État de l'art : l'influence du domaine sur la classification de l'opinion

Dis-moi de quoi tu parles, je te dirai ce que tu penses

Morgane Marchand[1,2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Centre Nano-Innov Saclay, 91191 Gif-sur-Yvette Cedex
(2) LIMSI-CNRS, Univ. Paris-Sud
91403 Orsay Cedex
morgane.marchand@cea.fr

RÉSUMÉ

L'intérêt pour la fouille d'opinion s'est développé en même temps que se sont répandus les blogs, forums et autres plate-formes où les internautes peuvent librement exprimer leur opinion. La très grande quantité de données disponibles oblige à avoir recours à des traitements automatiques de fouille d'opinion. Cependant, la manière dont les gens expriment leur avis change selon ce dont ils parlent. Les distributions des mots utilisés sont différentes d'un domaine à l'autre. Aussi, il est très difficile d'obtenir un classifieur d'opinion fonctionnant sur tous les domaines. De plus, on ne peut appliquer sans adaptation sur un domaine cible un classifieur entraîné sur un domaine source différent. L'objet de cet article est de recenser les moyens de résoudre ce problème difficile.

ABSTRACT

State of the Art : Influence of Domain on Opinion Classification

The interest in opinion mining has grown concurrently with blogs, forums, and others platforms where the internauts can freely write about their opinion on every topic. As the amounts of available data are increasingly huge, the use of automatic methods for opinion mining becomes imperative. However, sentiment is expressed differently in different domains : words distributions can indeed differ significantly. An effective global opinion classifier is therefore hard to develop. Moreover, a classifier trained on a source domain can't be used without adaptation on a target domain. This article aims to describe the state-of-the-art methods used to solve this difficult task.

MOTS-CLÉS : État de l'art, Fouille d'opinion, Multi-domaines, Cross-domaines.

KEYWORDS: State of the art, Opinion mining, Multi-domain, Cross-domain.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 177–190,
Grenoble, 4 au 8 juin 2012.

1 Introduction

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'opinion en tout genre.
Les applications possibles de la fouille d'opinion sont multiples (Pang et Lee, 2008). 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.

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).
L'analyse de l'opinion peut se situer au niveau du texte entier, du paragraphe, de la phrase ou bien du fragment selon les applications envisagées.

Dans cet article, nous nous intéresserons uniquement à la tâche de classification de l'axiologie de l'opinion, dont nous donnons un aperçu des problèmes dans la section 2, de façon générale et dans le cas particulier des domaines différents. En section 3, nous expliciterons brièvement pourquoi les techniques classiques pour la constitution des ressources ou des classifieurs sont moins efficaces lorsque l'on change de domaine d'expression. Nous dresserons alors, dans la section 4, un état de l'art des recherches actuelles visant à améliorer les performances des classifieurs dans ce cas particulier. Dans une dernière partie, nous évoquerons les travaux préliminaires effectués ainsi que les perspectives d'exploration.

2 La subjectivité dans le langage

2.1 Présentation et définition

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 exprimant 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.

2.2 Différences d'expression selon les domaines ou les niveaux de langage

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 (Pupier, 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 et Wiebe, 2003). 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 lexicale (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 existant (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 langage. 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 classiques 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.

3 Les problèmes d'adaptation des ressources et des classifieurs classiques

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 et Lee, 2008).

3.1 Les ressources

Pour la constitution de ressources, on distingue deux grandes familles d'approches. La première consiste à utiliser des dictionnaires. A partir d'un petit ensemble de mots, appelés mots racines, le lexique est étendu en utilisant les relations de synonymie et d'antonymie (Kim et Hovy, 2005; Esuli et Sebastiani, 2006) ou bien les définitions (Andreevskaia et Bergler, 2006). La seconde consiste à s'appuyer sur un corpus. Le lexique de mots racines est étendu en s'appuyant sur plusieurs indices comme les conjonctions et/mais (Hatzivassiloglou et McKeown, 1997), la cooccurrence entre mots(Turney et 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 et 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 lexique 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.

3.2 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 *et al.*, 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 utilisées qui peuvent être des n-grammes (Pang *et al.*, 2002), des arbres de relations syntaxiques (Kudo et 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 score 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 et 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*).

4 Ressources et techniques pour l'adaptation au domaine

Afin de surmonter les défauts de performance des méthodes classiques, la première possibilité est de s'attacher à développer des ressources générales plus performantes (section 4.1). 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 (section 4.2). Enfin, lorsque l'on dispose déjà de ressources ou d'outils adaptés à un domaine particulier, on peut les adapter à un domaine proche (section 4.3).

4.1 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.
Il existe un flou sur ce qu'on appelle 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 et 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 dépendance 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, si on veut utiliser des classifieurs fondés uniquement sur des méthodes d'apprentissage statistique tout en étant les plus généraux possible, il faut des données d'apprentissage venant du plus grand nombre de domaines possible. En effet, quand on a un peu de données annotées dans plusieurs domaines, on peut faire en sorte que les domaines s'aident les uns les autres. 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 et Zong, 2008; Li *et al.*, 2011). La fusion la plus efficace dans ces travaux 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 donnant 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.

4.2 Passer automatiquement 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 racines du lexique initial.
The Double Propagation method, 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.

Enfin, sans étendre ou restreindre le vocabulaire, on peut 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 et 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 les mots et la polarité des expressions qui les contiennent afin d'adapter la polarité *a priori* des mots.

4.3 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.

4.3.1 Mélanger les corpus ou les traits

Si l'on dispose d'un corpus annoté suffisamment grand, la méthode donnant les meilleurs résultats repose, de façon naturelle, sur un entraînement direct sur les données du domaine. En revanche, dans le cas où l'on ne dispose pas de données annotées, il devient utile de s'entraîner sur d'autres corpus. 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 certain mot dépend ou non du domaine du texte où se trouve le mot. Ainsi, ils construisent automatiquement des dictionnaires valués pour chaque domaine.
Afin de s'adapter plus précisément au domaine cible, des poids peuvent être attribués aux exemples (Bickel *et al.*, 2007) ou aux traits (Satpal et Sarawagi, 2007). Ces méthodes s'appliquent également à l'extraction d'information générale (Gupta et Sarawagi, 2009).

Un problème peut 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. A notre connaissance, il n'y a pas de travaux en classification d'opinion traitant ce problème particulier.

4.3.2 Domaine de représentation commune

Une autre approche est d'essayer de détecter des pivots, des structures communes entre deux domaines. La méthode développée dans (Blitzer *et al.*, 2006), le *Structural Correspondance Learning* (SCL) se fonde sur la recherche de pivots entre les deux domaines permettant de comparer les histogrammes de répartition des différents termes des domaines. Elle est motivée par un algorithme d'apprentissage multitâches, ASO (*Alternating Structural Optimization*), proposé par (Ando et Zhang, 2005). Cette méthode a été appliquée à la recherche d'opinion dans (Blitzer *et al.*, 2007), travaux que nous reproduisons dans la partie 5. 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 (Blitzer *et al.*, 2011), les auteurs s'intéressent plus spécifiquement au cas où les supports

des domaines source et cibles (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 au deux domaines mais également par rapport à des mots spécifiques à un seul domaine.

Un travail plus récent à ce sujet est celui de (Pan *et al.*, 2010). Ils se servent également comme pivots de mots indépendants du domaine sélectionné 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) montre que la méthode SFA obtient de meilleurs résultats en exactitude que d'autres méthodes, dont SCL.

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 et Gamon, 2005).

4.3.3 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 et 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 sources et cibles 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 s'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.

5 Pistes de recherche et travaux préliminaires

Notre thème de recherche concerne l'adaptation au domaine pour la fouille d'opinion et la constitution automatique de lexiques pour ce problème. Les travaux cherchant à projeter deux corpus de domaines différents dans un espace commun semblent prometteurs. Aussi, nous nous sommes attachés à reproduire les travaux présentés dans (Blitzer *et al.*, 2007). 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.

5.1 Description de la méthode SCL

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 sac 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 obtenus trois étoiles n'ont pas été pris en compte à cause de leur polarité ambigüe.
Les travaux des auteurs cherchent à reproduire la situation réelle où l'on dispose d'un grand nombre de données non annotées à la fois pour le domaine cible et pour le domaine source, mais seulement 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.

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, co-occurrent 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 traits qui apparaissent fréquemment dans les deux domaines. Ces traits 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. Seuls les 1000 plus informatifs sont conservés. Ces traits

pivots sont donc fréquents dans les deux domaines et relativement utile à la tâche de classification de l'opinion pour le domaine source (par exemple "a-must", "loved-it", "weak", "awful", etc.)

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 singulière. 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. Les auteurs utilisent un classifieur linéaire dont les coefficients sont obtenus par descente stochastique de gradient.

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.

5.2 Nos travaux de reproduction

Nous utilisons deux des corpus constitués et utilisés par les auteurs : les corpus *DVDs* et *kitchen* du *Multi-Domain Sentiment Dataset*. Le corpus source *DVDs* contient 5586 critiques et le corpus cible *kitchen* 7945 critiques également réparties entre négatives et positives. Comme dit précédemment, le domaine source contient 1000 critiques positives et 1000 critiques négatives pour lesquelles on connaît les étiquettes. En moyenne, les critiques du corpus *kitchen* contiennent 145 unigrammes et bigrammes, celles de *DVDs*, 269.

Nous 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 également nos résultats avec ceux présentés dans (Blitzer *et al.*, 2007).

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. Il nous a semblé intéressant d'étudier l'influence de la sélection des traits pivots sur la performance en précision pour les deux classes.
Nous avons 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 critère précédent.

Le tableau 1 présente les résultats obtenus pour un classifieur entraîné sur *DVDs* et testé sur *kitchen* ainsi que les références présentées plus haut.
De plus, dans (Blitzer *et al.*, 2007), les auteurs normalisent les nouveaux traits afin que leur norme moyenne équivaille à α fois celle des anciens traits. Ils obtiennent de cette façon de meilleurs résultats. La dernière ligne du tableau présente donc les résultats avec un seuil α de pondération que nous avons expérimentalement fixé à 0,5.

	Blitzer et al.	Exactitude (*Accuracy*)	Précision classe positive	Précision classe négative	Rappel classe positive	Rappel classe négative
Réf. source->cible	74,0	78,5	79,4	77,6	76,5	80,4
Réf. source->source	82,4	81,8	80,3	83,4	84,6	79,0
Réf. cible->cible	87,7	87,7	88,4	87,0	86,4	88,9
Pivots : fréquence	79,4	79,8	80,9	78,7	77,6	81,9
Pivots : MI	.	79,6	85,0	75,7	71,6	87,6
Pivots : mixte	.	79,9	83,9	76,7	73,6	86,1
Pivots : mixte pond.	81,4	80,7	82,5	79,13	77,6	83,8

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

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 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.
Nous observons donc une difficulté particulière à la classe négative. Plus de textes positifs sont faussement classés en négatif que l'inverse. Une difficulté similaire a été notée par (Vernier *et al.*, 2009) pour la détection précise de passages subjectifs négatifs. Il faudra donc porter une attention particulière au traitement des opinions négatives.

5.3 Perspectives

L'utilisation de la matrice de projection créée par la méthode SCL est donc utile à 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 se rapportant aux sujets 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 informatifs. Lorsque le classifieur est adapté du domaine hétérogène vers le domaine homogène, il manque donc les informations contenus 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.

L'utilisation d'une matrice de projection obtenue par une décomposition en valeurs singulières rend l'interprétation des résultats plus difficiles car les traits finaux ne sont plus des unigrammes ou des bigrammes. Nous aimerions pouvoir rendre cette méthode plus interprétable, c'est-à-dire garder des traits liés aux mots de façon directe. Notre idée serait d'utiliser une méthode s'inspirant de (Pan *et al.*, 2010). Une fois les traits sources et cibles projetés dans l'espace commun nouvellement créé, on peut les regrouper en *clusters*. Ce sont ces *clusters* qui seraient alors les nouveaux traits.
Une autre possibilité est d'utiliser l'hyperplan séparateur du classifieur afin de donner des scores d'opinion aux termes cibles qui serait la distance à cet hyperplan séparateur. Nous faisons l'hypothèse que les mots réellement polarisés auront une grande distance à cet hyperplan.

6 Conclusion

Nous nous sommes intéressés dans cet article à la fouille d'opinion et plus particulièrement à la classification de l'opinion et nous avons présenté un état de l'art des différentes méthodes utilisées pour cette tâche, en particulier pour traiter le problème de l'adaptation au domaine. Nous avons vu que l'expression de l'opinion prend des formes très variées et qui dépendent du domaine où l'on se place. Un mot ayant une polarité neutre dans un certain contexte peut avoir une polarité positive dans un autre. C'est pourquoi il est très difficile de mettre au point un classifieur ayant de bonnes performances dans tous les domaines.

Les pistes étudiées pour pallier ce problème sont multiples. On peut tout d'abord améliorer les ressources générales, notamment en créant des lexiques contextuels précis. Une autre approche est de développer des techniques pour particulariser automatiquement des ressources ou des classifieurs généraux à l'aide d'un corpus mono-domaine spécifique. Enfin, une troisième possibilité est de travailler sur l'adaptation entre domaines. Pour cela, on peut projeter l'espace cible sur l'espace source ou bien projeter les deux espaces dans un espace commun. La difficulté réside alors dans la détermination de cet espace de projection.

Nous avons également présenté nos premières pistes de recherche pour définir cet espace de projection de telle sorte qu'il reste lié à un lexique, pour rester interprétable.

Références

ANDO, R. et ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853.

ANDREEVSKAIA, A. et BERGLER, S. (2006). Mining wordnet for fuzzy sentiment : Sentiment tag extraction from wordnet glosses. *In EACL*, volume 6.

AUE, A. et GAMON, M. (2005). Customizing sentiment classifiers to new domains : A case study. *In Recent Advances in Natural Language Processing*.

BEN-DAVID, S., BLITZER, J., CRAMMER, K. et PEREIRA, F. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137.

BENAMARA, F., CESARANO, C., PICARIELLO, A., REFORGIATO, D. et SUBRAHMANIAN, V. (2007). Sentiment analysis : Adjectives and adverbs are better than adjectives alone. *In ICWSM*.

BENVENISTE, E. (1966). *Problèmes de linguistique générale I*. Gallimard.

BICKEL, S., BRÜCKNER, M. et SCHEFFER, T. (2007). Discriminative learning for differing training and test distributions. *In 24th international conference on Machine learning*. ACM.

BLAIR-GOLDENSOHN, S., HANNAN, K., MCDONALD, R., NEYLON, T., REIS, G. et REYNAR, J. (2008). Building a sentiment summarizer for local service reviews. *In WWW Workshop on NLP*.

BLITZER, J., DREDZE, M. et PEREIRA, F. (2007). Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification. *In Annual Meeting of the ACL*.

BLITZER, J., FOSTER, D. et KAKADE, S. (2011). Domain adaptation with coupled subspaces. *Journal of Machine Learning Research - Proceedings Track*, 15:173–181.

BLITZER, J., MCDONALD, R. et PEREIRA, F. (2006). Domain adaptation with structural correspondence learning. *In EMNLP*.

CHOI, Y. et CARDIE, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *In EMNLP*.

DANG, Y., ZHANG, Y. et CHEN, H. (2010). A lexicon enhanced method for sentiment classification : An experiment on online product reviews.

DAUMÉ, H. (2007). Frustratingly easy domain adaptation. *In Annual Meeting of the ACL*.

DENECKE, K. (2009). Are sentiwordnet scores suited for multi-domain sentiment classification ? *In International Conference on Digital Information Management*.

ESULI, A. et SEBASTIANI, F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. *In LREC*.

FERRARI, S., CHARNOIS, T., MATHET, Y., RIOULT, F. et LEGALLOIS, D. (2009). Analyse de discours évaluatif, modèle linguistique et applications. *RNTI*, E-17:71–93.

GINDL, S., WEICHSELBRAUN, A. et SCHARL, A. (2010). Cross-domain contextualisation of sentiment lexicons. *European Conference on Artificial Intelligence*.

GUPTA, R. et SARAWAGI, S. (2009). Domain adaptation of information extraction models. *ACM SIGMOD Record*, 37(4):35–40.

HARB, A., DRAY, G., PLANTIÉ, M., PONCELET, P., ROCHE, M., TROUSSET, F. *et al.* (2008). Détection d'opinion : Apprenons les bons adjectifs ! *Atelier FOuille des Données d'OPinions*.

HATZIVASSILOGLOU, V. et MCKEOWN, K. (1997). Predicting the semantic orientation of adjectives. *In EACL*.

HOFFMAN, J., SAENKO, K., KULIS, B. et DARRELL, T. (2011). Domain adaptation with multiple latent domains. *In NIPS Domain Adaptation Workshop*.

JIJKOUN, V., RIJKE, M. et WEERKAMP, W. (2010). Generating focused topic-specific sentiment lexicons. *In Annual Meeting of the ACL*.

KIM, S. et HOVY, E. (2005). Automatic detection of opinion bearing words and sentences. *In International Joint Conference on Natural Language Processing*, pages 61–66.

KUDO, T. et MATSUMOTO, Y. (2004). A boosting algorithm for classification of semi-structured text. *In EMNLP*.

LI, S., HUANG, C. et ZONG, C. (2011). Multi-domain sentiment classification with classifier combination. *Journal of Computer Science and Technology*, 26:25–33.

LI, S. et ZONG, C. (2008). Multi-domain sentiment classification. *In Annual Meeting of the ACL*.

MANSOUR, Y., MOHRI, M. et ROSTAMIZADEH, A. (2009). Domain adaptation : Learning bounds and algorithms. *In Conference on Learning Theory*.

NAVIGLI, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys*.

PAN, S., NI, X., SUN, J., YANG, Q. et CHEN, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. *In International Conference on World Wide Web*.

PANG, B. et LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrival*, 2:1–2.

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques.

PUPIER, P. (1998). Une première systématique des évaluatifs en français. *Revue québécoise de linguistique*, 26(1).

QIU, G., LIU, B., BU, J. et CHEN, C. (2009). Expanding domain sentiment lexicon through double propagation. *In International Joint Conference on Artificial Intelligence*.

QIU, G., LIU, B., BU, J. et CHEN, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37:9–27.

RILOFF, E. et WIEBE, J. (2003). Learning extraction patterns for subjective expressions. *In EMNLP*.

SATPAL, S. et SARAWAGI, S. (2007). Domain adaptation of conditional probability models via feature subsetting. *In Knowledge Discovery in Databases : PKDD*.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K. et STEDE, M. (2011). Lexicon-based methods for sentiment analysis. *In Computational linguistics*.

TAKAMURA, H., INUI, T. et OKUMURA, M. (2005). Extracting semantic orientations of words using spin model. *In ACL*.

TURNEY, P. et LITTMAN, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Erb-1094, Institute for Information Technology, Canada.

TURNEY, P., LITTMAN, M. *et al.* (2003). Measuring praise and criticism : Inference of semantic orientation from association. *In ACM Transactions on Information Systems*.

VERNIER, M., MONCEAUX, L. et DAILLE, B. (2010). Learning subjectivity phrases missing from resources through a large set of semantic tests. *In LREC*.

VERNIER, M., MONCEAUX, L. et DUBREIL, E. (2009). Catégorisation sémantico-discursive des évaluations exprimées dans la blogosphère. *In TALN*.

WILSON, T., WIEBE, J. et HOFFMANN, P. (2009). Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35:339–433.

YOSHIDA, Y., HIRAO, T., IWATA, T., NAGATA, M. et MATSUMOTO, Y. (2011). Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polaritys. *In Porceedings of AAAI*.