur est positif pour le domaine *films*, il sera appris comme trait positif par le classifieur. Et s'il est neutre pour le domaine *jeux vidéo*, cela provoquera une erreur de transfert. Mais dans l'autre sens, en s'entraînant sur *jeux vidéo*, rien ne sera appris pour ce trait puisqu'il est neutre. Ainsi, il n'y aura pas d'erreur de transfert bien que l'on perde de l'information. Par exemple, le marqueur "brillant" a une positivité de 0,89 sur le corpus parlant des films mais de 0,50, donc neutre, sur le corpus des jeux vidéo. Le classifieur s'entraînant sur le corpus de films l'apprendra comme trait discriminant et classera faussement certaines critiques de jeux. Par contre, le classifieur s'entraînant sur le corpus de jeux vidéo n'apprendra rien à propos de "brillant" et ne contribuera pas à attribuer une polarité erronée aux critiques de films sur ce point. De façon opposée, "pauvreté est neutre pour les films (positivité de 0,50) et très négatif pour les jeux vidéo (positivité de 0,07).
Ainsi, sur le corpus français, 92 marqueurs multi-polaires sont utilisés à l'origine dans le sens *films* vers *jeux vidéo* mais seulement 55 dans le sens *jeux vidéo* vers *films*. C'est pourquoi l'intégration des marqueurs multi-polaires permet d'éviter plus d'erreurs dans un sens que dans l'autre.

8.4 CONCLUSION

En conclusion, nous avons montré dans ce chapitre que la prise en compte des marqueurs multi-polaires entre un domaine source et un domaine cible permet d'améliorer les résultats de manière notable, surtout sur le corpus français. Comme nous l'avions remarqué à la partie 5.2, on note qu'en moyenne, parmi les premiers 100 et 200 mots choisis par BoosTexter en tant que classifieurs faibles, et donc ceux avec le plus de poids pour la classification, 11 % sont détectés comme marqueurs multi-polaires par notre méthode. Les améliorations observées dans ce chapitre reflètent l'absence de propagation d'erreur par l'intermédiaire de ces marqueurs. Nos résultats montrent également une amélioration d'autant plus marquée qu'initialement la performance du classifieur en transfert était faible. Ce phénomène s'explique par le fait que, si initialement le classifieur entraîné sur le domaine source donne de piètres performances sur le domaine cible, cela est dû entre autres à la présence de nombreux marqueurs multi-polaires. Si le transfert se passe sans anicroches, cela indique une faible utilisation par les classifieurs de marqueurs multi-polaires. Ainsi, en plus d'être intéressants d'un point de vue linguistique, les

marqueurs multi-polaires que nous détectons le sont pour l'adaptation au domaine lors de classification de l'opinion au niveau du texte.

9

CLASSIFICATION D'OPINION SUR UN CORPUS MULTI-DOMAINE

Dans le chapitre précédent, nous avons montré que la prise en compte, même minimale, des marqueurs multi-polaires permettait un meilleur transfert des classifieurs d'opinion d'un domaine à un autre. Cependant, les marqueurs multi-polaires peuvent également se présenter dans des corpus généraux qui contiennent des textes faisant référence à plusieurs domaines. En effet, si le corpus contient plusieurs sous-domaines, et qu'un mot a une connotation positive dans tous les sous-domaines sauf un, les classifieurs apprendront ce mot comme trait positif et feront une erreur de classification dans le dernier sous-domaine.
Dans ce chapitre, nous nous intéressons au cas où l'on dispose d'un corpus annoté au niveau du texte en positif/négatif comprenant des textes de plusieurs domaines disposant d'étiquettes de domaines. Les textes du corpus de test présentent la même répartition de domaines et disposent également d'étiquettes. C'est par exemple le cas lorsque l'on collecte les textes d'un blog abordant plusieurs sujets annotés par un système d'étiquettes thématiques (*tags*). Il est possible de vouloir créer un classifieur pour un seul des thèmes abordés mais de ne pas disposer de suffisamment de textes annotés de ce domaine particulier pour pouvoir construire directement un classifieur lui correspondant. Il est alors intéressant de prendre en compte tous les textes annotés disponibles, mais s'il n'est pas prêté attention aux marqueurs multi-polaires à l'intérieur même du corpus d'apprentissage, ils risquent d'induire le classifieur d'opinion en erreur.

9.1 UTILISER PLUSIEURS DOMAINES

Pour pouvoir utiliser des classifieurs fondés uniquement sur des méthodes d'apprentissage statistique tout en étant les plus généraux possible, il est recommandé de disposer de données d'apprentissage venant du plus grand nombre de domaines possible. En effet, si l'on dispose d'un peu de données annotées dans plusieurs domaines, il est utile d'utiliser les informations contenues dans un domaine pour aider l'apprentissage sur les autres domaines, et réciproquement. C'est ce qu'on appelle de l'apprentissage multitâches. Dans ce cadre, fusionner les classifieurs fonctionne mieux que fusionner directement les données d'apprentissage [Li and Zong, 2008; Li et al., 2011]. Dans

ces travaux, la fusion la plus efficace est réalisée par la somme pondérée des résultats des différents classifieurs, les poids de cette somme étant appris sur un petit corpus de développement du domaine cible. Cette approche donne un classifieur obtenant de bons résultats sur plusieurs domaines si l'on dispose d'un peu de données annotées pour tous. Néanmoins, il est impossible de garantir des résultats pour des domaines complètement nouveaux.

Dans Yoshida et al. [2011], les auteurs étudient l'influence du nombre de domaines source et cible, allant jusqu'à quatorze domaines différents. Plus le nombre de corpus source est élevé, plus les résultats sur un corpus cible différent sont bons. De plus, leur modèle probabiliste génératif permet de déterminer si la polarité inférée pour un mot dépend ou non du domaine du texte où se trouve le mot. Ainsi, ils construisent automatiquement des dictionnaires valués pour chaque domaine.
Les auteurs de Mansour et al. [2013] montrent également que pour les langues disposant déjà de nombreux corpus libres annotés au niveau du texte selon l'opinion de l'auteur, s'entraîner sur l'intégralité des ressources disponibles permet de créer un classifieur robuste aux nouveaux domaines. Le problème de transfert continue néanmoins à se poser pour les langues moins dotées ou bien pour la création de ressources lexicales. Les méthodes que nous développons, bien que testées sur l'anglais et le français, visent à être utilisables pour un grand nombre de langues différentes.

Un problème peut également se poser lorsque les corpus sont hétérogènes et couvrent plusieurs domaines. Dans le domaine de la classification d'image, Hoffman et al. [2011] s'attaquent au problème de plusieurs domaines sources dont on ne connaît pas *a priori* les étiquettes. Ils séparent d'abord les domaines sources à l'aide d'une variante de l'algorithme des *k-means* avant de poursuivre plus classiquement en combinant les classifieurs appris sur les domaines ainsi séparés. À notre connaissance, il n'y a pas de travaux en classification d'opinion traitant ce problème particulier. Nous étudions ce problème dans le cadre de la classification de l'opinion au chapitre 10.

9.2 PRÉSENTATION DE LA MÉTHODE

Nous étudions donc dans ce chapitre comment la prise en compte des marqueurs multi-polaires intra-corpus peut influer sur la classification de l'opinion. Il faut tout d'abord détecter ces marqueurs entre les sous-domaines du corpus d'entraînement avant de les utiliser afin de particulariser le corpus d'entraînement. Nous entraînons ainsi plusieurs classifieurs, chacun particularisé pour un sous-domaine particulier. Le déroulement de notre méthode est le suivant :

DÉTECTION DES MARQUEURS MULTI-POLAIRES INTRA-CORPUS Le corpus d'entraînement est séparé en sous-corpus thématiques. La détection des marqueurs multi-polaires s'effectue entre un sous-domaine et tous les autres. Nous obtenons donc plusieurs collections de marqueurs multi-polaires, une pour chaque sous-domaine.

PARTICULARISATION DU CORPUS D'ENTRAÎNEMENT Pour chaque sous-domaine, est créée une version légèrement modifiée du corpus d'entraînement en prenant en compte la collection de marqueurs multi-polaires correspondante.

ENTRAÎNEMENT DE PLUSIEURS CLASSIFIEURS Un classifieur est entraîné pour chaque sous-domaine sur le corpus d'entraînement particularisé adapté.

CLASSIFICATION DE NOUVEAUX TEXTES Pour classer un nouveau texte, on utilise le classifieur correspondant à son domaine.

9.3 CORPUS UTILISÉS

Pour cette expérience, nous avons besoin d'un corpus contenant plusieurs domaines. Aussi, pour le français, avons nous utilisé une partition différente du corpus issu de la campagne d'évaluation DEFT 2007. Nous avons effectué notre évaluation sur l'intégralité du corpus *AvoirAlire* de DEFT que nous avons manuellement séparé en fonction des objets abordés. Il y a donc 5 domaines, présentés au tableau 31, contenant trois classes non équilibrées (55 % de textes positifs, 30 % de neutres et 15 % de négatifs).

Nom du domaine	Nombre de textes
Livres	757
Bandes dessinées	387
Films	1623
Musique	343
Théâtre	289

TABLE 31: Composition du corpus français utilisé après séparation thématique manuelle des critiques issues du corpus *AvoirAlire* de DEFT 2007.

Pour l'anglais, nous avons de nouveau utilisé le corpus MDSD (8000 critiques annotées en positif/négatif réparties en quatre domaines : *DVDs*, *books*, *electronics* et *kitchen*).

Pour chaque corpus, nous avons réalisé une validation croisée. Le corpus est séparé aléatoirement en dix parties, neuf d'entre elles servant successivement de corpus d'entraînement et la dixième de cor-

pus de test. Les résultats présentés sont les résultats moyens des dix expériences. Les textes sont toujours représentés en sacs de mots des unigrammes et bigrammes des formes fléchies.
La métrique d'évaluation utilisée lors de cette expérience est la F-mesure moyenne des classes positives et négatives. Cette mesure est en effet celle utilisée lors de la campagne d'évaluation SemEval 2013 à laquelle nous avons participé. Notre contribution à cette campagne est détaillée au chapitre 10.

9.4 DÉTECTION DES MARQUEURS

Le processus de détection des marqueurs multi-polaires pour un corpus multi-domaine est présenté dans la figure 22. Le corpus d'entraînement est séparé en plusieurs sous-parties, chacune correspondant à un domaine particulier. Pour détecter les marqueurs multi-polaires, nous utilisons les étiquettes positives et négatives des données d'entraînement, comme décrit dans la section 5. Les paramètres de sélection sont réglés de façon à maximiser le nombre de marqueurs multi-polaires détectés tout en gardant un risque de faux positif à 1 %. Pour être candidat, un mot ou groupe de mots doit apparaître au moins 5 fois dans la sous-partie du domaine concerné. Finalement, il est retenu si sa différence de positivité entre le sous-corpus et son complémentaire est au moins 0,1 mais avec une p-valeur de 0,001. Ces paramètres sont étudiés plus en détail au chapitre 10

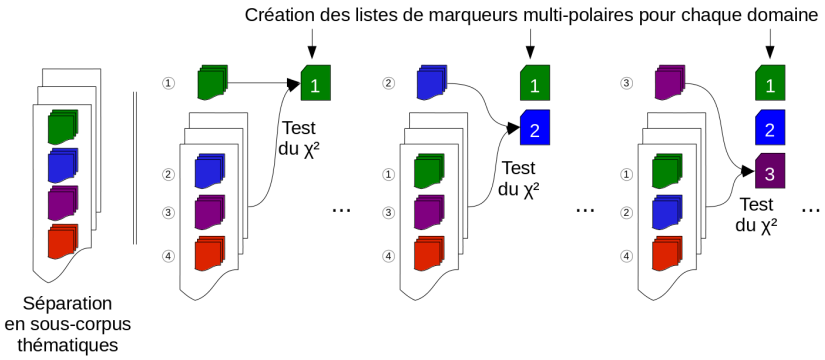


FIGURE 22: Détection des marqueurs multi-polaires entre les sous-parties thématiques du corpus d'entraînement.

Nous effectuons cette détection pour chaque sous-partie. À chaque fois, nous détectons les mots qui changent de polarité entre une sous-partie particulière du corpus d'entraînement et son complément (tous les autres textes). À la fin de cette procédure, nous avons plusieurs collections de marqueurs multi-polaires (une collection différente pour chaque sous-partie).

Il serait envisageable de détecter les marqueurs multi-polaires entre deux sous-parties, c'est à dire en *one-vs-one* plutôt qu'en *one-vs-all*. Les particularisations décrites au paragraphe suivant devraient alors être légèrement modifiées.
Cette approche serait peut-être intéressante pour les versions de suppression ou de différenciation partielle. En effet, si un mot est détecté comme marqueur multi-polaire entre le sous-domaine 1 et uniquement le sous-domaine 2, il peut être intéressant de ne supprimer ou particulariser ce mot que dans le sous-domaine 2. Ainsi, les occurrences de ce mot présentes dans les autres sous-domaines contribuent à entraîner sans erreurs le classifieur spécifique au sous-domaine 1. En détection *one-vs-all*, dans le cas de notre exemple, le mot considéré ne serait vraisemblablement pas détecté en tant que marqueur multi-polaire entre le sous-domaine 1 et tous les autres. Il ne serait donc ni supprimé, ni différencié et son comportement différent dans le sous-domaine 2 devrait être contrebalancé par les autres domaines et donc ne pas entraîner d'erreurs. Ainsi, il n'est pas clair que la détection *one-vs-one* apporte de franches améliorations pour les particularisations de type suppression ou différenciation partielle.
De plus, pour la particularisation de type suppression totale, qui est celle donnant en moyenne les meilleurs résultats, disposer de plusieurs listes à la place d'une seule lors de la particularisation du corpus pour un sous-domaine particulier ne changera pas fondamentalement les résultats puisque tous les mots sont retirés de tous les sous-corpus.
La détection *one-vs-all* étant plus facile à gérer que celle en *one-vs-one*, car créant moins de listes de marqueurs multi-polaires, c'est cette dernière qui a été retenue.

9.5 CORPUS D'ENTRAÎNEMENT PARTICULARISÉS

Nous créons un corpus d'entraînement différent pour chaque domaine par modification du corpus original en utilisant la liste de marqueurs multi-polaires associée à ce domaine. Au chapitre 8, le corpus d'entraînement n'était pas séparé en sous-corpus. Aussi, les seules solutions de particularisation étaient la suppression ou la différentiation des marqueurs sur l'intégralité du corpus d'entraînement. Dans le cas qui nous occupe ici, il est par contre possible d'effectuer les modifications uniquement sur des sous parties. Les différentes particularisations possibles sont donc en définitive les suivantes :

SPÉCIFIQUE AU DOMAINE Le corpus d'entraînement est séparé en ses sous-corpus non modifiés et un apprentissage différent est fait pour chaque sous-corpus. Le problème de cette approche est que les nouveaux corpus d'entraînement sont de taille trop restreinte.

SUPPRESSION TOTALE Pour chaque domaine, les marqueurs multi-polaires détectés sont supprimés de l'intégralité du corpus. Il y a donc un corpus légèrement différent par domaine, de la même taille que le corpus d'entraînement original (figure 23).

SUPPRESSION PARTIELLE Pour chaque domaine, les marqueurs multi-polaires détectés sont supprimés du corpus sauf de la sous-partie associée à ce domaine. En effet, les marqueurs pourraient être porteurs d'un indication polaire importante dans cette sous-partie. Il y a donc un corpus légèrement différent par domaine, de la même taille que le corpus d'entraînement original (figure 23).

DIFFÉRENCIATION PARTIELLE Pour chaque domaine, les marqueurs multi-polaires détectés sont différenciés : *marqueur-domainX* dans la sous partie associée à ce domaine et *marqueur-autre* dans le reste du corpus. Il y a donc un corpus légèrement différent par domaine, de la même taille que le corpus d'entraînement original (figure 23).

DIFFÉRENCIATION CONCOMITANTE La différenciation des marqueurs de toutes les listes se fait en même temps. Il n'y a donc au final qu'un seul corpus d'entraînement modifié avec cette version (figure 24).

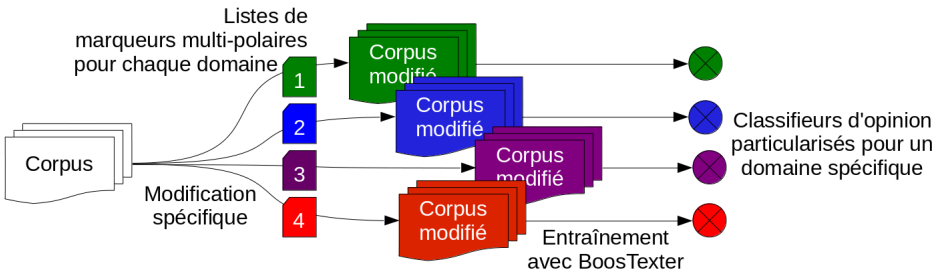


FIGURE 23: Processus de création de plusieurs classifieurs thématiques en particularisant le corpus d'entraînement. Version suppression totale, suppression partielle et différenciation partielle.

Pour cette expérience, nous avons testé uniquement la suppression totale des marqueurs multi-polaires. En effet, cette modification donne globalement de meilleurs résultats dans les expériences menées sur le corpus de tweets qui sont présentés dans le chapitre 10 et qui proposent une évaluation comparative plus détaillée de ces différentes approches.

Nous entraînons ensuite un classifieur sur ce corpus modifié et obtenons ainsi un classifieur spécifiquement adapté pour chaque domaine.

Nous obtenons ainsi plusieurs classifieurs différents, chacun particularisé pour un domaine particulier (figure 23). Un texte du corpus

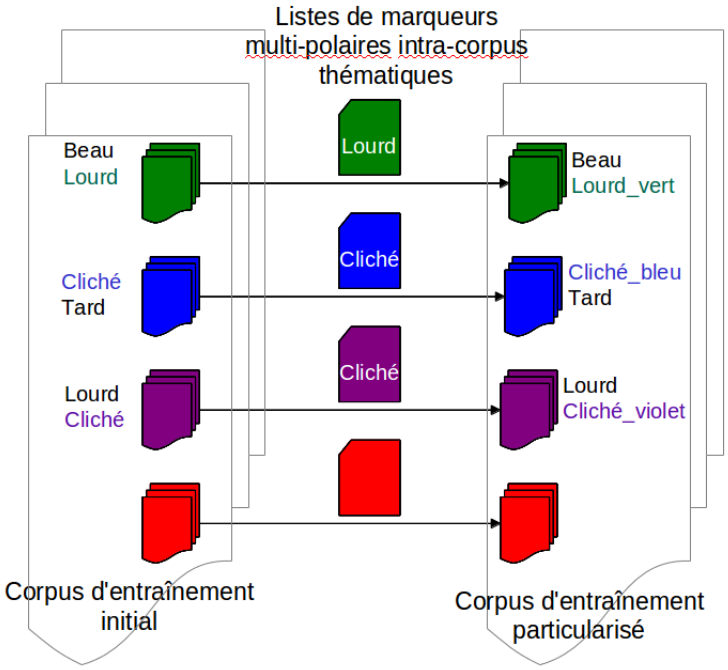


FIGURE 24: Processus de particularisation du corpus d'entraînement pour la version différenciation concomitante. Chaque mot apparaissant dans une sous-partie thématique du corpus et détecté comme marqueur multi-polaire entre cette sous-partie et les autres est modifié en "mot-domaine". Si le mot n'est pas détecté comme marqueur multi-polaire pour la sous-partie dans laquelle il se trouve, il n'est pas modifié, même s'il est détecté comme marqueur multi-polaire pour une autre sous-partie.

de test est ensuite classifié en utilisant le classifieur propre à son domaine.

9.6 PRÉSENTATION DES RÉSULTATS

Les résultats, présentés à la figure 25, montrent que la prise en compte des marqueurs multi-polaires dans le cas d'un corpus contenant plusieurs domaines amène des améliorations plus faibles que celles obtenues lors de l'adaptation d'un domaine à un autre (cf. section 8). Ces améliorations demeurent néanmoins significatives pour le français.

9.7 CONCLUSION

Dans ce chapitre, nous avons étudié la situation où les corpus d'entraînement et de test ont la même répartition de domaines. Notre méthode permet d'éviter des erreurs lorsqu'un mot a une certaine polarité dans tous les domaines sauf dans un, où il a une polarité

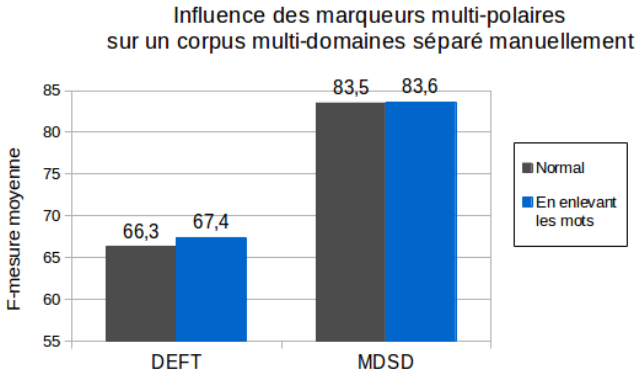


FIGURE 25: Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus multi-domaines français (DEFT) et anglais (MDSD).

différente. Un apprentissage global assignera à ce mot la polarité dominante. Les erreurs ne se présenteront que dans la sous-partie du corpus de test associée avec le domaine isolé alors que, pour les autres parties du corpus de test, il n'y aura pas d'erreur. Les erreurs que l'on peut éviter avec notre méthode sont donc moins nombreuses dans ce cas que lors de l'adaptation d'un domaine à un autre présentée dans la partie 8. Néanmoins, différencier les marqueurs multi-polaires n'est pas très difficile à mettre en place et se conjugue aisément avec les autres méthodes de classification de l'opinion en leur permettant d'éviter un certain nombre d'erreurs. De plus, dans une perspective de création de ressources, la détection des marqueurs multi-polaires intra-corpus peut potentiellement amener une amélioration qualitative.

Au chapitre suivant, nous nous intéresserons au cas où l'appartenance des textes à un domaine particulier n'est pas explicite.

10 CLASSIFICATION D'OPINION SUR UN CORPUS EN DOMAINE OUVERT

Dans le chapitre précédent, nous avons fait l'hypothèse que la répartition des textes en différents domaines était connue. Or, ce n'est pas forcément le cas : certaines collections de textes contiennent des documents de différents domaines sans séparation ni indication explicite des domaines couverts. C'est par exemple le cas de corpus collectés automatiquement sur des médias particuliers, comme Twitter, qui présentent pourtant en général un grand intérêt pour des systèmes de veille d'opinion.

10.1 DIFFÉRENCE AVEC LE CAS DU MULTI-DOMAINES

Le contexte applicatif des domaines ouverts est très semblable à celui du multi-domaine. La seule différence par rapport au chapitre 9 est en effet l'absence d'étiquette de domaine pour les textes des corpus. Ainsi, s'il est possible de séparer automatiquement les textes disponibles en différentes thématiques, le problème est ramené au cas précédent. La détection des marqueurs multi-polaires et leur intégration afin d'entraîner plusieurs classifieurs peuvent alors être effectués selon la méthode présentée dans la section 9. Néanmoins, dans le cas des domaines ouverts, l'appartenance d'un texte à un domaine n'est pas une information binaire : chaque texte possède en général un poids d'association avec les différents domaines. Pour chaque texte, les résultats des différents classifieurs spécifiques aux domaines doivent donc être fusionnés en tenant compte de ces poids pour obtenir la classification finale de l'opinion.
Comme nous l'avons évoqué au chapitre 9, le problème des corpus multi-sources dont on ne connaît pas *a priori* les étiquettes a été traité dans le domaine de la classification d'image par Hoffman et al. [2011]. Les différents domaines du corpus sont séparés à l'aide d'une variante de l'algorithme des *k-means* avant de combiner les classifieurs appris sur les domaines ainsi séparés. Cependant, à notre connaissance, ce problème n'a pas été abordé dans des travaux traitant de classification d'opinion.

Afin de pouvoir comparer les résultats en domaine ouvert avec ceux obtenus au chapitre précédent avec une séparation en domaines explicite, nous avons, dans un premier temps, utilisé les mêmes corpus français (DEFT) et anglais (MDSD) sans tenir compte des éti-

quettes de domaine. De façon complémentaire, nous avons également utilisé le corpus anglais de tweets issu de la campagne d'évaluation SemEval 2013 pour la tâche 2 d'annotation de l'opinion [Wilson et al., 2013]. Ce dernier corpus est représentatif d'une collection de documents en domaine ouvert et permet de varier le type de textes sur lequel appliquer notre méthode.

10.2 GÉNÉRATION DE DOMAINES

Comme le corpus initial n'a pas d'étiquette de domaine, nous devons tout d'abord identifier les domaines sous-jacents et assigner chaque texte à un domaine. Nous avons utilisé dans ce but une méthode automatique de détection de thèmes (*Topic Models*), à savoir la méthode d'allocation de Dirichlet latente (LDA) [Blei et al., 2003].
Dans le cadre de la détection d'opinion, la méthode LDA a déjà été utilisée pour l'analyse de critiques focalisées sur un aspect, qui est proche de notre travail : dans Titov and McDonald [2008a,b], les auteurs introduisent un modèle fusionnant des thèmes locaux et globaux et utilisent les annotations manuelles des critiques afin d'améliorer l'identification des différents thèmes. D'autres travaux, tels que Zhang et al. [2013]; Li et al. [2010], combinent au modèle LDA des informations de sentiment ou bien des techniques de Naives Bayes afin de sortir du modèle en sac de mots.

10.2.1 *Séparation des corpus à l'aide de l'allocation de Dirichlet latente*

Pour notre expérience, nous avons utilisé l'implémentation de la méthode LDA proposée dans Mallet [McCallum, 2002], qui utilise la méthode d'échantillonnage de Gibbs afin d'inférer la distribution utilisée pour la création des modèles de thèmes.

Comme pour la classification, les mots des textes sont au préalable passés en minuscules et certains mots, présents sur une liste de mots vides (*stopwords*), sont retirés. L'allocation de Dirichlet latente est ensuite appliquée à ces textes.
Sans adaptation du poids relatif des thèmes, on obtient sur le corpus *MDSD* -8.95 de perplexité contre -8.70 avec une évaluation des poids des différents thèmes à chaque pas d'optimisation. La perplexité mesure le degré d'adaptation du modèle aux données par une expression normalisée de la log-vraisemblance du corpus. Il s'agit ici d'une auto-évaluation sur le corpus d'apprentissage lui-même. La perplexité est toujours négative et une valeur proche de 0 est le signe d'un bon modèle. Aussi avons nous gardé l'optimisation des poids à chaque pas pour la suite des expérimentations.

Il est nécessaire d'indiquer le nombre de thèmes désirés pour que la

méthode LDA puisse fonctionner. Il existe des extensions de cette méthode qui essaient de déterminer automatiquement le nombre de domaines sous-jacents, ainsi que d'autres méthodes spécialement conçues pour la catégorisation de textes de petite taille comme les tweets [Shrestha et al., 2012]. Tester l'efficacité de ces méthodes serait une prolongation intéressante de nos travaux. En l'occurrence, nous avons indiqué le nombre exact de domaines pour les corpus *MDSD* et *DEFT*, à savoir, respectivement 4 et 5. Pour le corpus de tweets, nous avons effectué plusieurs essais et conservé celui donnant de meilleurs résultats en terme de F-mesure moyenne.

Après avoir déterminé les thèmes à l'aide du corpus d'entraînement, chaque texte est représenté par un vecteur dont la taille est le nombre de thèmes, et dont chaque composante est la proportion de mots du texte qui appartient au thème associé à la dimension correspondante. Le corpus d'entraînement est ensuite séparé en sous-parties, ou domaines, chacun d'entre eux associé avec l'un des thèmes sous-jacents détectés. Un texte est simplement associé au thème avec lequel il a le plus d'affinité. Par exemple, si sa proportion de mots appartenant à un thème est 55 %, il fera partie de la sous-partie du corpus associée au domaine correspondant.

10.2.2 *Adéquation entre la séparation manuelle et la séparation par LDA*

Dans cette section, nous allons comparer la séparation obtenue automatiquement par LDA sur les corpus *DEFT* et *MDSD* avec celle obtenue manuellement. L'allocation de Dirichlet latente comprend une initialisation aléatoire aussi et les thèmes obtenus sont, par conséquent, légèrement différents à chaque fois. Les chiffres présentés dans cette section sont obtenus sur une réalisation particulière mais sont représentatifs des thèmes que l'on peut obtenir sur nos corpus de travail.
Le modèle LDA est créé à partir de la partie de ces corpus destinées à l'entraînement. C'est à dire 8000 textes pour *MDSD*, équilibrés au niveau des classes positive et négative et initialement répartis en quatre domaines. Le corpus d'entraînement issu de *DEFT* contient initialement 5 domaines de respectivement 669, 354, 1454, 314 et 260 textes répartis de manière non équilibré entre les trois classes positive, négative et neutre. Le nombre exact de domaines a été précisé à LDA à chaque fois.

Le tableau 32 montre les mots les plus représentatifs des thèmes trouvés. Pour le corpus *MDSD*, on reconnaît aisément les livres, l'électronique, les films ainsi que les ustensiles de cuisine. Cela est confirmé par la matrice de confusion (tableau 33). Pour le corpus *DEFT*, on peut également associer les thèmes créés avec les domaines de la musique, des livres, des films, du théâtre et des bandes dessinées

MDSD	Thème 0	book read books num people reading time life author world good quot work great written found make don't story
	Thème 1	num good great product quality sound work bought don't buy time back price works unit problem phone amazon
	Thème 2	movie film num good story time great dvd don't characters love watch make plot character people movies bad year
	Thème 3	num great time good don't coffee easy pan love bought buy product set water make clean made years
DEFT	Thème 0	album rock groupe musique bien voix disque pop premier morceaux années ainsi électro déjà leurs titres scène toujours hop
	Thème 1	vie roman histoire ans monde homme jamais bien jeune amour père mère femme passe toute va temps leurs celui
	Thème 2	film argument avis cinéma bien cinéaste réalisateur scène personnages films œuvre spectateur grand métrage années histoire scenes genre acteurs
	Thème 3	bien scène vie histoire théâtre temps femme comédie fran toujours monde amour dire pièce mise jeune jean petit rien
	Thème 4	bien histoire série album dessin personnages premier temps monde grand auteur héros aventures tome leurs petit toujours toute bande

TABLE 32: Mots les plus représentatifs pour les 5 thèmes trouvés par LDA sur les corpus mixés de *MDSD* (haut) et *DEFT* (bas).

bien que le thème 2 n'ait guère que le mot "roman" qui dénote clairement le monde des livres. On remarque cependant que certains mots se retrouvent dans plusieurs thèmes. "histoire" en particulier se retrouve dans les thèmes 1, 2, 3 et 4. "scène" appartient aux thèmes 0, 2 et 3, "personnages" se retrouve dans le 2 et le 3. En effet, sur les 5 domaines du corpus français, 4 sont liés au fait de raconter des histoires avec des personnages. On remarque au passage que, quelle que soit la langue, les histoires d'amour ont beaucoup de succès ("love", "amour"). Ces domaines se ressemblent plus comme l'atteste la matrice de confusion : le domaine de la musique est particulièrement bien distingué par le découpage en thèmes alors que les autres domaines se mélangent plus. On remarque également que les textes parlant de théâtre ou de bandes dessinées se retrouvent presqu'intégralement dans un seul thème, certainement en raison d'un vocabulaire plus technique. Par contre, les livres et les films, bien qu'ayant un thème privilégié, sont plus répartis. En effet, la méthode LDA a tendance à créer des thèmes de même taille environ. Or, pour le corpus français, les domaines sont de tailles très variables. Les livres et les films, en plus grande quantité, se trouvent répartis dans les différentes catégories. Il est cependant intéressant de remarquer que cette répartition ne semble pas se faire au hasard. En effet, les livres se retrouvent plutôt avec les bandes dessinées et l'on peut aisément imaginer des commentaires sur les illustrations ou la couverture sans parler des romans graphiques. Les films, quant à eux, sont mal classés dans les catégories rassemblant des livres, la description de l'intrigue devant être assez similaire, et des pièces de théâtre, sans doute dans ce dernier cas en raison de commentaires sur les jeux d'acteurs. Néanmoins, il serait intéressant de poursuivre nos travaux en testant des méthodes de séparation en thèmes qui ne soient pas sensibles à ce déséquilibrage des classes.

Nous avons également regardé l'adéquation des thèmes détectés à la réalité en ne prenant en compte que les critiques qui sont attribuées à plus de 75 % à un thème. Pour chaque corpus et chaque thème, cela a pour effet d'augmenter la précision et de diminuer le rappel. Pour le corpus anglais, l'*accuracy* globale est de 86,70 % si l'on attribue un thème à toutes les critiques. Si on n'attribue un thème qu'aux critiques dépassant 75 %, l'*accuracy* est de 74,55 % mais de 93,32 % si on ne prend en compte que les critiques attribuées. Pour le corpus français, plus difficile à traiter comme nous l'avons vu, l'*accurracy* globale est de 69,85 % avec toutes les critiques et de 87,34 % si on ne prend en compte que les critiques attribuées à plus de 75 %, beaucoup n'étant dans ce cas pas classées.

Au final, la méthode LDA retrouve assez bien les différents domaines à travers son profil de thèmes. Néanmoins elle comporte deux

MDSD

	Books	DVDs	Electronics	Kitchen
Thème 0	**1433**	127	3	6
Thème 2	423	**1746**	10	7
Thème 1	53	107	**1931**	161
Thème 3	91	20	56	**1826**

DEFT

	Musique	Livres	Films	Théâtre	BDs
Thème 0	**306**	9	17	1	3
Thème 2	1	**449**	279	9	30
Thème 1	0	4	**826**	0	0
Thème 3	4	78	233	**244**	15
Thème 4	3	129	99	6	**306**

TABLE 33: Table de confusion entre les thèmes détectés par LDA et les séparation manuelles en domaines. Corpus *MDSD* en haut et *DEFT* en bas.

MDSD

	Books	DVDs	Electronics	Kitchen
Thème 0	**1010**	46	1	3
Thème 2	158	**1446**	5	3
Thème 1	11	44	**1842**	99
Thème 3	27	5	25	**1666**
Non attribué	794	459	127	229

DEFT

	Musique	Livres	Films	Théâtre	BDs
Thème 0	**284**	0	1	0	0
Thème 2	0	**309**	89	0	4
Thème 1	0	1	**394**	0	0
Thème 3	1	16	30	**159**	2
Thème 4	1	43	5	0	**185**
Non attribué	28	300	935	101	163

TABLE 34: Table de confusion lorsque l'attribution d'une critique à un domaine est d'au moins 75%. Corpus *MDSD* en haut et *DEFT* en bas.

limitations qu'il serait intéressant d'étudier : l'indication par l'utilisateur du nombre de thèmes ainsi que la création de thèmes nécessairement équilibrés. L'efficacité de cette séparation en domaines via les thèmes LDA pour la détection et la prise en compte des marqueurs multi-polaires est présentée à la section des résultats (10.5).

10.2.3 *Génération de domaines sur un corpus de tweets*

En plus des corpus *MDSD* et *DEFT*, utilisés sans prendre en compte les étiquettes de domaine, nous avons voulu tester notre méthode sur un corpus plus représentatif des sources en domaine ouvert que l'on peut trouver sur internet, en l'occurrence un corpus de tweets. Ce corpus est celui utilisé lors de la tâche 2B de la campagne d'évaluation SemEval 2013 à laquelle nous avons participé. Le résultat de notre participation à cette campagne est décrit plus en détail à la section 10.6.

Pour ce corpus, des pré-traitements spécifiques ont été appliqués. Tout d'abord, les liens ou adresses mails ont été retirés, la proportion de liens ayant un titre explicite et exploitable étant trop limitée par rapport au bruit généré. Les émoticônes sont repérés par expressions régulières en se fondant sur la liste des émoticônes occidentaux répertoriés sur Wikipédia. Ils sont extraits des tweets et le nombre d'occurrences de chaque type est ajouté en tant que trait additionnel pour l'apprentissage. Le tableau 35 présente quelques types d'émoticônes utilisés. Le corpus contenait initialement 10402 tweets répartis en 3 classes (3855 positifs, 4914 objectifs ou neutres et 1633 négatifs). Nous avons sélectionné au hasard une sous-partie équilibrée de ce corpus sur laquelle nous avons travaillé (1633 de chaque classe). Enfin, pour ce corpus uniquement, nous avons lemmatisé les mots du texte en utilisant l'analyseur linguistique LIMA [Besançon et al., 2010]. En effet, les tweets étant de très courts textes, les formes fléchies ont peu d'occurrences. Comme pour les autres corpus, nous utilisons un sac de mots des unigrammes et bigrammes.

Pour ce corpus, nous ne connaissons pas *a priori* le nombre de domaines sous-jacents. Nous avons donc effectué des tests avec plusieurs nombres de thèmes : 5, 10, 15 et 20. Il se trouve que les résultats se sont montrés plus probant avec 5 thèmes. Ce sont donc ces résultats qui sont présentés plus bas. Le tableau 36 présente les mots les plus fréquents relatifs à chaque thème que nous avons nommés pour plus de lisibilité.

:-) :) =) X) x)	Sourire
:-(:(=(	Tristesse
:-D :D =D X-D XD x-D xD :')	Rire
;-) ;)	Clin d'œil
< 3	Cœur
:'-(:'(='(	Larme

TABLE 35: Quelques types classiques d'émoticônes.

Thème	mots représentatifs
Spectacle	tonight watch time today good love make saturday friday sunday
Obama	win vote sunday award obama black make fight today saturday
Sport	game play tonight win team today sunday season saturday start
Informatique	time oct nov news monday thursday apple international sun anderson
Show	friday ticket show saturday open october center club live november

TABLE 36: Description des thèmes obtenus sur le corpus de tweets.

10.3 DÉTECTION DES MARQUEURS MULTI-POLAIRES ET PARTICULARISATION

Une fois les corpus partitionnés en domaines via le modèle de thèmes produit par LDA, la détection des marqueurs multi-polaires s'effectue de la même façon qu'à la partie 9. Pour les corpus DEFT et MDSD, la particularisation se passe également comme décrit à la partie 9, par suppression totale des marqueurs multi-polaires détectés pour chaque domaine issu du modèle de thèmes LDA. Pour le corpus de tweets, qui a par essence une structure différente des deux autres corpus, nous avons testé l'intégralité des particularisations présentées à la partie 9.5. Au final, la version donnant les meilleurs résultats est la version de suppression totale.

10.4 UNE CLASSIFICATION PAR FUSION

La seconde différence entre le cas multi-domaines et domaine ouvert se situe lors de la classification des nouveaux textes. En effet, les textes du corpus de test n'ont pas d'étiquette de domaine. Nous devons tout d'abord déterminer leur profil de thèmes en utilisant le

modèle de thèmes de la LDA. Ensuite, nous appliquons tous les classifieurs sur les nouveaux textes et obtenons plusieurs réponses différentes, une pour chaque classifieur spécifique à un sous-domaine. Nous fusionnons ces réponses en utilisant comme pondération les poids de leur profil de thèmes. Nous avons testé différentes stratégies de fusion, décrites ci-dessous.

Nous désignons par $\mathcal{T}$ l'ensemble des thèmes détectés par LDA. Un thème particulier, et par extension le domaine qui lui est associé, est noté t. t_{max} désigne le thème avec lequel un texte donné a le plus d'affinités selon le modèle de thèmes issu de la LDA. Le rôle du classifieur est de classer un texte dans une classe notée c. Pour ce faire, il calcule plusieurs scores pour un texte donné que nous définissons de la manière suivante :

- $score_t^c$: score donné par le classifieur entraîné pour le domaine t pour la classe c.
- $score_{ini}^c$: score obtenu pour la classe c par le classifieur entraîné sur l'intégralité du corpus initial sans changement.

Nous définissons également $pond_t$ comme étant la pondération dérivée du modèle LDA indiquant le degré d'association du texte avec le thème t. Enfin, $NBtexte_t$ désigne le nombre de textes servant à l'entraînement du classifieur sur le domaine t et $NBtexte_{max}$ le nombre maximal de textes utilisés pour entraîner un des classifieurs, tout domaine confondu.

Les différentes stratégies de fusion que nous avons testées sont donc les suivantes :

CHOISIR LE CLASSIFIEUR LE PLUS ADAPTÉ Seul les scores du classifieur entraîné sur le domaine majoritairement relié au texte à classer est utilisé.

$$score_1^c = score_{t_{max}}^c$$

COMBINER LES SCORES DES CLASSIFIEURS THÉMATIQUES Tous les classifieurs thématiques sont utilisés sur le texte à classer. Pour chaque classe (positif/négatif), les scores obtenus avec les différents classifieurs sont ajoutés avec différentes pondérations. Au final, la classe totalisant le score le plus élevé est la classe choisie pour le nouveau texte.

$$score_2^c = \sum_{t \in \mathcal{T}} pond_t * score_t^c$$

$$score_3^c = \sum_{t \in \mathcal{T}} e^{pond_t} * score_t^c$$

COMBINER ÉGALEMENT LE CLASSIFIEUR NON MODIFIÉ Le classifieur entraîné sur l'intégralité du corpus sans prise en compte

des marqueurs multi-polaires, c'est à dire sans aucune modification, est également intégré au score composé.

$$score_4^c = score_{ini}^c + \sum_{t \in \mathcal{T}} pond_t * score_t^c$$

$$score_5^c = score_{ini}^c + \sum_{t \in \mathcal{T}} e^{pond_t} * score_t^c$$

COMPENSATION DE LA TAILLE DES DIFFÉRENTS SOUS-CORPUS THÉMATIQUES

La partition automatique du corpus initial ne donne pas nécessairement des sous-corpus de taille équivalente. On peut vouloir compenser cet état de fait et redonner de l'importance aux petites sous-parties.

$$score_6^c = \sum_{t \in \mathcal{T}} exp\left(pond_t * (1 - e^{-\frac{NBtexte_t}{NBtexte_{max}}})\right) * score_t^c$$

10.5 ÉVALUATION DES RÉSULTATS ET INFLUENCE DES DIFFÉRENTS PARAMÈTRES

Nous avons réalisé plusieurs études croisées afin de déterminer les paramètres optimaux. Sur le corpus de tweets utilisé par SemEval 2013, la combinaison de paramètres ayant permis d'obtenir les meilleurs résultats est la suivante : pour la détection des marqueurs multi-polaires, une p-value de 0.05, une différence minimale de positivité de 0.1 (minDiff) et un nombre d'occurrences minimal dans les corpus source et cible de 10 (minOcc).
La méthode de particularisation fonctionnant le mieux est celle consistant à retirer pour chaque sous-domaine les marqueurs sélectionnés de l'intégralité du corpus (suppression totale, voir 9.5). La méthode de mixage des résultats la plus stable a été d'utiliser l'exponentielle du score LDA (voir 10.4 - score 4).
Avec ces mêmes méthode et fusion, les paramètres de sélection optimaux pour les corpus *DEFT* et *MDSD* ont été p-value=0,01, minDiff=0,1 et minOcc=20 bien que l'influence de ce dernier soit très faible. Sauf mention contraire explicite, les résultats présentés ci-dessous sont obtenus avec ces paramètres. L'influence plus spécifique de chaque paramètre est étudiée aux paragraphes 10.5.4 et 10.5.5.

10.5.1 *Un comportement semblable au cas des corpus multi-domaines*

La figure 26 montre que l'utilisation d'une partition automatique avec la LDA n'a pas modifié le comportement que l'on obtenait en utilisant une partition manuelle. Nous obtenons toujours une amélioration modeste sur le corpus français et des résultats similaires

sur le corpus MDSD. L'utilisation d'une séparation automatique en domaines par LDA peut donc remédier à l'absence d'étiquette de domaine sans perte de performance.

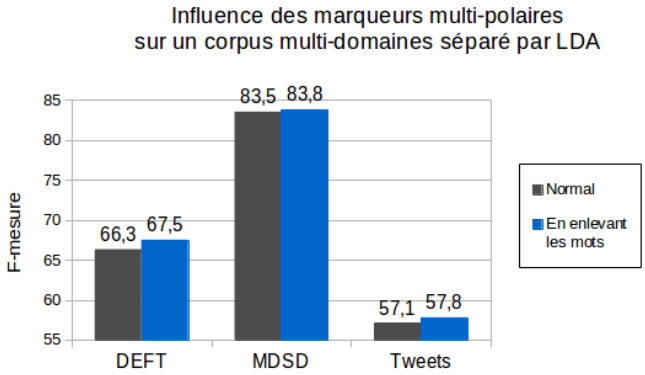


FIGURE 26: Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus en domaine ouvert français (DEFT) et anglais (MDSD et Tweets). La séparation en sous-domaines thématiques a été effectuée par LDA.

Les différents tests de particularisation réalisés sur le corpus de tweets et présentés figure 27 montrent bien qu'enlever tous les marqueurs multi-polaires détectés est le plus efficace. Utiliser des classifieurs entraînés uniquement sur les sous-parties du corpus d'entraînement associées aux différents thèmes dégrade significativement les performances. En effet, les corpus d'entraînement sont dans ce cas bien plus petits. Ce test sert à montrer l'intérêt des marqueurs multi-polaires intra-corpus par rapport à uniquement prendre en compte les informations de domaines induites par le modèle de thèmes de la méthode LDA.

Pour ce qui est du corpus de tweets, nous obtenons une très faible amélioration (+0.7 %) qui est néanmoins significative (selon un test de significativité par échantillonnage aléatoire). Ce résultat doit être mis en relation avec le petit nombre de marqueurs multi-polaires détectés (en moyenne, 36 par domaine). Nous pensons que la taille du corpus, combinée aux 144 caractères des tweets, est trop petite pour que le test du χ_2 détecte beaucoup de marqueurs avec suffisamment de confiance. Pour comparaison, dans notre expérience sur les critiques en anglais, nous avions détecté 400 marqueurs multi-polaires par domaine. Nous nous sommes demandé si, pour ce corpus, des domaines plus focalisés sur un seul sujet pouvaient contrebalancer l'effet du manque de données.

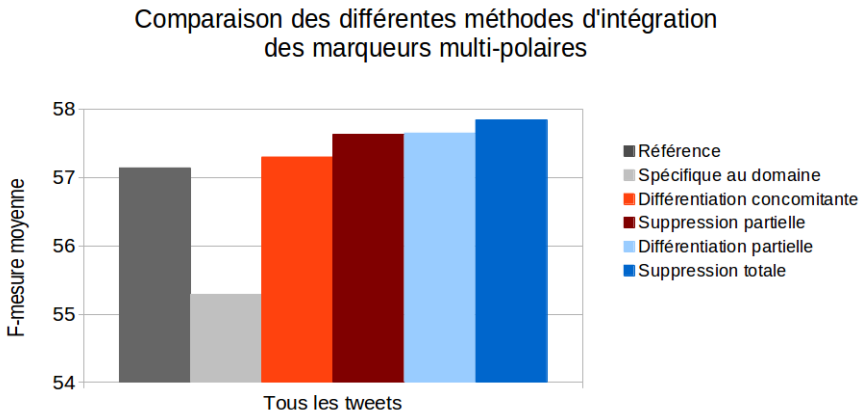


FIGURE 27: Comparaison des différentes techniques d'intégration des marqueurs multi-polaires sur le corpus de tweets.

10.5.2 *Recentrage des sous-domaines détectés*

Nous avons donc réalisé une seconde subdivision du corpus de tweets. Cette fois, un tweet n'est pris en compte que si plus de 75 % de ses mots appartiennent au même thème. Ainsi, un tweet dont seulement 55 % des mots appartiennent à un certain thème ne sera pas retenu. Dans cette version, les sous-parties du corpus d'entraînement obtenues sont plus focalisées sur un seul et même thème. En retour, elles contiennent moins de tweets et donc moins de données d'entraînement.

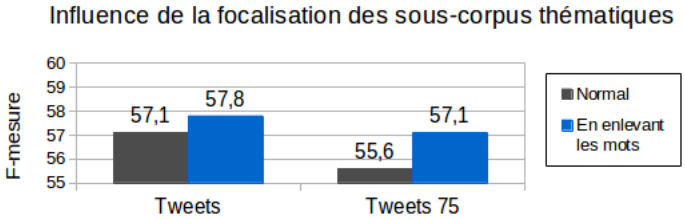


FIGURE 28: Deux corpus d'entraînement différents sont utilisés. *Tweet* contient l'intégralité des tweets tandis que *Tweets75* contient uniquement ceux qui sont focalisés sur un seul thème.

La figure 28 montre que pour l'expérience avec seulement les tweets les plus focalisés, l'amélioration est plus sensible (+1,46 % contre +0,70 %) bien que la valeur absolue du score reste inférieure en raison de la taille bien plus petite du corpus d'entraînement.

10.5.3 *Test des différents mixages*

Afin de comparer la performance des mixages envisagés, présentés à la section 10.4, sur le corpus de tweets, nous avons extrait des corpus d'entraînement équilibrés de 4000 textes. Nous avons choisi cette taille car cela permet une variation du corpus d'entraînement tout en conservant une taille de corpus suffisamment grande pour nos traitements statistiques. Les résultats présentés sont donc les valeurs moyennes obtenues sur ces différents corpus aléatoires.

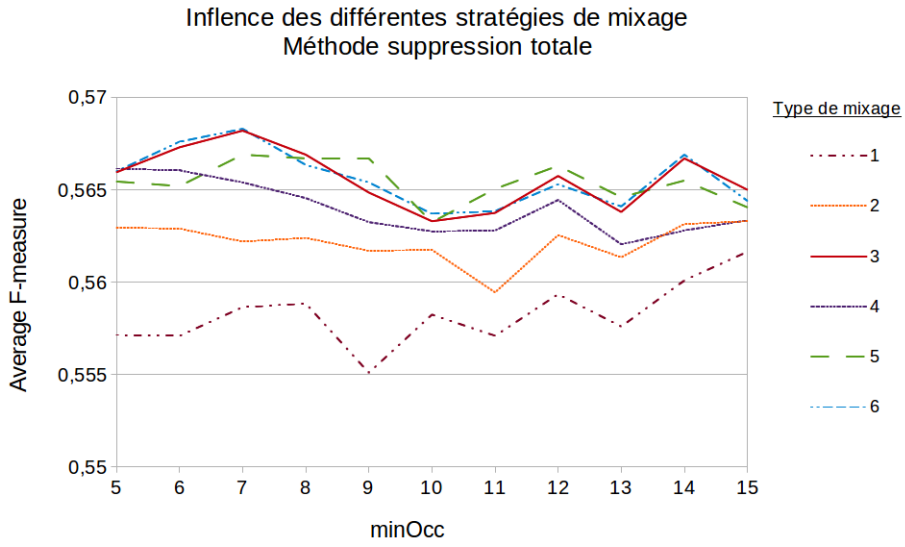


FIGURE 29: F-mesure moyenne en fonction de minOcc pour plusieurs types de mixages sur le corpus de tweets. Méthode : suppression totale.

Il pourrait sembler naturel d'utiliser le classifieur entraîné sur un corpus particularisé pour un domaine précis pour classer les textes se rapprochant le plus de ce domaine (score 1). Cependant, on observe à la figure 29 que c'est le mixage offrant les moins bons résultats. En effet, le nouveau texte se rapproche d'un mélange de domaines. Prenons l'exemple d'un texte classé à 40 % dans un thème et à 20 % dans trois autres. Il est compréhensible que le fait de n'utiliser que le classifieur entraîné sur le sous-domaine correspondant au premier thème résulte en une perte d'information et donc des résultats moins bons que lorsque l'on utilise un mélange de classifieurs.
Utiliser les poids des scores LDA des textes afin de pondérer les résultats des classifieurs est ainsi plus adapté. Toutefois nous observons une nette différence entre l'utilisation directe des pondérations LDA (score 2) et l'utilisation de l'exponentielle de ces pondérations (score 3) qui donne de meilleurs résultats. En effet, dans ce dernier cas, la priorité est donnée au classifieur entraîné sur le sous-domaine

qui est présent en majorité dans le nouveau texte, sans pour autant perdre les informations provenant des autres sous-domaines.
Ajouter les scores du classifieur entraîné sur le corpus non modifié améliore les résultats par rapport à la pondération utilisant les scores LDA bruts (score 4). Par contre, nous observons pas d'amélioration par rapport à la pondération utilisant les exponentielles des scores LDA (score 5).
Enfin, prendre en compte la taille des sous-corpus lors de la pondération exponentielle ne change pratiquement rien (score 6).

En conclusion, parmi les méthodes de mixage testées, trois se démarquent : celles utilisant en pondération l'exponentielle des poids LDA, c'est à dire le degré d'appartenance du texte à classer avec les thèmes sous-jacents calculés avec la méthode LDA. Rajouter le classifieur initial ou bien prendre en compte la taille des thèmes ne semble pas changer les résultats. Nous avons donc opté pour la formulation la plus simple, à savoir le mixage numéro 3, qui combine les classifieurs uniquement à l'aide de l'exponentielle des poids LDA.

10.5.4 *pVal et minDiff : description des marqueurs multi-polaires*

Le paramètre minDiff, différence minimale entre les scores de positivité entre les domaines source et cible afin qu'un mot puisse être envisagé comme marqueur multi-polaire, est directement lié à la définition des marqueurs multi-polaires. En effet, si l'on veut détecter les marqueurs multi-polaires entre deux domaines afin de produire un lexique utilisable lors d'une autre application linguistique, cette différence de positivité doit être fixée en fonction de ce que l'on recherche. Le paramètre pVal, la p-value du test du χ^2, doit également être fixé en fonction de la confiance que l'on souhaite accorder aux marqueurs sélectionnés. Le paramètre minOcc quant à lui, le nombre minimal d'occurrences à la fois dans le domaine source et le domaine cible, va *a priori* influencer la rareté ainsi que la confiance accordée aux mots sélectionnés. Cela est très lié à la taille du corpus et n'est pas un paramètre intrinsèque à la définition des marqueurs que l'on veut sélectionner.

Pour ce qui est des expériences d'intégration directe des marqueurs multi-polaires dans les classifieurs d'opinion, on constate que les deux paramètres minDiff et pVal sont liés. En effet, si l'on en relâche un, il faut rendre le second plus restrictif afin de conserver un niveau de performance équivalent. Ce comportement suggère un lien linéaire entre les deux paramètres. On peut le voir en regardant dans le corpus de tweets, quelles sont les valeurs de p-value et différences de positivité des marqueurs multi-polaires détectés entre le premier sous-domaine et les autres. On calcule alors sur ces données le coef-

ficient de corrélation de Pearson, permettant d'analyser les relations linéaires et le coefficient de corrélation de Spearman, plus adapté aux relations non-linéaires monotones. Il est à noter que ces coefficients ne renseignent pas sur le degré de significativité d'une relation. En prenant un seuil minOcc de 10, celui avec lequel les meilleurs résultats sont obtenus sur le corpus de tweets, le coefficient de Pearson est de -0,708 et celui de Spearman de -0,785. Plus la valeur absolue de ces coefficient est proche de 1, plus le lien entre les deux paramètres testés est avéré. Ici, il y a une corrélation modérée. Néanmoins, la faible variation entre les deux scores montre que cette corrélation est effectivement de type linéaire. Aussi, compenser l'assouplissement d'un paramètre par le renforcement de l'autre, au delà de la conservation des performances, va permettre de garder la collection de marqueurs multi-polaires détectés stable. Autrement dit, ce seront en grande majorité les mêmes mots qui seront détectés.

10.5.5 *minOcc : pertinence statistique*

Le but premier d'un seuil sur le nombre d'occurrences est de rendre le test du χ^2 pertinent d'un point de vu statistique. En effet, il s'agit d'un test à la limite qui ne peut donc pas être effectué sur de trop petits échantillons. Cela se voit d'ailleurs dans les résultats de nos expériences. S'il n'y a aucun seuil en nombre d'occurrences, les résultats chutent, même par rapport à un seuil faible de 5 occurrences. On peut l'observer sur l'exemple présenté à la figure 30 qui présente les résultats obtenus pour l'intégration de type suppression totale et le mixage 0. Quelque soit la combinaison des paramètres pVal et minDiff, l'absence de seuil sur le nombre minimal d'occurrences nécessaires pour être considéré comme marqueur multi-polaire (minOcc = 1) entraîne une baisse des performances, parfois jusqu'à 2 points de F-mesure.

Naturellement, le seuil d'occurrence va déterminer la possible rareté des marqueurs pris en compte. Il faut rappeler que des mots rares peuvent être très informatifs quant à la polarité d'un texte. Par contre, en fonction de la taille d'un corpus, un même seuil ne correspondra pas au même niveau de rareté. C'est pourquoi on pourrait penser que le seuil minOcc optimal puisse varier en fonction de la taille des corpus utilisés. Cependant, nos expériences pour essayer de mettre en évidence un tel effet ne montrent pas de stabilité à travers les tailles de corpus. Pour ces expériences, nous avons créé des corpus d'entraînement plus petits en sélectionnant au hasard un sous-corpus équilibré à partir du corpus total de tweets. Les tailles testées vont de 2000 à 4500 textes avec un pas de 500. Nous avons effectué plusieurs tirages pour chaque taille testée et considéré les moyennes des scores obtenus. Nous avons fait varier la valeur de minOcc de 5 à 25 par pas de 1. Seuls les corpus de petite taille (2000 textes) présentent une

Référence Pang Lee ?

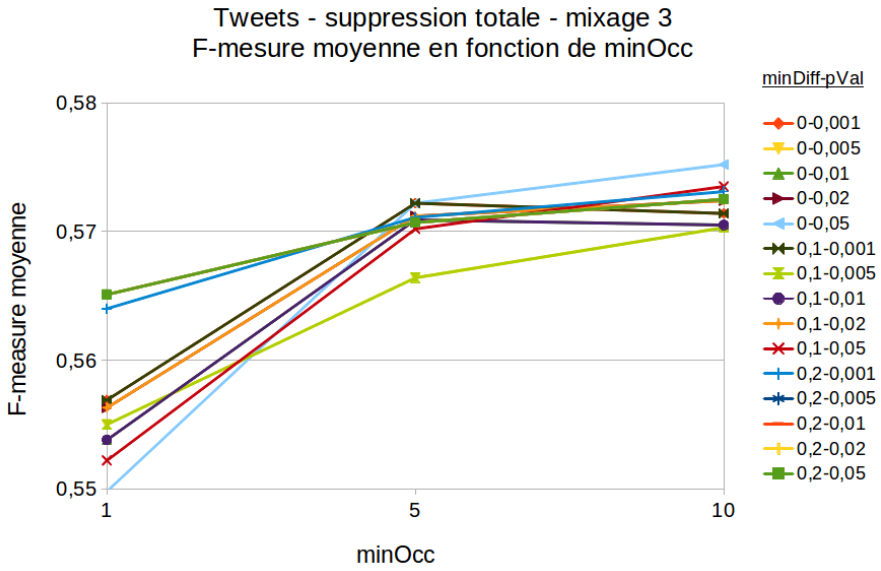


FIGURE 30: F-mesure moyenne en fonction de minOcc pour plusieurs combinaisons de pVal et minDiff sur le corpus de tweets. Méthode : suppression totale ; Mixage 3.

variation significative : les résultats décroissent avec l'augmentation de minOcc. Pour les autres, les résultats demeurent stables.

Il est donc important d'utiliser un seuil minOcc minimal afin d'assurer la pertinence du test statistique du χ^2. Cependant, des valeurs plus élevée n'ont pas d'effet détectable sauf en ce qui concerne les plus petits corpus.

10.6 PARTICIPATION À LA CAMPAGNE D'ÉVALUATION SEMEVAL 2013

La tâche 2B de la campagne d'évaluation SemEval 2013 [Wilson et al., 2013] consistait à indiquer la polarité globale d'un tweet entier parmi les classes positive, neutre ou négative. La métrique d'évaluation retenue par les organisateur a été la F-mesure moyenne des classes positive et négative. Trente-huit équipes ont participé à cette tâche 2B. Elles ont été réparties en deux catégories : système contraint et système non-contraint. Un système non-contraint peut s'entraîner sur des corpus de tweets supplémentaires ou rajouter des annotations. Un système contraint n'utilise pas de données d'entraînement supplémentaires. Par contre il peut utiliser des lexiques ou d'autres ressources lexicales ayant été au préalable établies sur d'autres tweets, annotés ou non.

Les pré-traitements utilisés par les différents participants sont très semblables.

- Suppression des liens, des noms d'utilisateurs, des mentions de retweet, parfois des étiquettes (*hashtags*).
- Normalisation des textes par correction orthographique en utilisant des outils développés spécifiquement pour twitter, lemmatisation, stemming, traduction des abréviations, suppression ou remplacement des émoticônes, traitement des voyelles répétées, réduction des signes de ponctuation répétés.
- Détection des parties du discours ou des relations syntaxiques

De même, les traits utilisés par les différents classifieurs sont souvent similaires. Plusieurs articles mentionnent le fait que les unigrammes en sac de mots contribuent le plus aux performances, suivis par l'utilisation des différents lexiques. Les points d'exclamation et d'interrogations semblent être très discriminant, surtout redoublés et en dernière position. Enfin, les émoticônes ainsi que les étiquettes (*hashtags*) porteuses d'opinion sont également très informatifs.

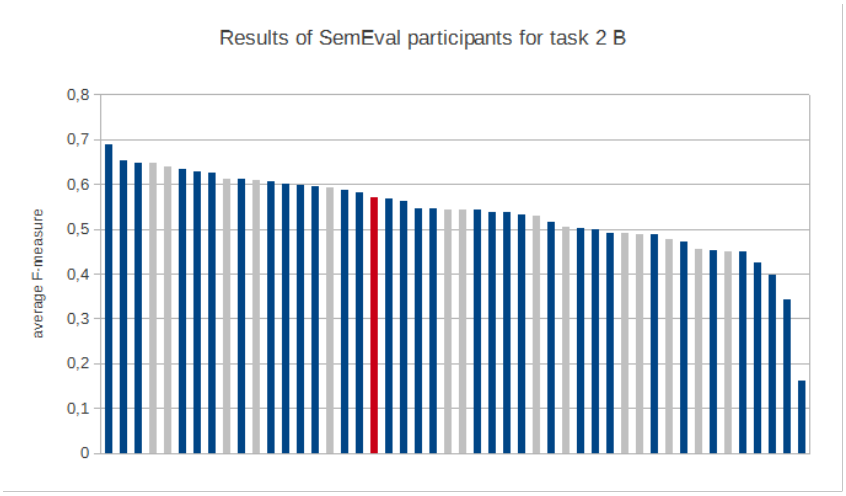


FIGURE 31: Notre participation à SemEval parmi celles des autres participants. Les barres bleues sont placées dans la catégorie contraint.

Lors de notre participation à cette campagne d'évaluation, nous avons considéré le corpus de tweets d'entraînement comme un corpus en domaine ouvert et avons appliqué notre méthode de détection des marqueurs multi-polaires intra-corpus comme décrit précédemment. La particularisation utilisée lors de la soumission a été celle de différenciation concomitante. La différenciation de tous les marqueurs multi-polaires détectés pour chaque sous-domaine se réalisant en même temps, il n'y a pas de fusion nécessaire. Nos résultats se placent au milieu des autres participants (tableau 31).

Notre système est naturellement contraint. Nous n'utilisons même aucune ressource supplémentaire alors que nombre de contributions s'appuient sur des lexiques d'opinion. Nous avons essayé à la lecture des articles des autres participants de nous comparer avec leur systèmes avant l'adjonction d'information extérieure. Cette information n'était présente que pour dix-sept équipes, qui ont parfois soumis plusieurs versions. Ces résultats sont présentés à la figure 32, qui met l'accent sur le type de classifieurs utilisés par les différentes équipes.

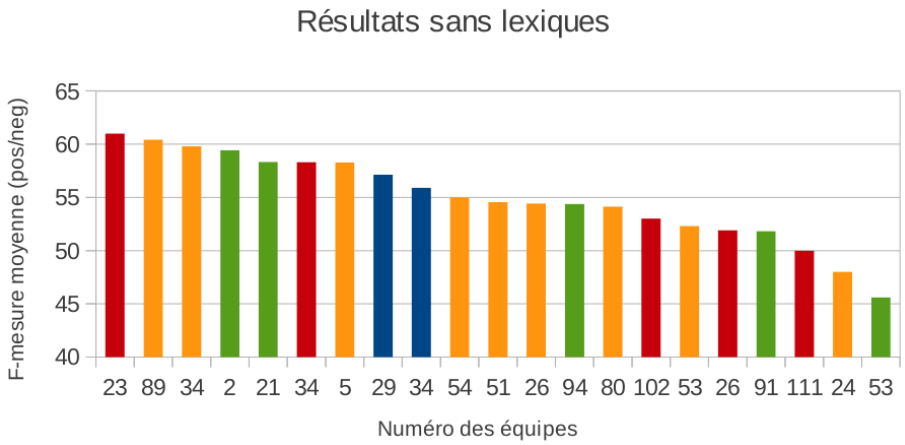


FIGURE 32: Résultats mentionnés par les différentes équipes dans leurs articles avant l'adjonction de lexiques. Rouge : Maximum d'entropie ; orange : SVM ; vert : Naives Bayes ; bleu : Boosting

Il semble difficile d'établir une hiérarchie des classifieurs. D'autant plus que, pour les équipes ayant testé plusieurs classifieurs, l'ordre des résultats n'est pas toujours le même. Dans la plupart des cas, le SVM l'emporte sur le Maximum d'Entropie qui l'emporte sur le boosting. Le SVM l'emporte également sur le Naives Bayes. Seulement, il y a toujours des équipes pour lesquelles le phénomène inverse se produit.

10.7 CONCLUSION

Cette étude conclut donc l'étude des impacts pratiques des marqueurs multi-polaires sur les classifieurs automatiques d'opinion en présentant le test de différentes méthodes de fusion ainsi qu'une étude de l'influence des différents paramètres de sélection des marqueurs. Les paramètres pVal et minDiff sont liés linéairement et permettent de définir les caractéristiques souhaitées pour les marqueurs détectés. Le paramètre minOcc, quant à lui, rend possible l'utilisation des tests statistiques.
Nous avons également montré dans ce chapitre qu'il est possible de prendre en compte les domaines des textes d'entraînement et de test même si ceux-ci ne sont pas explicitement spécifiés. En effet, une mé-

thode de séparation automatique en domaine, combinée à une fusion de classifieurs permet d'obtenir les mêmes résultats que pour le cas du multi-domaine.
Enfin, nous avons appliqué notre méthode sur un corpus de tweets avec les mêmes résultats que sur des critiques de produits. Sur ce dernier corpus, nous avons également mis en lumière le fait que la détection des marqueurs multi-polaires intra-corpus apporte plus d'améliorations potentielles lorsque les domaines sous-jacents sont plus distincts les uns des autres.

11 CONCLUSION

Au cours de cette thèse nous nous sommes intéressée à la problématique de l'adaptation au domaine pour la classification de l'opinion au niveau du texte. Le domaine désigne ici le type d'objet à propos duquel l'opinion est exprimée.
Lorsque l'on entraîne un classifieur sur un corpus afin de l'utiliser ensuite sur d'autres textes, l'hypothèse implicite est faite que la distribution des traits de représentation des textes dans le corpus d'entraînement est la même que dans les textes que l'on souhaite classer. En pratique, ce n'est pas toujours le cas, et nous nous penchons dans cette thèse sur ce problème, et plus particulièrement sur les mots dont la polarité d'opinion change d'un domaine à un autre.

TRANSMISSION DE POLARITÉ PAR CRÉATION DE CLUSTERS

Aborder le problème des mot inconnus par la création d'un espace de projection commun peut être très efficace pour effectuer une classification de l'opinion. Cependant il devient plus difficile d'interpréter les résultats car les mots n'existent plus en tant que tels dans la représentation des textes dans le nouvel espace. C'est pourquoi nous avons proposé la création de clusters de mots à l'intérieur de l'espace commun de projection. Afin d'être utiles, de tels clusters doivent être hétérogènes et purs, c'est à dire regrouper des mots appartenant aux deux domaines et être d'une même polarité. Nous avons mis en place deux mesures permettant d'évaluer ces caractéristiques.
Une fois constitués, ces clusters peuvent servir de traits de représentation additionnels pour les classifieurs. Leur principal intérêt réside cependant dans leur capacité à transmettre de l'information polaire du domaine source au domaine cible. Nos expériences suggèrent d'ailleurs que l'utilisation de bigrammes dans ces clusters permet de reconstruire des expressions multi-mots et ainsi d'effectuer le transfert de polarité également sur celles-ci et non uniquement sur des mots simples.

DÉTECTION DES MARQUEURS MULTI-POLAIRES

Notre recherche à propos de la création et l'utilisation de ces clusters nous a permis de mettre en lumière le problème des marqueurs multi-polaires. En effet, de tels mots vont non seulement perturber la création de l'espace commun de représentation mais vont également fausser le transfert d'information polaire à l'intérieur des clusters que

nous avons créés. Pour des raisons pragmatiques, nous avons fait le choix de ne pas présupposer de la forme que peuvent prendre les marqueurs multi-polaires et de les détecter directement à l'aide des corpus.

Nous avons proposé une première méthode de détection de ces marqueurs lorsque l'on dispose de corpus annotés dans tous les domaines d'intérêt. Nous avons montré qu'ils étaient effectivement utilisés par les classifieurs automatiques d'opinion, ce qui peut mener à des erreurs de classification. Nous avons également proposé une seconde méthode de détection de ces marqueurs sans utiliser d'annotation dans le domaine cible. Nous supposons l'existence d'une collection de mots pivots communs aux deux domaines, utiles à la classification de l'opinion au niveau du texte, au moins dans le domaine source, et ne changeant pas de polarité entre les deux domaines. La difficulté est ici de s'assurer de la constance de cette polarité. Nous avons proposé pour cela une auto-épuration de la collection de mots pivots. Les résultats obtenus présentent un bon rappel mais une faible précision.

Il faudrait poursuivre les travaux dans l'épuration des mots pivots ou bien des marqueurs sélectionnés. Plusieurs pistes sont envisageables. L'analyse des cooccurences des mots pivots entre eux, suggère qu'il serait plus efficace de les rassembler en quatre groupes : entièrement positif, majoritairement positif, majoritairement négatif, entièrement négatif. Les profils de cooccurence seraient calculés par rapport à ces quatre classes plutôt que par rapport à l'ensemble des mots. Une autre possibilité est de représenter les mots avec une succession de vecteurs calculés par exemple par des réseaux de neurones sur des fenêtres de contexte de différentes tailles.

COMPORTEMENT DES MARQUEURS MULTI-POLAIRES

Après avoir sélectionné ces marqueurs multi-polaires en nous appuyant sur les corpus annotés, nous les avons étudiés d'une manière plus fine. C'est pourquoi nous avons mis en place une expérience d'annotation. Après avoir vérifié sur une expérience à petite échelle que notre schéma d'annotation permettait d'attribuer une polarité à chaque marqueur au niveau de la phrase de manière relativement fiable, nous avons lancé l'expérience à plus grande échelle.
L'analyse des résultats obtenus nous a permis d'identifier les différents comportements qu'adoptent les marqueurs multi-polaires dans les corpus. Les marqueurs peuvent avoir un comportement indifférent s'ils apparaissent dans une phrase établissant un contexte ou une description. En dehors de ce cadre, les marqueurs multipolaires se retrouvent dans trois types de comportements qui se rapportent aux rôles des objets impliqués dans l'expression d'une opinion. Certains

sont des marqueurs d'expression d'opinion, avec ou sans spécification de l'orientation de l'opinion. D'autres font référence à l'objet sur lequel est exprimée une opinion. Enfin un marqueur multi-polaire peut désigner une caractéristique de l'objet que l'on évalue. Cette caractéristique peut être explicitement contenue dans le marqueur ou bien celui-ce peut faire référence à une action ou un sentiment induit par l'objet. Cette caractéristique peut être positive, négative ou habituelle.
Ces différents comportements se traduisent par une répartition attendue dans les critiques positives et négatives des corpus. Comme nous l'avons vu, il existe parfois des préférences dans les apparitions des marqueurs dans deux ou trois classes. Ces préférences s'expliquent parfois par un biais de corpus, auquel cas elles ne sont pas légitimes. De manière plus intéressante, elles peuvent être dues à des habitudes culturelles communes entre l'auteur et le lecteur, ces habitudes pouvant bien évidemment dépendre du domaine. Il peut simplement s'agir de tournures de phrases habituelles ou bien par exemple de caractéristiques des objets tellement évidentes qu'on ne songe à en parler que lorsqu'elles sont absentes.

Sur un corpus représentant un domaine donné, un marqueur va potentiellement adopter plusieurs comportements. Il peut y avoir un comportement préférentiel par rapport aux autres ou bien tous les comportements ont la même importance dans ce domaine. Chaque comportement induisant une certaine répartition dans des phrases positives, négatives ou neutres, ce que va en fait mesurer le score de positivité est le mélange de tous ces comportements sur un même domaine. Et pour qu'un mot soit effectivement un marqueur multipolaire, il faut que sa positivité globale dans un domaine soit différente de celle dans un autre domaine. Il faut donc que l'ensemble des comportements adoptés par le mot dans le premier domaine mène à une répartition distinguable par rapport à celle engendrée par l'ensemble des comportements du mot dans le second domaine.
Deux ensembles de comportements distinguables sont souvent un ensemble de comportements globalement neutre dans un domaine et un ensemble de comportements globalement polaire dans l'autre. Nous avons répertorié les associations les plus fréquentes.
La question la plus intéressante est, pour chaque paire de comportements unitaires, d'identifier la raison du changement de comportement entre les deux domaines. Nous en avons donc dressé une typologie. Une des raisons est que le marqueur est employé dans une phrase contextuelle d'un côté et dans une phrase évaluative de l'autre. Il est également possible que la différence soit tout simplement liée à un biais de corpus. Les autres catégories sont les suivantes : changement de sens, changement d'objet et changement d'utilisation. Dans cette dernière catégorie, nous avons observé des changements liés à

une utilisation physique de l'objet qui diffère et d'autres liés à une utilisation différente dans le temps.

Certains mots ont été retirés de l'étude car ils présentaient trop peu d'annotations. Il serait peut-être possible d'exploiter les données recueillies à l'aide de méthodes de types non paramétriques qui, bien que moins précises que les méthodes paramétriques lorsque l'on dispose de nombreuses données, permettent d'interpréter des résultats obtenus sur un très petit nombre d'instances. D'une manière générale, il serait très intéressant de reproduire cette expérience pour le français, et peut-être à plus grande échelle, afin de vérifier que les catégories que nous avons observées sur le corpus anglais s'y retrouvent également.
Enfin, une expérience complémentaire pourrait être de reproduire cette étude au niveau de la phrase et de comparer les marqueurs obtenus dans les deux cas. Il est vraisemblable que les effets des phrases de contexte soient ainsi atténués et que les trois catégories principales de cause de changement de polarité soient mis en exergue.

INFLUENCE DES MARQUEURS MULTI-POLAIRE SUR LES CLASSIFIEURS D'OPINION

Enfin, dans la dernière partie de notre thèse, nous nous sommes intéressée à l'influence concrète de la prise en compte des marqueurs multi-polaires sur la performance des classifieurs d'opinion automatiques au niveau du texte.
Le premier cas que nous avons considéré, est le cas classique d'adaptation au domaine. Nous avons expressément défini les marqueurs multi-polaires afin d'améliorer ce cas de figure. Une fois détectés, les marqueurs multi-polaires sont supprimés du corpus ou bien différenciés entre les deux domaines. Nous observons des améliorations de classification en terme d'exactitude allant jusqu'à cinq points. Pour certains sens d'adaptation cependant, nous n'observons pas d'amélioration même si les détériorations significatives restent rares. D'une manière générale, nous avons observé que plus une adaptation est au départ difficile, plus la prise en compte des marqueurs multi-polaires est potentiellement bénéfique à la classification. En effet, plus le nombre de marqueurs multi-polaires entre deux domaines est élevé, plus ils vont perturber le transfert et plus leur suppression aura de l'effet.

Le second cas auquel nous nous sommes intéressée, est celui d'un corpus d'entraînement multi-domaines. Il n'y a pas ici de nouveaux domaines, ceux apparaissant dans le corpus cible sont les mêmes que ceux déjà rencontrés dans le corpus source. Les marqueurs multi-polaires détectés sont donc intra-corpus. Puisque l'on considère tous

les domaines ensemble, nous ne disposons plus ici que de deux corpus d'évaluation. Les résultats obtenus présentent une petite amélioration sur le corpus DEFT et une stagnation sur le corpus MDSD. En effet, plus un classifieur à la possibilité de s'entraîner sur de nombreux domaines annotés plus il devient robuste. En terme de marqueurs multi-polaires cela se traduit par le fait que ce qui va ressortir est le comportement moyen sur plusieurs domaines. Ainsi, si un mot se comporte différemment dans un dixième du corpus d'apprentissage, c'est le comportement prédominant qui sera appris. Et au final cet apprentissage ne propagera potentiellement des erreurs que dans un dixième du corpus de test. Ainsi les erreurs évitées par la prise en compte des marqueurs multi-polaires intra-corpus sont bien moins nombreuses que dans le cas de l'adaptation au domaine. Les améliorations potentielles sont donc plus faibles. Cependant la détection des marqueurs n'étant pas très coûteuse et n'entraînant pas de dégradation des performances, il est toujours possible de l'intégrer au sein d'une chaîne de traitement plus complexe. De plus, malgré les moins bons résultats quantitatifs dans le cadre de la classification, nous pouvons espérer obtenir par la suite une amélioration qualitative utile pour une application d'acquisition de ressources lexicales d'opinion.

Le dernier cas auquel nous nous sommes intéressée, est le cas d'un apprentissage en domaine ouvert. Les différents domaines des textes des corpus source et cible sont ici inconnus. Il s'agit d'un cas particulier à notre connaissance non traité dans les travaux de traitement automatique des langues. Afin de se rapporter au cas précédent des multi-domaines, nous avons proposé d'utiliser des méthodes de détection automatique de thèmes de type LDA. Une fois le corpus d'entraînement séparé en sous-domaines à l'aide de cette détection de thèmes, nous entraînons le classifieur comme pour le cas précédent. La classification des nouveaux textes fait intervenir une fusion de classifieurs puisque leurs domaines d'appartenance est également inconnu. Nous avons testé plusieurs méthodes de mixage à partir des scores obtenus par la séparation en thèmes. Les résultats obtenus sont équivalents à ceux obtenus pour le corpus multi-domaine ce qui montre que l'utilisation d'une séparation automatique en thèmes associée à une fusion de classifieurs permet de pallier le problème de l'absence d'étiquette de domaine spécifique.

Il serait cependant intéressant d'appliquer des méthodes de détection de thèmes ne nécessitant pas la spécification préalable du nombre de thèmes recherchés et également de fusionner les classifieurs en utilisant des techniques de fusion d'experts plus élaborées. Nous avons également testé cette méthode sur le corpus de tweets de la campagne d'évaluation SemEval 2013 à laquelle nous avons participé. Nous avons montré que dans le cas d'une focalisation plus forte

des sous-domaines autour d'un même thème, le potentiel d'amélioration de notre méthode est plus élevé.

POUR ALLER PLUS LOIN - MARQUEURS MULTI-POLAIRES ET CLUSTERS

Dans cette thèse, nous avons proposé une méthode à l'aide de clusters afin de faire explicitement des liens entre mots de différents domaines pour aider à transférer des informations de polarité d'un domaine à l'autre et ainsi développer de nouvelles ressources lexicales. Ce faisant, nous avons mis en évidence le problème des marqueurs multi-polaires que nous avons donc étudiés. Nous en avons proposé des méthodes de sélection, élaboré une classification suite à une expérience d'annotation et évalué leur influence sur des classifieurs automatiques d'opinion dans différents cas.
Pour poursuivre dans la logique de cette thèse, la prochaine étape devra logiquement être l'intégration des marqueurs multi-polaires dans la création des clusters afin d'améliorer le transfert et de produire des ressources plus performantes dans le domaine cible.
En effet, l'utilisation de lexiques d'opinion, lorsqu'ils sont disponibles et de bonne qualité pour la langue et le domaine ciblés, est très efficace pour la fouille d'opinion. Ces lexiques sont d'autant plus utiles lorsque l'on s'intéresse par exemple à la détection d'opinion à grain fin, de la cible, de l'aspect ou de l'auteur de l'opinion. Or, la création manuelle de telles ressources peut être couteuse, en temps, en personne et en coût. L'étude de méthodes de création de ressources de la manière la plus automatique possible est donc un grand défi pour le domaine
Les méthodes que nous avons proposées et étudiées au cours de cette thèse sont très prometteuses de ce point de vue. En sus de permettre le développement de ressources pour de nouveaux domaines, nous les avons volontairement conçues de manière à ce qu'elles soient le plus indépendantes possible de la langue utilisée. Ainsi, elles sont en théorie utilisables pour beaucoup de langues, éventuellement après un pré-traitement de dés-agglutination ou de séparation de caractères. Cet aspect multilingue mérite d'être exploré plus en détail, afin de permettre la création de ressources pour des langues peu dotées.

A

EXEMPLE DE CRITIQUES

Cette section rassemble quelques exemples de critiques issues des corpus étudiés. Le corpus français contient des textes écrits par des critiques de métier, au contraire du corpus anglais qui recueille des avis d'internautes. Cela entraîne des différences en terme de vocabulaire et de taille de critique. Ce dernier critère varie également en fonction du domaine de la critique.

CORPUS FRANÇAIS ISSU DE DEFT 2007

Domaine : jeux vidéos

The Great Art Race
Les jeux de société n'ont que très rarement donné quelque chose de convaincant sur PC. Rappelez-vous de la série des Monopoly, dont les multiples adaptations n'ont pas éclipsé les jeux de plateaux ; toujours plus agréables à jouer. Bref, il faut vraiment que ces portages apportent un plus pour séduire les joueurs. Pour The Great Art Race ça semble bien parti puisque son concept est original. Hélas, le tableau se ternit au fil des heures, c'est le cas de le dire.
Le but du jeu est simple et tient en quelques mots : Walter, un riche amoureux de l'art pictural a convoqué ses cinq nièces et neveux, dont vous faites partie, pour leur faire une annonce. Atteint d'une grave maladie, il va bientôt mourir et voudrait, avant son départ pour l'au-delà, retrouver ses tableaux qui ont été volés. Ces derniers réapparaissent périodiquement dans des ventes aux enchères. Il va donc vous falloir parcourir le monde pour racheter les précieuses peintures. Pour ce faire, il vous faudra évidemment gagner de l'argent mais aussi éviter d'acquérir de simples copies. Et tout cela, vous devez le faire sans perdre de temps puisque ce sera la personne qui aura retrouvé le plus de tableaux qui héritera de la fortune de l'oncle Walter. C'est pas très gentil pour les autres, mais c'est comme ça.
Lorsqu'on entame une partie, il faut tout d'abord décider du temps que vous aurez pour accomplir votre mission : deux, quatre ou sept ans. Cela influe évidemment sur la longueur de la partie qui varie ainsi d'une bonne heure à quelques heures tout au plus. A la fin du temps imparti, votre score s'affichera et vous saurez si vous êtes celui qui, parmi les cinq nièces et neveux, a su rassembler le plus de peintures. Le jeu se joue donc à cinq. Problème : il n'y a aucun mode LAN, ni de prise en charge d'internet ou de jeu par e-mail. C'est ennuyeux car cela limite le multijoueur, essence même de ce jeu, à des

parties à tour de rôle sur le même PC. Vraiment dommage et très peu pratique. Heureusement, on peut aussi y jouer seul contre des adversaires dirigés par l'intelligence artificielle. Une IA qui est d'ailleurs bien moins efficace et beaucoup plus prévisible que de vrais joueurs. Mais passons sur ce (gros) défaut et voyons de quelle façon vous pourrez gagner assez d'argent pour acquérir les peintures. The Great Art Race se déroule en tour par tour sur une carte du monde sur laquelle plusieurs villes sont signalées. Vous pouvez vous rendre de l'une à l'autre en déplaçant votre pion à l'écran. Chaque ville a évidemment ses spécificités et est représentée par un écran fixe et plusieurs boutons qui sont autant d'options. Votre souci principal étant de gagner de l'argent, il vous faudra beaucoup voyager et vous investir dans différentes activités. La première est la culture de denrées. Dans certaines villes, vous pourrez établir des plantations (de coton, soie, cacao, tabac, café ou thé) et recruter du personnel pour cultiver la terre. Une fois que c'est fait, la production commencera et vous pourrez revendre vos biens sur le marché de Londres ou de New York en acheminant la marchandise par le biais de l'entreprise qui dessert la ville de production. Évidemment, le prix de chaque bien varie selon la loi de l'offre et de la demande. Si par exemple, tous les joueurs produisent du café, son prix va chuter. Il est donc judicieux de diversifier ses productions. Ce qui est problématique, c'est qu'il faut très régulièrement revenir dans les villes où vous avez des plantations pour payer le personnel et exporter la marchandise sous peine de voir vos ouvriers faire grève car ils n'ont pas été payés. Ces actions ne peuvent se faire que manuellement, il n'y a pas d'automatisation possible. Cela vous contraindra donc à faire de nombreux allers-retours entre vos centres de production et les marchés. Lassant au bout d'un moment. Autre moyen de gagner de l'argent : la spéculation. Vous pouvez acheter et revendre les actions des cinq sociétés de transport de marchandises. Le but étant évidemment d'acheter au prix le plus bas et de revendre au prix le plus haut pour faire de confortables bénéfices. Le prix de l'action est lui aussi intimement lié à l'offre et à la demande mais également à l'activité de la société. Ainsi, si elle doit transporter de grosses quantités de denrées, le prix de l'action de l'entreprise va monter. Pour faire une bonne affaire, il suffit donc d'acheter des actions lorsque l'entreprise n'a quasiment pas de travail, puis d'établir des plantations dans une ville qu'elle dessert pour qu'elle puisse travailler. Le prix de son action va monter ce qui vous permettra de les revendre. L'achat d'un hôtel ou emprunter à la banque vous permettra également de gagner un peu d'argent, mais ces méthodes ne sont pas forcément les plus rentables. Notamment la dernière puisque vous devrez à un moment ou à un autre rembourser votre emprunt auprès de la banque.

Au bout de quelques dizaines de minutes, on comprend très bien tous ces mécanismes et on n'a aucun mal à faire grimper son compte en

banque. Il sera alors temps de songer à se rendre dans les ventes aux enchères pour commencer à acheter les fameuses toiles. Les dates de ces événements vous sont communiquées à l'avance et il faut donc tenter d'arriver dans la ville en question au bon moment. Vous pourrez alors surenchérir, tout comme les autres neveux et nièces présents. Mais il faut faire attention à bien identifier le bon grain de l'ivraie. Pour cela, il faut prendre des cours. Ces cours (un pour chaque mouvement artistique) prennent deux semaines et sont évidemment payants. Mais une fois que vous aurez participé à ces séances, vous pourrez voir si le tableau est un original ou pas. Comment ? C'est très simple. A chaque vente d'un tableau dont vous avez étudié le courant pictural, s'affichera un code du genre "P3", un code qui ne s'affichera pas si vous n'avez pas de connaissances particulières. Il suffira alors de vous reporter au fichier pdf de décodage présent sur le CD du jeu pour voir à quoi correspond le sigle. En l'occurrence, P3 est une œuvre authentique. Bref, identifier une peinture est une activité qui se résume à lire un code sur un tableau. Pas très palpitant. On aurait aimé que les développeurs se lancent dans quelque chose de plus complet, en nous apprenant pourquoi pas à voir les défauts d'une œuvre. Bref, le contenu laisse à désirer et le joueur doit répéter constamment les mêmes actions, en l'espèce, faire des allers-retours entre les marchés, ses plantations et les ventes aux enchères. Un aspect répétitif qui est fatal au jeu qui s'en serait mieux sorti avec un gameplay un peu plus riche et un vrai mode multijoueur supportant internet.

Graphismes.

Les graphismes sont uniquement constitués de plans fixes. Certes, ces derniers sont plutôt jolis, mais ils manquent cruellement d'animation.

Jouabilité.

Simple à prendre en main, le jeu souffre néanmoins des possibilités trop faibles qu'il offre. Les manières de gagner de l'argent sont peu nombreuses et trop stéréotypées. De plus, l'intelligence artificielle n'est pas des plus convaincantes et le système de repérage des faux tableaux est trop simpliste.

Durée de vie.

C'est très répétitif. Il n'est vraiment pas certain que vous enchaîniez les parties en solo à cause d'une IA trop prévisible. C'est mieux en multijoueur, mais on regrette que cette fonctionnalité soit limitée à un mode hot seat et qu'il n'y ait aucune possibilité de jouer en LAN, sur internet ou par e-mail car The Great Art Race s'y serait prêté parfaitement.

Bande son.

La bande-son est pauvre. Il n'y a pas beaucoup d'effets sonores et de musiques différentes. Heureusement, le peu qu'il y a est plutôt convaincant.

Scénario.

Note Générale.
The Great Art Race pourrait presque être qualifié de jeu de société sur PC tant les mécanismes sont proches de ces derniers. Hélas, ses fonctionnalités paraissent trop limitées pour que le titre se fasse une place au soleil. En outre, l'absence de prise en charge d'internet est vraiment préjudiciable et nuit à la durée de vie puisque les réactions de l'IA en solo sont trop prévisibles.

Domaine : musique

Depuis trente-six ans, chaque nouvelle production de David Bowie est un événement. Heathen , ne fait pas exception à cette règle. On reconnaît instantanément la patte de son vieux compère Tony Visconti. La voix de Bowie est mise en avant. Agréable surprise, surtout qu'elle n'a rien perdu depuis ses débuts. Là, commence le voyage. Ambiance, mélange dosé des instruments. Dès l'ouverture de l'album avec Sunday , un sentiment étrange nous envahit. Comme si Bowie venait de rentrer d'un voyage expérimental au cœur même de la musique. Retour aux sources.
L'ensemble du disque est rythmé par cette pulsation dont le duo a le secret. Le tout saupoudré de quelques pincées d'électronique. Le groupe est réduit au minimum. Outre Bowie en chef d'orchestre et Visconti, David Torn ponctue les compositions de ses guitares aventureuses et Matt Chamberlain apporte de l'âme à la rythmique. Un quatuor à cordes fait une apparition, comme Pete Townshend (The Who) ou Dave Grohl (ex-batteur de Nirvana).
Avec trois reprises réarrangées et neuf compositions originales, le 25e album de Bowie est à l'image d'une cohérence artistique retrouvée.

Domaine : livres

Certains esprits chagrins ont osé prétendre qu'avec la fin de l'apartheid se tarirait la veine romanesque de Nadine Gordimer. C'était bien mal connaître le prix Nobel de littérature 1991, écrivain engagé s'il en est, qui citait ces mots de Garcia Márquez dans son discours de réception : "La meilleure façon pour un écrivain de servir la révolution, c'est d'écrire aussi bien qu'il peut."
Pour Nadine Gordimer, la révolution ne s'est pas arrêtée à l'avènement d'une ère nouvelle en Afrique du Sud. A bientôt quatre-vingts ans, la voici qui creuse de nouveaux sillons et aborde, à travers le récit d'un amour voué à l'échec, l'un des problèmes-clés du XXIe siècle, celui de l'immigration.
Casse-cou, a-t-on envie de crier à Julie, jeune femme de la bourgeoisie blanche du Cap, en rupture de ban avec son milieu, lorsqu'elle épouse Abdou, émigré sans papier. Que de malentendus entre ces deux-là,

que de choses à peine dites et encore moins entendues. Lui, en particulier, ne parvient pas à admettre son refus de l'aider puisqu'elle le peut, puisqu'elle connaît "des gens intéressants". Ceux qu'elle vomit, justement. Mais comment pourrait-il le comprendre, puisque son unique ambition est d'un jour leur ressembler ?
Dénoncé, obligé de quitter l'Afrique du Sud, Abdou rentre au pays, dépourvu de tout ce qu'un fils qui a réussi peut apporter à ses parents, "avec rien d'autre qu'une épouse - une étrangère". Et, pendant qu'il s'active à essayer d'obtenir un visa pour n'importe quel pays occidentalisé, elle, doucement, s'accoutume à ce paysage si loin de celui qu'elle imaginait, le désert au bout de l'unique rue du village, même pas un palmier. Elle avance sereinement avec sa différence, apprivoise les femmes et les enfants, accepte de comprendre les règles de cette société régie par la loi de l'Islam, fait son chemin, à l'écoute et tolérante, dans ce coin perdu qui n'est que ruine et pauvreté...
Un amant de fortune. Plongée en profondeur dans le quotidien d'une famille musulmane vivant dans un quelque part aride et incertain. Éblouissant portrait d'une femme prenant conscience d'elle-même dans un dénuement où pensées et sentiments retrouvent leur juste place. D'un homme pour qui ce même dénuement ne peut être qu'une humiliation permanente.
C'est le paradoxe magnifique de ce roman écrit d'une plume concise et vibrante : il est plus facile de renoncer à ce qu'on possède qu'à ce qu'on ne possède pas, tant il est vrai qu'il est impossible de renoncer à ses rêves...

Domaine : films

En retraçant les événements de janvier 1972 en Irlande du Nord , Paul Greengrass nous offre un formidable film coup de poing entre reconstitution historique et documentaire politique. Poignant.
Dimanche 30 janvier 1972, Derry, Irlande du Nord. Ivan Cooper, député catholique, organise une marche pacifique revendiquant l'égalité des droits civiques entre catholiques et protestants. Malgré ses négociations avec les autorités unionistes et l'armée britannique, le rassemblement dégénère. Le bilan est lourd : treize personnes sont tuées par l'armée. Ce "Dimanche Sanglant" restera gravé dans les mémoires et marquera le début de la guerre civile en Irlande du Nord.
Le film commence par deux conférences de presse symétriques : celle d'Ivan Cooper, bien décidé à manifester pacifiquement, et celle du commandant des parachutistes britanniques, résolu à empêcher toute manifestation contestant les autorités en place. Deux logiques opposées ne pouvant mener inexorablement qu'au clash final.
C'est dans cette spirale infernale que nous entraîne Paul Greengrass, caméra à l'épaule, pendant presque deux heures. La pellicule 35mm au grain rugueux nous permet de capturer les moindres émotions, hé-

sitations ou doutes des comédiens en tout point parfaits - une grande partie des figurants ont réellement vécu ces évènements.
Le style employé emprunte donc beaucoup à la technique du documentaire et à celle d'un travail journalistique à la manière d'un reportage de guerre. Le propos et les moyens utilisés s'y prêtent parfaitement.
Ce parti pris formel n'est absolument pas gratuit, on l'aura compris. Tout est là pour nous décrire le terrible engrenage de la violence, violence qui au final ne pourra que s'auto-engendrer. L'actualité, jour après jour, nous rappelle à quel point nous restons concernés.
C'est pour tout cela que ce film agit comme un électrochoc politique et nous remue jusqu'au fond des tripes.

Domaine : bandes-dessinées

Quels fous sommes-nous de craindre la mort plus que le pouvoir des songes ? Accompagnez donc le maître des rêves dans son royaume... Votre perception des choses pourrait bien changer à la lecture de cette œuvre riche et bouleversante...
Qui n'a jamais rêvé de prendre la mort au piège ? De la garder prisonnière ? En son pouvoir ? En voulant réaliser cette folie, un groupe d'apprentis sorciers capture le jeune frère de cette dernière. Dream, Morpheus, the Sandman... Emprisonné durant soixante-dix ans sous une cloche de verre, le maître des rêves rumine sa vengeance... Mais quand la libération survient, il lui faut reconstruire son royaume en ruines et récupérer les trois attributs de son pouvoir : son masque, sa bourse et son rubis. Commence alors une quête, entre réalité et monde du rêve... A la recherche de son pouvoir perdu, Dream devra rapprocher un petit privé d'un ancien amour aux frontières de la mort, mettre fin aux agissements d'un fou dangereux, et rencontrer le maître des enfers en personne ! Une aventure sombre, brumeuse et poétique commence dans ce premier volume de la saga culte de Neil Gaiman.
Neil Gaiman est surtout connu en France pour ses romans (Neverwhere , American gods , Stardust ; De bons présages - avec Terry Pratchett) mais ses premières créations étaient des bandes dessinées. Après une première tentative d'édition (de feu les éditions Le Téméraire), Delcourt nous livre enfin l'intégralité de son oeuvre majeure, une série décalée et créative, pleine de poésie sombre et de personnages hors du commun. Les aventures de ces immortels qui président chaque instant de nos vies sont touchantes d'humanité et de sensibilité. Dream, le maître des rêves, surpuissant mais fragile, mélancolique, parfois puéril... Death, sa sœur aînée, lolita gothique pleine de charme et de peps... Delirium, fragile, à fleur de peau... Desire, avenant, souriant, séduisant mais terriblement égoïste... Et aussi Destruction, Despair, Destiny... Tous profondément humains...

Un lecteur non averti pourra être dérouté par le découpage et la mise en couleur de cette bande dessinée atypique, et par les fréquents changements de dessinateurs, mais il serait bien dommage de ne pas passer outre cette première impression. Sandman est une œuvre d'une prodigieuse richesse, un concentré de légendes de tous temps et de toutes cultures, où chaque artiste apporte sa vision du personnage... Un monument de la bande dessinée, déjà culte dans son pays d'origine... A découvrir de toute urgence !

Domaine : théâtre

"Malgré le dégoût, malgré l'horreur, malgré les erreurs féroces et les crimes, je vais à l'enfant, je prends le nouveau-né : il est l'espoir misérable de l'avenir humain." A l'image de Romain Rolland se contraignant à choisir le communisme, unique rempart, selon lui, au fascisme, les personnages du Jour du destin ont tous été happés par le mouvement impitoyable de l'histoire du XXe siècle et de ses idéologies exterminatrices. Et onze ans après la fin de la Guerre d'Espagne, les cicatrices n'ont pas fini de se refermer. Sans doute ne se renfermeront-elles jamais, d'ailleurs. Car ce qui stupéfie dans cette histoire adaptée d'un épisode de La nuit du décret , le roman de Michel del Castillo, c'est le renouvellement d'un art que beaucoup croyaient disparu : la tragédie.
Prisonnier des méandres hérités d'une jeunesse passée sur le front, Avelino Pared, un agent charismatique de la police secrète de Franco, réussit enfin à appréhender Ramon Puig, l'un des fers de lance de la résistance anarchiste. Quelques heures plus tard, pris entre un amour filial pour son supérieur et une admiration respectueuse pour ce prisonnier doté d'une érudition rare, un jeune inspecteur, Laredo, débarque dans la brigade. Plus les jours, les semaines et les mois se succèderont, plus il découvrira, grâce à l'interrogatoire du prévenu, à quel point l'âme et l'honneur de son idole ont été corrompus par le temps. Idéaliste, jusqu'au-boutiste, il ne pourra s'empêcher de lui renvoyer le dégoût que ses méthodes lui inspirent. Toute cette haine projetée à la face de Pared s'apparente aux scènes de crachat, celles-là même qui conduisent au point de non-retour entre un père et un fils dans le théâtre antique.
Cette atmosphère pesante est parfaitement rendue par la mise en scène de Jean-Marie Besset et de Gilbert Désveaux. Le bureau où se déroule toute cette dramatique ressemble, en effet, à une ancienne église où l'on entendrait presque le bruit continu et stressant des gouttes qui perlent au plafond et qui viennent s'écraser sur le sol. Au centre, un escalier descend jusqu'à la geôle de Puig où l'on ne voit ce qui s'y passe que par ombre chinoise. Tout est prêt pour un affrontement placé sous le signe de la confession. Dans le rôle de Pared, Michel Aumont, toujours aussi impressionnant, est la pierre angulaire

du duel qu'il livre successivement à Puig, Christophe Malavoy, et à Laredo, Loïc Corbery - difficile de ne pas penser ici à la confrontation entre Raskolnikov et le juge Pétrovich dans Crime et châtiment . Poussé par un texte sobre, leur jeu discipliné et irréprochable s'apparente à celui des films de Jean-Pierre Melville : sans superflu, sans incantation inutile. Un brin de lyrisme aurait toutefois pu élever cette pièce au rang de chef-d'œuvre. Bien entendu, les comédiens n'y sont pour rien.

CORPUS ANGLAIS ISSU DE MDSD

Domaine : books

Charlie Bone isn't your typical nine-year-old boy. Although he didn't know that himself. But when photographs get mixed up and he ends up with a picture of a man and a little girl that talks to him, he knows everything that was normal about his life is gone forever. Tracking down the picture, Charlie finds out it belongs to Miss Ingledew, a woman who owns a bookshop and has been looking for her missing niece for years. She gives Charlie a mysterious package (that he can't open) and sets him on his journey that will change everything in his world. As it turns out, Charlie is one of the descendants of the Red King, a mysterious person who had wondrous magical powers who went into hiding. The Red King's children have been equally divided between good and evil ever since. At Bloor's Academy, Charlie finds friends and enemies, and the challenge of his lifetime as he tracks down secrets others don't want revealed.
Jenny Nimmo is the author of five books in the Charlie Bone series and has written several other children's books and fantasies.
Many people compare Charlie Bone to Harry Potter, and that's a good comparison. But the two series are different. Harry's family is really non-supportive, but Charlie has a loving mother, a doting grandmother, and Uncle Paton, who turns out to be something of a hero. The book is a fun, fast-paced whirlwind of mysteries, magic, and friendship. Even at 400 pages, it feels like it's over much too soon.
There are a few jarring instances where the point-of-view was disconnected, moving from one character and one scene into another without warning. They were easily overlooked due to the pacing of the story, but noticeable all the same.
Fans who are waiting for the next Harry Potter book who haven't tried the Charlie Bone series are encouraged to do so. Charlie's story, although similar, has much to offer in reading excitement.

Domaine : dvds

This entire movie could have run in only 20 minutes and you wouldn't miss anything and might even enjoy it. Unfortunately it ran 88 minutes too long and I couldn't wait for it to end. I saw it in the theater and the people all around me were all complaining how boring it was. At least a quarter of them walked out before the end. It's that bad. It's a shame, I love a good suspense/horror movie and the decent actors in this movies were waisted.

Domaine : electronics

I had DCS-900W from the same product line. It craped out after 9 months, and now I am on the fifth call to try to resolve the issue and get it shipped back. I don't know where their support is located, but the line quality is very bad, they ran the same drill over and over, but just don't want to issue you an order for you to ship back.

This is the first D-Link product I own and probably the last. The resolution looks crappy even when it's new, and now it's totally dead, even reset won't work on it.

Domaine : kitchen and houseware

This cover is very rich looking. The fabric is wonderful. I have a kitten that was pulling himself onto the bed and snagging it, but I just snipped the snags and you cannot tell. It is easy to wash and looks great on the bed.

B

CRÉATION DE CLUSTERS

Les clusters obtenus en projetant le plus grand vocabulaire de mots sont présentés ci-dessous. Comme expliqué à la section 4.3, les grands clusters ont une faible pureté alors que les petits clusters ont une faible mixité. Les clusters intéressants pour le transfert d'information polaire sont donc les clusters de moyenne taille.

```
at_least least difficult highly highly_recommend i_highly
pleased disappointed very_disappointed very_pleased expected
i_expected recommend would_recommend poor poor_quality
disappointed_with not_recommend would_not an_excellent
excellent second the_second excited i_was a_wonderful
wonderful acting the_acting quality attempt would_highly
anyone anyone_who happy very_happy a_very very for_anyone
script the_script was_very very_poor disappointed_in
highly_recommended recommended his of_his fails zero a_must
favorite my_favorite must_have this_set beautiful worse felt
man dull i_could is_very both this_knife three of_my movie
movie_is was_really young trying outstanding collection
is_excellent durable one_star performance john sharp
story_of and_was weak a_classic classic he_is enjoyable
allows wanted_to is_wonderful and_he performances found_the
definitely_not amazing plot experience poorly who_has people
keeps gain i_purchased and_his that_was ridiculous
beautifully enjoy more shows the_old about_the war overall

difficult_to easy easy_to spout the_spout impossible
impossible_to to_clean clean_up ended_up a_little little
plastic the_plastic cleans cleans_up fits ended the_unit
unit machine this_machine to_use thing_is very_easy knife
tried use make and_easy to_work work clean_and try size
it_back than_a table value evenly seemed seemed_to steam
having_to leaks makes come_out bought_this i_bought kitchen
cookware before_it burned i_used large comes when_it become
perfectly easier feature cooking heat food leak is_easy
time_i the_problem stuck clean i_use have_used bit_of solid

don't_waste waste_your of_money waste_of a_waste waste
ashamed be_ashamed money worst money_on your_money save
save_your the_money and_money my_money money_back of_time
what_a your_time a_disappointment disappointment junk
of_junk not_waste leaked zero_stars product this_product
```

piece_of dont_waste paid total_waste a_total my_time
destroyed vin garbage sorry_i sorry horrible trash pathetic
threw wasting worst_movie useless of_garbage no_stars
wasted dumb not_worth doesn't_do boring buy_a thinking
awful joke is_<num> bad

then then_it after after_the months <num>_months weeks
stopped stopped_working first the_first after_about i_loved
loved broke_after after_<num> after_using broke it_worked
worked started started_to after_less working even_after
month new_one died after_one months_of it_started back
few_uses less_than mechanism disappointing a_new again
lasted first_time replace avoid_this it_broke process soon

read read_the book the_book a_favor favor_and based
based_on favor reviews the_reviews guess do_yourself
it_then reviews_here book_then returning returning_it must
bad_movie probe thermometer reviewers sending_it enjoyed
carefully skip buying_this a_bad to_try you_will badly
this_unit a_different cheaper the_same mediocre huge
powerful purchasing wrote information is_terrible instead

we_move wrath story_begins she_then hostile his_partner
feelings_for able able_to of_our our a_replacement
replacement past the_past we_had family husband rating
we_thought who year failed baby ways for_his we_use
nicely wife day care cooks quiet age followed either space

i_love love love_it love_this love_the in_love love_story
absolutely_love love_them love_these love_and will_love
so_easy fantastic you'll_love sweet favorites musical
be_disappointed absolutely fabulous perfect

the_worst i've i've_never a_long long_time that_i've
<num>_times times i've_owned months_now far most the_most
now_and worst_movies of_times always weight fact_that
every my butter

a_lot lot lot_of i_thought thought i_guess it_didn't
thought_it thought_that tough end rock doesn't_work regular
special son idea cuts maybe fun aid

a_great great_movie great is_great great_for great_story
great_product the_great works_great great_job one_day
great_price great_and great_little worked_great
great_value great_film it's_great every_day is_perfect

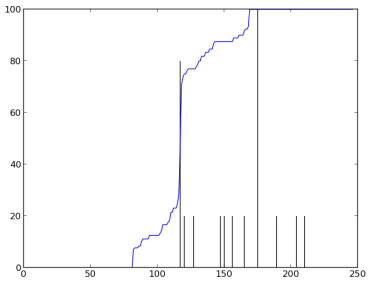
```
as_well well_as very_well well well_worth well_made
well_done is_well and_well well_written how_well so_well
well_that can_use wood useful works and_very simple

minutes <num>_minutes hours <num>_hours first_<num>
minutes_i <num>_seconds seconds hours_of <num> it_took
took long_and about_<num> bored jack takes slow

best the_best is_one one_of best_of best_movies for_best
by_far ever oscar at_best superb song hilarious season
and_clean awesome part_of

worry worry_about don't i_don't to_worry don't_buy
don't_have you_don't bother don't_bother need need_to
know_how why two_months

i_turned turned_it off off_i broke_off dangerous off_after
off_a taylor the_blade flimsy cheesy unfortunately turned

better better_than compared compared_to are_better
even_better much_better than_this better_off do_better
than_i so_much cheap

buy_this not_buy do_not buy go_ahead plugged going_back
<num>_uses breaks this_ went_to  warranty

get to_get get_a what not_what what_was not_get get_what
and_you course pay lucky

only the_only it_only not_only only_<num> redeeming problem
was_only only_complaint only_wish otherwise only_reason

should should_have didn't i_didn't i_should hard_to
mistake terrible back_and  is_hard going

i_like like like_a looks_like more_like look_like you'll
krups nothing  looked

a_bunch bunch_of bunch and_some some along along_with
extras ok

is_not not does_not did_not was_not not_work not_fit
not_enough not_very

is_still still still_a waiting waiting_for today strong
```

```
also is_also also_the are_also comes_with <year>

a_bit bit you_can larger add complaints

want you_want see_a you_really

all at_all for_all all_over

no no_problems effort

i_had and_i i

supposed supposed_to

even not_even
```

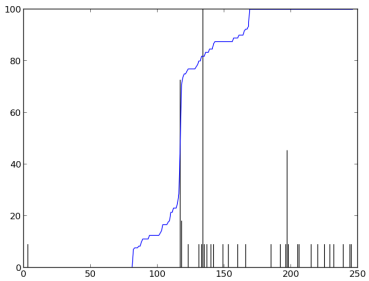
C

MOTS PIVOTS ET COOCCURRENCES

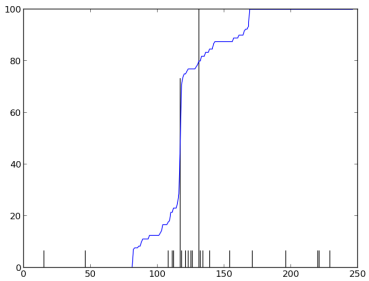
Les mots sélectionnés comme pivots pour la sélection semi-supervisée des marqueurs multi-polaires à la section 5.3 sont présentés ci-dessous. Les graphiques associés représentent les cooccurrences entre mots pivots. La courbe bleue, identique sur chaque schéma, montre la polarité de l'ensemble des marqueurs, triés du plus négatif au plus positif. On observe un décrochage au milieu de la courbe correspondant au passage des pivots positifs (positivité entre 0 et 0.25) et des pivots négatifs (positivité entre 0.75 et 1). Les traits verticaux représentent quant à eux le pourcentage de textes contenant le mot pivot considéré qui contiennent également un autre mot pivot. A l'emplacement du mot pivot considéré, il y a donc nécessairement un trait vertical complet. Ce trait permet ainsi de repérer la positivité du mot pivot considéré. Le pourcentage de cooccurrence avec le mot "*I*" ("je" en anglais) est également représenté. Ce mot possède une positivité de 0.5. Il est donc neutre et le trait associé permet de mieux visualiser la séparation entre mots pivots positifs et mots pivots négatifs.



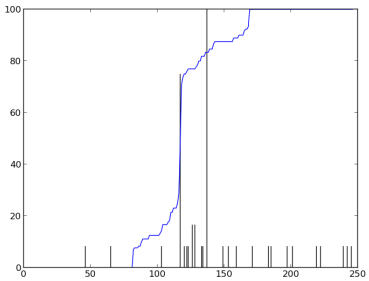
an-outstanding



finest



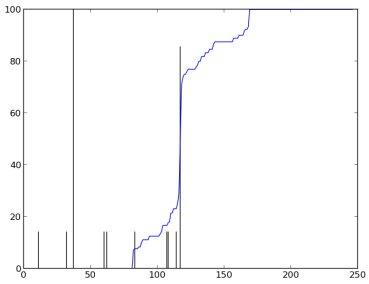
an-amazing



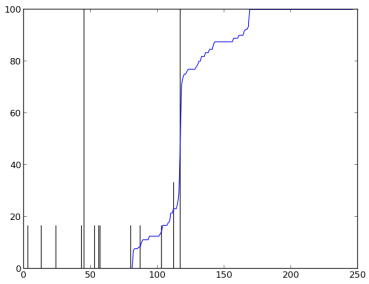
are-worth

On remarque que les mots fortement valués (positivité de 0 ou 1) ne cooccurrent qu'avec d'autres mots de la même polarité. Par contre,

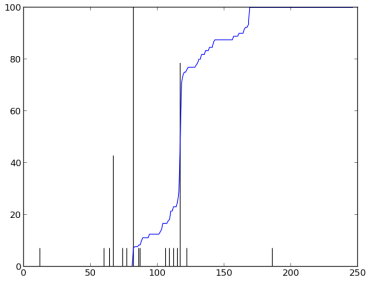
parmi les mots pivots plus nuancés (et donc apparaissant dans des critiques positives comme négatives), certains cooccurrent plusieurs fois avec des mots pivots de la polarité opposée.
De plus, il est flagrant que les pivots positifs ne cooccurrent pas tous entre eux. Il en est de même pour les négatifs. C'est pourquoi nous avons proposé de regrouper les mots pivots en quatre groupes : fortement négatif, négatif, positif et fortement positif.



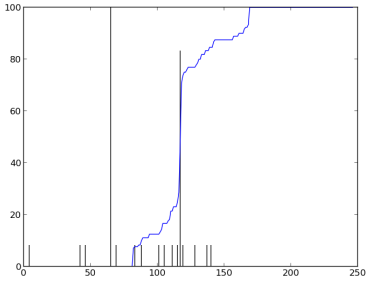
a-refund



destroyed



ashamed



mediocre

```
<num>-<year> a-<year> a-hit a-joke a-place a-refund a-solid
a-step a-weak absolutely-nothing actually-have addicted
adding-to after-about all-are all-things an-amazing
an-incredible an-outstanding and-almost and-completely
and-especially and-family and-money and-pick and-quality
and-starts anniversary any-good appreciate-the are-doing
are-perfect are-worth arrives as-was ashamed asleep at-amazon
aware-of
b bad-the be-ashamed better-as biggest-problem blast bones
but-nothing but-one buying-it
can't-go can't-wait check-it chilling chunks coming-off
complaint-is concept-of considering-the contemporary
continues-to could-use crappy
daddy definitely-a definitely-worth destroyed did-they
didn't-do didn't-like didn't-make dissapointment do-better
doesnt downside dutch
```

enjoys excellent-the expecting-a
far-from far-too features-are feeling-of felt-as felt-that
finest first-saw five-stars for-best for-yourself freely
from-them from-what frustrating fun-to
gave-this give-away gives-the gives-you global go-along
goes-off going-for grew-up grossly
had-never half-hour halfway happened-in has-always has-many
hated-it have-bought hooked how-well huh
i i'd-rather i-hated i-keep i-kept i-returned in-more
in-particular instead-the insult-to irrelevant is-superb
it's-like it's-own it-became
job-as
kidding kudos
loosely love-of loved-this low-price
mediocre most-importantly muscle must-for my-experience
my-family my-head
naturally neighbors no-reason noisy noted-that
old-son only-complaint only-drawback original-and out-some
overrated
paid-for performing performs power-of
really-really reason-to right-off ruins
school-and sharp-and shows-on silent skip-this so-bad
so-excited so-that's something-different sorry-but spend-a
splendid spoiled stars-but stinks sucked superbly
talked tedious terrible-the tested that-seems the-beauty
the-finest the-individual the-many the-opportunity
the-packaging the-reality the-system the-trash there's-nothing
they-even they-used thin-and this-collection this-problem
this-wonderful time-around time-she time-when time-which
timeless times-better to-become to-consider to-drive
to-review to-worry today's top-notch turned-into twists
two-hours two-stars
uncomfortable unnecessary
very-happy very-satisfied
want-more was-being was-definitely was-excellent was-horrible
was-terrible wasn't-the web-site when-in who-had who-will
whole-family will-always with-ease without-being working-on
worthless wrong-this
you'll-find you're-not you-say your-body your-face your-head
your-life

D

EXEMPLES DE MARQUEURS MULTI-POLAIRES

MARQUEURS MULTI-POLAIRES DANS LE CORPUS MDSD

Tous les mots détectés comme marqueurs multi-polaires pour les paires de corpus issues du corpus MDSD et effectivement utilisés par le classifieur BoosTexter comme classifieurs faibles sont listés ci-dessous. Sur chaque ligne est présenté l'écart de positivité entre les deux domaines, le marqueur, la positivité dans le premier domaine ainsi que la positivité dans le second domaine (dans l'ordre du titre de sous-section).
Les améliorations constatées lors des expériences sont parfois modestes, mais elles sont à mettre en regard du faible nombre de marqueurs détectés pour certaines paires, et donc de la petite quantité de modifications apportées aux corpus initiaux.

DVD - Books

0,60	profound	0,20	0,80
0,57	one-with	0,67	0,10
0,53	asking	0,17	0,69
0,47	the-house	0,56	0,08
0,46	drug	0,63	0,17
0,45	finish	0,61	0,17
0,43	it-if	0,57	0,13
0,43	to-like	0,43	0,00
0,42	accept	0,53	0,11
0,36	no-matter	0,39	0,75
0,35	blow	0,55	0,90
0,35	not-very	0,42	0,06
0,35	british	0,40	0,75
0,34	be-better	0,45	0,11
0,34	used-to	0,65	0,32
0,34	source	0,35	0,69
0,33	life-is	0,64	0,31
0,33	universal	0,57	0,90
0,31	i-recommend	0,56	0,87
0,31	this-i	0,42	0,11
0,30	honestly	0,50	0,20
0,30	the-price	0,63	0,33
0,29	just-so	0,36	0,08
0,28	know-what	0,33	0,61

0,28	some-good	0,28	0,00
0,27	helps	0,56	0,83
0,27	found-myself	0,32	0,59
0,27	teacher	0,82	0,55
0,26	humor	0,51	0,77
0,25	sad	0,49	0,23
0,25	awesome	0,69	0,94
0,25	regret	0,88	0,64
0,24	artist	0,62	0,38
0,24	i-did	0,49	0,25
0,24	minutes	0,30	0,54
0,23	weak	0,27	0,04
0,23	care-about	0,33	0,10
0,23	to-begin	0,40	0,17
0,23	beautifully	0,77	1,00
0,23	trilogy	1,00	0,77
0,23	starting	0,53	0,76
0,23	not-just	0,79	0,57
0,22	loved-this	0,78	1,00
0,22	too-long	0,38	0,16
0,21	awful	0,04	0,25
0,19	way-too	0,26	0,07
0,18	kidding	0,18	0,00
0,17	dull	0,17	0,00
0,16	one-star	0,16	0,00
0,15	very-disappointed	0,15	0,00
0,15	boring-and	0,15	0,00
0,12	enjoyed-this	0,88	1,00

Books - Electronics

0,62	been-able	0,92	0,31
0,60	the-house	0,08	0,68
0,57	someone-who	0,35	0,92
0,55	loved-it	0,95	0,40
0,55	internal	0,82	0,27
0,54	easier	0,23	0,77
0,54	one-with	0,10	0,64
0,52	thin	0,21	0,73
0,49	no-matter	0,75	0,26
0,48	work-on	0,82	0,33
0,46	starting	0,76	0,30
0,46	flat	0,04	0,50
0,45	the-price	0,33	0,79
0,45	universal	0,90	0,45

0,44	life-is	0,31	0,75
0,44	it-seems	0,23	0,67
0,42	are-much	0,08	0,50
0,42	loved	0,73	0,30
0,42	straight	0,79	0,38
0,40	blow	0,90	0,50
0,40	more-on	0,20	0,60
0,38	period	0,68	0,30
0,37	love-this	0,91	0,54
0,37	makes-it	0,85	0,48
0,36	used-to	0,32	0,68
0,36	very-little	0,22	0,58
0,35	the-internet	0,12	0,46
0,32	is-just	0,24	0,56
0,32	process	0,71	0,39
0,32	be-better	0,11	0,43
0,32	care-about	0,10	0,42
0,31	experience	0,66	0,35
0,30	this-i	0,11	0,41
0,29	weak	0,04	0,33
0,29	doesnt	0,00	0,29
0,29	basically	0,19	0,49
0,29	know-what	0,61	0,32
0,29	for-all	0,72	0,43
0,28	lose	0,68	0,40
0,27	photos	0,48	0,74
0,26	info	0,24	0,50
0,26	soon	0,76	0,50
0,25	i-did	0,25	0,50
0,25	fan	0,31	0,56
0,25	stick	0,21	0,46
0,24	ordered	0,71	0,47
0,24	none-of	0,17	0,42
0,24	the-two	0,67	0,43
0,23	lack	0,27	0,50
0,23	not-very	0,06	0,29
0,23	internet	0,26	0,49
0,23	enjoy	0,51	0,73
0,23	while-i	0,27	0,50
0,21	tried-to	0,33	0,12
0,21	great-for	0,92	0,72
0,20	used-in	0,13	0,33
0,18	one-star	0,00	0,18
0,18	address	0,08	0,26
0,16	worst	0,20	0,04

Books - Kitchen

0,80	loved-this	1,00	0,20
0,57	loved-it	0,95	0,38
0,55	the-house	0,08	0,63
0,53	no-matter	0,75	0,22
0,51	easier	0,23	0,74
0,50	soon	0,76	0,26
0,47	experience	0,66	0,19
0,46	flat	0,04	0,50
0,44	this-i	0,11	0,55
0,44	finish	0,17	0,61
0,42	dull	0,00	0,42
0,41	very-little	0,22	0,64
0,39	effort	0,33	0,73
0,39	so-far	0,55	0,94
0,38	stick	0,21	0,59
0,37	the-price	0,33	0,71
0,36	while-the	0,64	0,29
0,35	ordered	0,71	0,36
0,35	starting	0,76	0,42
0,31	loved	0,73	0,42
0,31	want-a	0,30	0,61
0,30	liked-the	0,71	0,41
0,30	enjoy	0,51	0,81
0,29	upon	0,56	0,26
0,28	waiting-for	0,38	0,10
0,28	tried-to	0,33	0,06
0,27	basically	0,19	0,46
0,27	process	0,71	0,44
0,26	absolutely	0,55	0,82
0,26	used-to	0,32	0,58
0,26	totally	0,34	0,60
0,26	once-you	0,90	0,64
0,26	thus	0,53	0,27
0,25	honestly	0,20	0,45
0,24	i-did	0,25	0,48
0,22	for-all	0,72	0,93
0,20	poor	0,28	0,08
0,20	it-if	0,13	0,33
0,18	amazing	0,74	0,92
0,17	a-wonderful	0,95	0,79
0,16	not-very	0,06	0,23
0,13	beautifully	1,00	0,87

DVD - Electronics

0,72	one-day	0,92	0,20
0,67	i-loved	0,76	0,09
0,63	b 0,08	0,71	
0,53	found-the	0,22	0,75
0,50	case-and	0,20	0,70
0,43	look-at	0,60	0,17
0,43	loved	0,73	0,30
0,43	times-and	0,76	0,33
0,41	is-well	0,86	0,45
0,39	times-i	0,80	0,41
0,39	is-still	0,82	0,43
0,38	i-turned	0,00	0,38
0,37	the-video	0,33	0,70
0,36	movie	0,46	0,82
0,34	cant	0,62	0,27
0,34	looks-like	0,16	0,50
0,34	regret	0,88	0,55
0,33	city	0,71	0,38
0,33	biggest	0,31	0,64
0,31	compared	0,50	0,81
0,31	thing-is	0,14	0,45
0,30	serious	0,62	0,32
0,30	are-great	0,50	0,80
0,29	keeps	0,65	0,36
0,28	simple	0,45	0,74
0,28	love-this	0,82	0,54
0,28	hooked	1,00	0,72
0,27	just-not	0,06	0,33
0,26	different	0,59	0,33
0,25	tell	0,54	0,29
0,25	otherwise	0,27	0,51
0,25	comments	0,37	0,62
0,24	times	0,60	0,35
0,23	along	0,68	0,44
0,23	numbers	0,73	0,50
0,22	ship	0,42	0,20
0,14	terrible	0,16	0,02
0,12	stay-away	0,00	0,12

DVD - Kitchen

0,68	i-loved	0,76	0,08
0,55	times-and	0,76	0,21
0,51	times-i	0,80	0,29

0,43	thing-is	0,14	0,58
0,40	close-to	0,79	0,39
0,39	smart	0,89	0,50
0,39	are-great	0,50	0,89
0,38	getting-the	0,80	0,42
0,37	looks-like	0,16	0,53
0,36	single	0,33	0,69
0,35	mad	0,53	0,18
0,32	otherwise	0,27	0,59
0,32	special	0,58	0,89
0,32	loved	0,73	0,42
0,32	simple	0,45	0,77
0,30	actual	0,33	0,64
0,29	look-at	0,60	0,31
0,29	biggest	0,31	0,60
0,29	the-most	0,50	0,79
0,29	book	0,36	0,64
0,28	awesome	0,69	0,97
0,28	above	0,74	0,46
0,27	just-not	0,06	0,33
0,27	enough-to	0,38	0,65
0,27	really-is	0,87	0,60
0,26	times	0,60	0,34
0,25	serious	0,62	0,37
0,25	stay	0,30	0,54
0,25	dull	0,17	0,42
0,25	fails	0,15	0,40
0,25	holes	0,17	0,41
0,24	shape	0,86	0,62
0,23	compared	0,50	0,73
0,23	i-decided	0,69	0,46
0,23	is-still	0,82	0,59
0,23	although-i	0,76	0,53
0,22	his	0,59	0,81
0,22	not-only	0,60	0,82
0,22	loves	0,74	0,96
0,21	my-favorite	0,69	0,89
0,19	regret	0,88	0,69
0,19	a-total	0,29	0,10
0,16	terrible	0,16	0,00
0,13	highly-recommended	1,00	0,87
0,13	outstanding	0,87	1,00
0,13	disappointing	0,13	0,00

Electronics - Kitchen

0,55	suction	0,92	0,38
0,53	places	0,73	0,20
0,52	floor	0,93	0,41
0,51	delivery	0,80	0,29
0,48	drops	0,10	0,58
0,44	these-things	0,23	0,67
0,43	keeps	0,36	0,79
0,40	every-time	0,18	0,59
0,40	worked-for	0,10	0,50
0,37	good-for	0,69	0,32
0,36	walls	1,00	0,64
0,36	clear	0,73	0,37
0,33	expected	0,77	0,44
0,31	the-top	0,74	0,43
0,31	obviously	0,22	0,53
0,31	buy-one	0,56	0,87
0,31	essential	0,60	0,91
0,31	and-got	0,04	0,35
0,30	often	0,37	0,68
0,30	okay	0,33	0,64
0,30	came-out	0,75	0,45
0,30	so-good	0,42	0,71
0,28	considering	0,64	0,36
0,28	and-bought	0,19	0,47
0,27	and-go	0,14	0,42
0,26	business	0,36	0,10
0,26	them-but	0,14	0,40
0,25	i-put	0,21	0,46
0,25	fails	0,15	0,40
0,25	no-way	0,04	0,29
0,24	thinking	0,50	0,26
0,23	plenty	0,77	1,00
0,23	the-rest	0,20	0,43
0,23	beware	0,10	0,32
0,22	takes-a	0,82	0,60
0,22	this-thing	0,18	0,39
0,21	everyday	0,90	0,69
0,19	so-far	0,75	0,94
0,17	months-now	0,83	1,00
0,15	great-product	0,80	0,95
0,13	garbage	0,13	0,00
0,12	poorly	0,12	0,00
0,11	disappointing	0,11	0,00

MARQUEURS MULTI-POLAIRES DANS LE CORPUS DEFT

Certains mots détectés comme marqueurs multi-polaires pour l'extrait du corpus français DEFT et effectivement utilisés par le classifieur BoosTexter comme classifieurs faibles sont listés ci-dessous. Sur chaque ligne est présenté l'écart de positivité entre les deux domaines, le marqueur, la positivité dans le domaine des films ainsi que la positivité dans le domaine des jeux vidéos. Comme le corpus français utilisé est de taille plus faible que le corpus anglais, les marqueurs sélectionnés sont plus sujet au biais de corpus. Il semble par exemple que la bande dessinée batman ait été bien adaptée en film mais trahie dans les jeux vidéos. Pour cette même raison, le nombre de marqueurs sélectionnés est bien plus important que pour les corpus anglais. Aussi, ne sont présentés ici que les marqueurs ayant un écart de positivité de plus de 0.5.

0,94	tarantino	0,94	0,00
0,81	batman	0,91	0,10
0,81	vole	0,90	0,09
0,78	park	0,85	0,07
0,77	ne-existe	1,00	0,23
0,75	obtenu	0,92	0,17
0,67	le-aide	0,20	0,87
0,67	gentils	0,00	0,67
0,66	cruel	0,75	0,09
0,66	pervers	0,73	0,07
0,66	neurones	0,15	0,81
0,65	road	0,73	0,08
0,64	michael	0,64	0,00
0,63	honorable	0,10	0,73
0,63	scénario-qui	0,19	0,82
0,63	ludique	1,00	0,37
0,63	spider	0,33	0,96
0,63	répétition	0,89	0,27
0,63	de-aborder	0,18	0,81
0,62	les-filles	0,70	0,08
0,62	efficaces	0,27	0,89
0,61	rebondissements	0,20	0,81
0,61	rome	0,70	0,09
0,61	la-gueule	0,20	0,81
0,61	du-héros	0,00	0,61
0,59	unité	0,27	0,86
0,58	chanson	0,68	0,10
0,58	vers-les	0,80	0,22
0,58	une-poignée	0,77	0,19
0,57	substance	0,18	0,75
0,57	allemands	0,10	0,67

0,56	poignée-de	0,79	0,22
0,56	les-enfants	0,64	0,08
0,56	le-excellent	0,29	0,86
0,56	se-prend	0,17	0,73
0,56	chambre	0,73	0,17
0,56	des-morts	0,26	0,82
0,55	rien-que	0,25	0,80
0,55	au-sérieux	0,18	0,73
0,55	justice	0,76	0,21
0,55	héritage	0,45	1,00
0,54	enfants	0,73	0,18
0,54	celle-qui	0,75	0,21
0,54	pauvres	0,77	0,23
0,54	poignée	0,79	0,25
0,53	tigre	0,82	0,29
0,53	déçoit	0,00	0,53
0,53	accent	0,20	0,73
0,53	et-fait	0,93	0,40
0,53	cette-année	0,29	0,81
0,53	sont-à	0,90	0,37
0,53	essaie-de	0,71	0,19
0,53	lenteur	0,71	0,18
0,53	de-créer	0,29	0,81
0,52	pour-mettre	0,25	0,77
0,52	minute	0,67	0,14
0,52	trace	0,65	0,13
0,52	chaotique	0,80	0,28
0,52	la-sauce	0,23	0,75
0,52	confusion	0,77	0,25
0,52	intéressante	0,00	0,52
0,52	gang	0,13	0,64
0,51	un-remake	0,21	0,73
0,51	respect	0,83	0,32
0,51	au-sens	0,88	0,36
0,51	anglaise	0,36	0,88
0,51	celles-qui	0,90	0,39
0,51	les-parents	0,69	0,18
0,51	des-enfants	0,79	0,29
0,50	descente	0,69	0,19
0,50	les-bonnes	0,07	0,57
0,50	aux-personnages	0,30	0,80
0,50	de-effets	0,19	0,69
0,50	les-gens	0,70	0,20
0,50	fbi	0,00	0,50
0,50	adultes	0,67	0,17
0,50	la-façon	0,33	0,83

```
0,50  un-fait          0,92  0,42
```

E

PHRASES ANNOTÉES LORS DE LA CAMPAGNE D'ÉVALUATION

Quelques exemples de phrases utilisées lors de la campagne d'annotation (voir chapitre 6) sont présentées ci-dessous. Chacune contient donc un marqueur multi-polaire potentiel. Ce marqueur est ici signalé entre deux balises bien que lors de l'annotation, cette indication ne figurait pas. Pour chaque phrase, le domaine ainsi que la polarité de la critique originale est également indiquée, suivis du nom du produit considéré.

```
books : negative
Lost in a Good Book (Thursday Next Novels)
<MPW>comfortable</MPW>
And just when one is comfortable with Thursday's role as a
"SpecOps Litratech", and that whole milieu, she's thrown
into an entirely new one as a member of "Jurisfiction", a
kind of police comprised of book characters who move around
in different literary works and maintain order...

books : negative
Expanded Universe
<MPW>no matter</MPW>
Though I respect _Starship Troopers_, it's never going to be
my favorite Heinlein novel no matter how many times we
quibble over the precise definition of "fascism" -- and
I'm not going to have much respect for the nonfiction in
this collection.

books : negative
The Complete Clutter Solution: Organize Your Home for Good
<MPW>return</MPW>
Get it at the library and then return it after you don't read it.

books : negative
Kingdom Come
<MPW>return</MPW>
Wonder Woman approaches Superman in his self-imposed exile
in the nuclear wastleland that is Kansas and asks him
to return and lead the former heroes in returning order.

kitchen : negative
Black & Decker FP1500 Power Pro II food processor
```

<MPW>loved it</MPW>
I used it the first time to grate cheese and loved it.

kitchen : negative
Dyson Root 6
<MPW>pain</MPW>
First off, having to press the button continuously to make it work is a pain.

kitchen : positive
Henckels Four Star 6-Inch High Carbon Stainless Steel Utility/Sandwich Knife
<MPW>comfortable</MPW>
A good knife but Quality Control was poor The knife is solid and very comfortable in hand, however, when I got it new, the blade is slightly bent.

TABLE DES FIGURES

FIGURE 1 Haut : Évolution du nombre de clusters en fonction du seuil k. Bas : Répartition de la taille des clusters pour différentes valeurs de k pour un lexique de 100 mots. Petits clusters : de 2 à 4 mots ; moyens clusters : de 5 à 19 mots ; grands clusters : 20 mots ou plus. 33

FIGURE 2 Évaluation manuelle de la polarité des mots constituant le grand cluster (gauche) et les clusters de taille moyenne (droite) obtenus lors de la projection du vocabulaire de 100 mots. 37

FIGURE 3 Évaluation à l'aide de SentiWordNet de la polarité des mots constituant le grand cluster (gauche) et les clusters de taille moyenne (droite) obtenus lors de la projection du vocabulaire de 100 mots. 38

FIGURE 4 Comparaison de la pureté calculée des clusters en utilisant à la fois les annotations des corpus source et cible ou bien seulement celles du corpus source (cas réel). 40

FIGURE 5 Nombre de mots changeant de polarité parmi les mots sélectionnés par boostexter par tranche de 100 mots 50

FIGURE 6 Procédure de détection des marqueurs multipolaires à l'aide d'une collection de mots pivots. 53

FIGURE 7 Procédure d'épuration de la collection de mots pivots. 54

FIGURE 8 Phrases fournies en exemple aux annotateurs. 61

FIGURE 9 Quelques récompenses pour les annotateurs courageux. 62

FIGURE 10 Différentes échelles d'interprétation du coefficient κ. D'après Fort [2012] 66

FIGURE 11 Écart de positivité selon le score calculé au niveau des textes sur le corpus et le score phrastique calculé grâce aux annotations entre les domaines *Books* et *Kitchen*. 72

FIGURE 12 Écart de positivité selon le score calculé au niveau des textes sur le corpus et le score phrastique calculé grâce aux annotations entre les domaines *DVDs* et *Electronics*. 72

FIGURE 13 Exemple de profil de répartition des occurrences d'un marqueur dans les phrases positives, négatives et neutres. Ici, le marqueur apparaît beaucoup dans des phrases négatives, un peu dans des phrases positives et presque pas dans des phrases neutres. 75
FIGURE 14 Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil uniquement neutre. 81
FIGURE 15 Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil mélangeant les trois classes. 83
FIGURE 16 Comportement globalement polaire par opposition à un comportement globalement neutre en raison d'un profil mélangeant les classes positives et négatives. 85
FIGURE 17 Comportement globalement positif par opposition à un comportement globalement négatif. 88
FIGURE 18 *Accuracy* pour un classifieur BoosTexter entraîné sur un domaine source et testé sur un domaine cible en français (DEFT). 100
FIGURE 19 *Accuracy* pour un classifieur BoosTexter entraîné sur un domaine source et testé sur un domaine cible en anglais (MDSD) ; D : DVDs, B : books, E : electronics, K : kitchen. 101
FIGURE 20 *Accuracy* pour un classifieur SVM entraîné sur un domaine source et testé sur un domaine cible en français (DEFT). 101
FIGURE 21 *Accuracy* pour un classifieur SVM entraîné sur un domaine source et testé sur un domaine cible en anglais (MDSD) ; D : DVDs, B : books, E : electronics, K : kitchen. 102
FIGURE 22 Détection des marqueurs multi-polaires entre les sous-parties thématiques du corpus d'entraînement. 108
FIGURE 23 Processus de création de plusieurs classifieurs thématiques en particularisant le corpus d'entraînement. Version suppression totale, suppression partielle et différenciation partielle. 110

FIGURE 24 Processus de particularisation du corpus d'entraînement pour la version différenciation concomitante. Chaque mot apparaissant dans une sous-partie thématique du corpus et détecté comme marqueur multi-polaire entre cette sous-partie et les autres est modifié en "mot-domaine". Si le mot n'est pas détecté comme marqueur multi-polaire pour la sous-partie dans laquelle il se trouve, il n'est pas modifié, même s'il est détecté comme marqueur multi-polaire pour une autre sous-partie. 111

FIGURE 25 Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus multi-domaines français (DEFT) et anglais (MDSD). 112

FIGURE 26 Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus en domaine ouvert français (DEFT) et anglais (MDSD et Tweets). La séparation en sous-domaines thématiques a été effectuée par LDA. 123

FIGURE 27 Comparaison des différentes techniques d'intégration des marqueurs multi-polaires sur le corpus de tweets. 124

FIGURE 28 Deux corpus d'entraînement différents sont utilisés. *Tweet* contient l'intégralité des tweets tandis que *Tweets75* contient uniquement ceux qui sont focalisés sur un seul thème. 124

FIGURE 29 F-mesure moyenne en fonction de minOcc pour plusieurs types de mixages sur le corpus de tweets. Méthode : suppression totale. 125

FIGURE 30 F-mesure moyenne en fonction de minOcc pour plusieurs combinaisons de pVal et minDiff sur le corpus de tweets. Méthode : suppression totale ; Mixage 3. 128

FIGURE 31 Notre participation à SemEval parmi celles des autres participants. Les barres bleues sont placées dans la catégorie contraint. 129

Figure 32 Résultats mentionnés par les différentes équipes dans leurs articles avant l'adjonction de lexiques. Rouge : Maximum d'entropie ; orange : SVM ; vert : Naives Bayes ; bleu : Boosting 130

LISTE DES TABLEAUX

Table 1 Résultats pour un classifieur entraîné sur *DVDs* et testé sur *kitchen* 26
Table 2 Clusters obtenus avec le vocabulaire de 100 mots 34
Table 3 Scores d'exactitude obtenus sur une tâche de classification de l'opinion pour différentes tailles de vocabulaire projeté. 34
Table 4 Mixité de clusters de différentes tailles pour une taille de vocabulaire projeté de 100. 36
Table 5 Présentation du corpus anglais MDSD. Les tailles des vocabulaires et des textes sont exprimées en nombre d'unigrammes et bigrammes. 47
Table 6 Séparation manuelle des textes du corpus DEFT en différents domaines. 48
Table 7 Présentation du corpus français issu de DEFT. Les tailles des vocabulaires et des textes sont exprimées en nombre d'unigrammes et bigrammes. 48
Table 8 Pourcentage de présence de cinq exemples de marqueurs dans les critiques positives pour deux domaines. 49
Table 9 Mots utilisés par un classifieur qui sont détectés comme fortement multi-polaire (différence de positivité de plus de 0,5) 51
Table 10 Comparaison des mots sélectionnés par la méthode automatique avec ceux sélectionnés par la méthode supervisée selon deux jeux de mots pivots, présélectionnés en privilégiant soit l'information mutuelle source-cible (S-C), soit l'information mutuelle positif-négatif (P-N). Domaine source : *DVDs* ; domaine cible : *kitchen*. 55
Table 11 Accords calculés sur différents groupes d'annotateurs lorsque tous les choix sont pris en compte (3 classes) avec ou sans pondération ou bien seulement les choix positifs et négatifs (2 classes). Haut : Accord observé (avec plusieurs annotateurs). Bas : κ de Fleiss (avec plusieurs annotateurs). 66

Table 12 Haut : π de Scott entre deux annotateurs pour trois classes. Bas : κ de Cohen entre deux annotateurs pour trois classes. 67
Table 13 Haut : π de Scott entre deux annotateurs pour deux classes. Bas : κ de Cohen entre deux annotateurs pour deux classes. 68
Table 14 Répartition des choix de classe des annotateurs en pourcentage. 68
Table 15 Comparaison du score de positivité calculé à l'aide des étiquettes au niveau du texte par rapport à celui calculé à l'aide des annotations au niveau des phrases. Haut : *I loved*. Bas : *return*. 69
Table 16 Liste des marqueurs choisis. Ils présentent une différence de positivité d'au moins 0,5 entre les deux domaines auxquels ils sont liés. 70
Table 17 71
Table 18 Profils de répartition dans les phrases positives, négatives et neutres des marqueurs utilisés dans des phrases de contextes. 76
Table 19 Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une expression d'opinion. 77
Table 20 Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une propriété digne d'opinion. 78
Table 21 Profils de répartition dans les phrases positives, négatives et neutres des marqueurs désignant une cible d'opinion. 80
Table 22 return, returning, pain et region 82
Table 23 comfortable, easier, no matter et broken 83
Table 24 loved it, loved this, i loved et found the 86
Table 25 expected et today 88
Table 26 89
Table 27 Exemple de changement de polarité dû à l'appartenance d'une des phrases à une description de contexte. 92
Table 28 Exemple de changement de polarité dû à un changement de sens du marqueur multi-polaire. 93
Table 29 Exemple de changement de polarité dû au fait que le marqueur multi-polaire caractérise un objet d'un type différent. 94
Table 30 Exemple de changement de polarité dû au fait que le marqueur multi-polaire caractérise un objet qui s'utilise de manière différente. 95

Table 31 Composition du corpus français utilisé après séparation thématique manuelle des critiques issues du corpus *AvoirAlire* de DEFT 2007. 107
Table 32 Mots les plus représentatifs pour les 5 thèmes trouvés par LDA sur les corpus mixés de *MDSD* (haut) et *DEFT* (bas). 116
Table 33 Table de confusion entre les thèmes détectés par LDA et les séparation manuelles en domaines. Corpus *MDSD* en haut et *DEFT* en bas. 118
Table 34 Table de confusion lorsque l'attribution d'une critique à un domaine est d'au moins 75%. Corpus *MDSD* en haut et *DEFT* en bas. 118
Table 35 Quelques types classiques d'émoticônes. 120
Table 36 Description des thèmes obtenus sur le corpus de tweets. 120

Revue québécoise de linguistique, 1998. URL http://id.erudit.org/iderudit/603144ar.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. Subjectivity word sense disambiguation. In *EMNLP*, pages 190–199, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D09/D09-1020.

R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6 :1817–1853, 2005.

A. Andreevskaia and S. Bergler. Mining wordnet for fuzzy sentiment : Sentiment tag extraction from wordnet glosses. In *EACL*, volume 6, pages 209–216, 2006.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596, 2008.

A. Aue and M. Gamon. Customizing sentiment classifiers to new domains : A case study. In *RANLP*, 2005.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings EMNLP*, pages 355–362. Association for Computational Linguistics, 2011.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19 :137, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2) :151–175, 2010.

F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian. Sentiment analysis : Adjectives and adverbs are better than adjectives alone. In *ICWSM*, 2007.

Edward M Bennett, R Alpert, and AC Goldstein. Communications through limited-response questioning. *Public Opinion Quarterly*, 18 (3) :303–308, 1954.

Emile Benveniste. *Problèmes de linguistique générale I*. Gallimard, 1966.

Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Faiza Gara, Olivier Mesnard, Meriama LaÃ¯b, and Nasredine Semmar. Lima : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *Proceedings of LREC'10*, may 2010. ISBN 2-9517408-6-7.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *24th international conference on Machine learning*. ACM, 2007.

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP*, 2008.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3 :993–1022, 2003.

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128. Association for Computational Linguistics, 2006.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders : Domain adaptation for sentiment classification. In *ACL*, 2007.

John Blitzer, Dean Foster, and Sham Kakade. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.

Yee Seng Chan and Hwee Tou Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 89–96. Association for Computational Linguistics, 2006.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer : Little data can help a lot. *Computer Speech & Language*, 20(4) :382–399, 2006.

Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *EMNLP*, 2009.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :213–220, 1960.

Yan Dang, Yulei Zhang, and Hsinchun Chen. A lexicon enhanced method for sentiment classification : An experiment on online product reviews. In *IEEE*, 2010.

Hal Daumé. Frustratingly easy domain adaptation. In *ACL*, 2007.

Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 241–249. Association for Computational Linguistics, 2010.

K. Denecke. Are sentiwordnet scores suited for multi-domain sentiment classification ? In *ICDIM*, pages 1–6. IEEE, 2009.

Barbara Di Eugenio and Michael Glass. The kappa statistic : A second look. *Computational linguistics*, 30(1) :95–101, 2004.

Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of WSDM*, pages 231–240. ACM, 2008.

Katrin Erk. Vector space models of word meaning and phrase meaning : A survey. *Language and Linguistics Compass*, 6(10) :635–653, 2012.

A. Esuli and F. Sebastiani. Sentiwordnet : A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

Angela Fahrni and Manfred Klenner. Old wine or warm beer : Target-specific sentiment analysis of adjectives. In *Symposion on Affective Language in Human and Machine, AISB Convention*, 2008.

Charnois T. Mathet Y. Rioult F. et Legallois D. Ferrari, S. Analyse de discours évaluatif, modèle linguistique et applications. *Revue des Nouvelles Technologies de l'Information*, E-17 :71–93, 2009.

Jenny Rose Finkel and Christopher D Manning. Hierarchical bayesian domain adaptation. In *NAACL*, pages 602–610. Association for Computational Linguistics, 2009.

Karën Fort. *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris-Nord-Paris XIII, 2012.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

Michel Généreux, Thierry Poibeau, Moshe Koppel, et al. Sentiment analysis using automatically labelled financial news. In *LREC 2008 Workshop on Sentiment Analysis : Emotion, Metaphor, Ontology and Terminology*, 2008.

Eugenie Giesbrecht. Using product review sites for automatic generation of domain resources for sentiment analysis : Case studies. *Methods for the automatic acquisition of Language Resources and their evaluation methods*, page 43, 2010.

S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualisation of sentiment lexicons. *ECAI*, 2010.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification : A deep learning approach. In *ICML*, 2011.

Cyril Grouin, Jean-Baptiste Berthelin, Sarra El Ayari, Thomas Heitz, Martine Hurault-Plantet, Michele Jardino, Zohra Khalis, and Michel Lastes. Présentation de deft'07 (défi fouille de textes). *Actes du troisième DEfi Fouille de Textes*, page 3, 2007.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. Domain adaptation with latent semantic association for named entity recognition. In *NAACL*, pages 281–289. Association for Computational Linguistics, 2009.

R. Gupta and S. Sarawagi. Domain adaptation of information extraction models. *ACM SIGMOD Record*, 37(4) :35–40, 2009.

A. Harb, G. Dray, M. Plantié, P. Poncelet, M. Roche, F. Trousset, et al. Détection d'opinion : Apprenons les bons adjectifs ! *FODOP*, 2008.

Zellig S Harris. Distributional structure. *Word*, 1954.

V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *EACL*, pages 174–181. Association for Computational Linguistics, 1997.

J. Hoffman, K. Saenko, B. Kulis, and T. Darrell. Domain adaptation with multiple latent domains. In *NIPS*, 2011.

Fei Huang and Alexander Yates. Biased representation learning for domain adaptation. In *EMNLP*, pages 1313–1323, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D12-1120.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, volume 7, pages 264–271. Citeseer, 2007a.

Jing Jiang and ChengXiang Zhai. A two-stage approach to domain adaptation for statistical classifiers. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 401–410. ACM, 2007b.

V. Jijkoun, M. Rijke, and W. Weerkamp. Generating focused topic-specific sentiment lexicons. In *ACL*, pages 585–594, 2010. URL http://www.aclweb.org/anthology/P10-1060.

S.M. Kim and E. Hovy. Automatic detection of opinion bearing words and sentences. In *IJCNLP*, pages 61–66, 2005.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis : The good the bad and the omg ! *ICWSM*, 11 : 538–541, 2011.

Klaus Krippendorff. *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA., USA,, 1980.

Klaus Krippendorff. *Content Analysis : An Introduction to Its Methodology, second edition*, chapter 11. Sage, Thousand Oaks, CA., USA,, 2004.

T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. In *EMNLP*, 2004.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. Sentiment analysis with global topics and local dependency. In *Proceedings of the 24nd AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010.

Shoushan Li and Chengqing Zong. Multi-domain sentiment classification. In *ACL*, 2008.

Shoushan Li, Churen Huang, and Chengqing Zong. Multi-domain sentiment classification with classifier combination. *Journal of Computer Science and Technology*, 26 :25–33, 2011.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1) :1–167, 2012.

Riham Mansour, Nesma Refaei, Michael Gamon, Ahmed Abdul-Hamid, and Khaled Sami. Revisiting the old kitchen sink : Do we need sentiment domain adaptation ? In *RANLP*, pages 420–427, 2013.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation : Learning bounds and algorithms. In *COLT*, 2009.

Andrew Kachites McCallum. Mallet : A machine learning for language toolkit. 2002.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop at ICLR*, 2013.

R. Navigli. Word sense disambiguation : A survey. *ACM Computing Surveys*, 2009.

A. Pak and P. Paroubek. Construction d'un lexique affectif pour le franÃ§ais à partir de twitter. In *TALN*, 2010.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10) :1345–1359, 2010.

Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2) :199–210, 2011.

S.J. Pan, X. Ni, J.T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, pages 751–760. ACM, 2010.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. In *ACL*, 2002.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrival*, 2 :1–2, 2007.

G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In *AAAI*, pages 1199–1204. Morgan Kaufmann Publishers Inc., 2009.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37 :9–27, 2011.

Randolph Quirk and David Crystal. *A comprehensive grammar of the English language*, volume 6. Cambridge Univ Press, 1985.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *EMNLP*, pages 105–112, 2003.

Sandeepkumar Satpal and Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *PKDD*, 2007.

R.E. Schapire and Y. Singer. Boostexter : A boosting-based system for text categorization. *Machine learning*, 39(2) :135–168, 2000.

William A Scott. Reliability of content analysis : The case of nominal scale coding. *Public opinion quarterly*, 1955.

Prajol Shrestha, Christine Jacquin, and Béatrice Daille. Clustering short text and its evaluation. In *Computational Linguistics and Intelligent Text Processing*, pages 169–180. Springer, 2012.

Fangzhong Su and Katja Markert. From words to senses : a case study of subjectivity recognition. In *International Conference on Computational Linguistics*, 2008.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. In *Computational linguistics*, 2011.

H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *ACL*, 2005.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, 2008a.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, 2008b.

P. Turney, M.L. Littman, et al. Measuring praise and criticism : Inference of semantic orientation from association. In *ACM Transactions on Information Systems*, 2003.

P.D. Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Arxiv preprint cs/0212012*, 2002.

Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. Latent vector weighting for word meaning in context. In *EMNLP*, pages 1012–1022. Association for Computational Linguistics, 2011.

Matthieu Vernier. *Analyse à granularité fine de la subjectivité*. PhD thesis, Université de Nantes, 2011.

Matthieu Vernier, Laura Monceaux, and Béatrice Daille. Learning subjectivity phrases missing from resources through a large set of semantic tests. In *LREC*, 2010.

Paul von Bünau, Frank C Meinecke, Franz C Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103(21) :214101, 2009.

Janyce Wiebe and Rada Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics, 2006.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3) :165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, 2005. URL http://acl.ldc.upenn.edu/H/H05/H05-1044.pdf.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35 :339–433, 2009.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Alan Ritter, Sara Rosenthal, and Veselin Stoyanov. Semeval-2013 task 2 : Sentiment analysis in twitter. In *7th International Workshop on Semantic Evaluation*, 2013.

Theresa Ann Wilson. *Fine-grained subjectivity and sentiment analysis : recognizing the intensity, polarity, and attitudes of private states*. ProQuest, 2008.

Yunfang Wu and Peng Jin. Semeval-2010 task 18 : Disambiguating sentiment ambiguous adjectives. In *5th International Workshop on Semantic Evaluation*, pages 81–85, 2010.

Ruifeng Xu, Jun Xu, and Xiaolong Wang. Instance level transfer learning for cross lingual opinion analysis. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 182–188. Association for Computational Linguistics, 2011.

Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 627–634. ACM, 2008.

Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto. Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polaritys. In *AAAI*, 2011.

Yong Zhang, Dong-Hong Ji, Ying Su, and Hongmiao Wu. Joint naÃ¯ve bayes and lda for unsupervised sentiment analysis. In *PAKDD (1)*, pages 402–413, 2013.

Tian Tian Zhu, Fang Xi Zhang, and Man Lan. Ecnucs : A surface information based system description of sentiment analysis in twitter in the semeval-2013 (task 2). *Atlanta, Georgia, USA*, page 408, 2013a.

Zhemin Zhu, Djoerd Hiemstra, Peter Apers, and Andreas Wombacher. Ut-db : An experimental study on sentiment analysis in twitter. In *7th International Workshop on Semantic Evaluation*, 2013b.