Classification de Sentiments Multi-Domaines en Contexte Hétérogène & Passage à l'Echelle

Abdelhalim Rafrafi — Vincent Guigue — Patrick Gallinari

Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie, Paris 6 - 4 place Jussieu F-75252 PARIS cedex 05
{abdelhalim.rafrafi, vincent.guigue, patrick.gallinari}@lip6.fr

RÉSUMÉ. *La classification de sentiments multi-domaines est un problème complexe: en effet, les distributions de caractéristiques sont alors différentes dans les ensembles d'apprentissage et de test. Différentes propositions permettent de limiter la baisse de performance inhérente à ce cadre. Cependant, la classification de sentiments est une tâche très particulière: le web participatif nous donne accès à une quasi-infinité de données étiquetées pour apprendre des modèles supervisés. Cela soulève de nouvelles questions: à partir de quel volume de données les distributions d'apprentissage et de test convergent elles? Quand est ce que l'intérêt des techniques de transfert disparait? Dans cet article, nous étudions le taux de reconnaissance en sentiments par rapport la taille des ensembles d'apprentissage. D'abord, nous utilisons le corpus Amazon puis nous introduisons un nouveau cadre, la classification de sentiments multi-domaines hétérogènes (à partir des corpus movie reviews et TripAdvisor).*

ABSTRACT. *Multi-domain sentiment classification is known to be a difficult task in the literature since the feature distributions are different on training and testing sets. Thus, different transfer learning techniques have been proposed to cope with the induced lack of performance in recent years. However, the sentiment classification task is a particular supervised task where the labeled data are almost infinite (on the web 2.0). As a consequence, a new question emerged: if we have enough labeled data, does the train distribution converge to the test distribution? When does the transfer learning benefit vanish? In this article, we study the sentiment classification accuracy wrt the learning set size on the Amazon dataset. Afterwards, we consider different heterogeneity levels of transfer, we tackle the problem of keeping efficiency while learning on movie review and TripAdvisor datasets.*

MOTS-CLÉS : *Classification de Sentiments, Transfert Multi-Sources, Fouille d'Opinion*

KEYWORDS: *Multi-Source Domain Adaptation, Sentiment Classification, Opinion Mining*

1. Introduction

La fouille d'opinion s'est développée avec le web participatif (2.0) et les contenus utilisateurs. Les forts enjeux économiques (e-réputation, détection de buzz...) et politiques (sondages, identification des leaders d'opinion...) expliquent l'essor rapide de la littérature scientifique sur le sujet depuis une dizaine d'années. Comme le résument (Pang *et al.,* 2008) dans leur étude, deux problèmes principaux restent ouverts en classification de sentiments : la prise en compte de la structure dans l'expression des sentiments et le développement de modèles robustes en multi-domaines. Nous nous intéressons dans cet article au second problème en étudiant le comportement de techniques d'apprentissage automatique sur plusieurs domaines en classification de sentiments.

Les contenus utilisateurs (revues, commentaires...) sont souvent postés avec une note explicite (par exemple en étoiles) et les techniques d'apprentissage permettent d'exploiter ces données. L'efficacité de ces modèles en classification de sentiments a été démontré dès (Pang *et al.,* 2002). Plusieurs améliorations ont ensuite été proposée via des caractéristiques de haut niveau allant des n-grammes (Dave *et al.,* 2003) au codage des négations (Das *et al.,* 2001) et de l'analyse d'arbre syntaxiques (Matsumoto *et al.,* 2005). Mais toutes ces approches sont dépendantes du domaine et construire des classifieurs plus robustes est vite devenu un enjeu majeur (Blitzer *et al.,* 2007). La problématique est appelée *adaptation multi-domaine*. Pour tester la robustesse des solutions, le protocole consiste à apprendre sur un jeu de données (source) et tester les performances sur des données d'un autre domaine (cible). Les propositions se répartissent en trois grands ensembles : utiliser le cadre de la régularisation pour améliorer le pouvoir de généralisation des modèles(Dredze *et al.,* 2010a, Rafrafi *et al.,* 2012), apprendre un espace sémantique codant des informations complexes et générales qui facilitent le passage à des données inconnues (Liu *et al.,* 2007, Maas *et al.,* 2011) ou optimiser un modèle explicite de transfert qui minimise la distance entre les distributions source et cible (Blitzer *et al.,* 2007, Pan *et al.,* 2010).

Cet article décrit une double contributions expérimentales. Nous nous intéressons aux modèles d'adaptation à partir de sources multiples comme (Mansour *et al.,* 2008, Whitehead *et al.,* 2009) : plusieurs bases d'apprentissage de différents domaines (ici des sous-corpus d'Amazon) sont fusionnées pour apprendre un modèle qui est testé sur de nouvelles données (la cible appartient à un domaine inconnu en apprentissage). Nous montrons que dans ce cadre, il est possible d'obtenir systématiquement des performances meilleures qu'en intra-domaine (apprentissage et test réalisés sur la cible). A notre connaissance, il s'agit de la première étude aboutissant à cette conclusion. Notre hypothèse est que la taille des ensembles d'apprentissage est directement liée à la performance en reconnaissance de sentiments tandis que la plupart des études précédentes se focalisent sur de petits jeux de données qui ne permettent pas de modéliser correctement la distribution de mots caractéristiques des sentiments. Afin d'expliquer les bons résultats obtenus, nous analysons en détail l'évolution des performances de nos systèmes en fonction de la taille des corpus d'apprentissage. Nous proposons

également une étude empirique asymptotique des performances basée sur le jeu de donnée *huge Amazon* (Jindal *et al.,* 2008) (5,8 millions de revues).

Cette première série d'expérience repose sur des revues Amazon qui sont issus de domaines relativement proches (Blitzer *et al.,* 2007). Notre seconde série d'expériences décrit l'évolution des performances en contexte hétérogène. L'apprentissage est alors effectué sur les bases 50k *movie reviews* (Maas *et al.,* 2011) et 50k *TripAdvisor* (Wang *et al.,* 2010). Alors que les techniques de transfert explicite étaient devenues inutiles dans l'expérience Amazon, notre conclusion est différente dans pour l'adaptation hétérogène : dans ce dernier cas, ces techniques apportent stabilité et efficacité aux résultats.

En section 2, nous présentons les travaux connexes. Notre approche est détaillée en section 3. Enfin, toutes nos expériences sont décrites en section 4 : nous analysons successivement le cadre intra-domaine, les transferts mono-source puis multi-sources et l'adaptation hétérogène.

2. Travaux Connexes

Les algorithmes multi-domaines ont été largement étudiés en classification de textes, cependant, les applications en analyse de sentiments sont plus récentes (Blitzer *et al.,* 2007). Nous décrivons les principales propositions émises depuis 2007.

2.1. *Classification de Sentiments Multi-Domaines*

L'hypothèse classique i.i.d. n'est pas valable dans le cas multi-domaines : chaque domaine est en effet associé à une distribution de mots caractéristique. Cela explique l'écart de performance entre les systèmes mono et multi-domaines. Les techniques d'adaptation consistent à améliorer la généralisation des algorithmes en utilisant différentes stratégies :

2.1.1. *Régularisation*

La plupart des classifieurs de sentiments sont linéaires et utilisent une représentation en sac de mots (combinée à un codage présentiel) (Pang *et al.,* 2008). L'optimisation des poids du classifieurs peut être régularisée pour améliorer la généralisation. Des formulations spécifiques à la classification de sentiments ont été proposées (Crammer *et al.,* 2009, Rafrafi *et al.,* 2012). De plus, lorsque plusieurs sources de données sont utilisées en apprentissage, il est possible d'ajouter des termes de régularisation pour contraindre le classifieur à avoir un comportement cohérent (Dredze *et al.,* 2010a). (Daumé-III, 2007) propose un cadre particulier basé sur l'enrichissement de la représentation, chaque source/cible étant décrite à l'aide de caractéristiques spécifiques (par réplication du dictionnaire). L'apprentissage en grande dimension est délicat et repose alors sur une régularisation efficace. Comme dans la référence précédente, la régularisation peut également intervenir pour limiter les écarts de comporte-

ment entre les sources. Notons que cette approche requiert des données étiquetées du domaine cible.

2.1.2. *Alignement Explicite*

Une autre possibilité consiste à chercher les points communs entre les distributions source et cible pour améliorer le transfert : (Blitzer *et al.,* 2007, Pan *et al.,* 2010) proposent de chercher les mots qui ont le même comportement dans les deux domaines. Une fois identifiés les mots *pivots*, ils utilisent des techniques de factorisations matricielles pour apprendre une matrice de projection Θ qui permet de construire de nouvelles caractéristiques robustes aux transferts. Ces approches nécessitent des données non-étiquetées venant de la cible.

2.1.3. *Apprentissage d'un Espace Sémantique*

Le concept d'espace sémantique, permettant de projeter les mots dans un système métrique est récurrent en classification de documents. Formellement, il s'agit d'utiliser de gros corpus de documents pour apprendre les positions des mots dans l'espace. La classification de sentiments consiste ensuite à faire le lien entre des zones de l'espace et des opinions positives ou négatives. Dans un tel système, les mots de la cible qui n'apparaissaient pas dans la source ont une position et participent à la décision ce qui améliore les performances en transfert. Les premières approches se sont basée sur des dérivés de PLSA (Liu *et al.,* 2007, Lin *et al.,* 2009), puis ces modèles ont été dépassés par ceux dérivés de LDA (Gerrish *et al.,* 2011). Les travaux plus récents utilisent beaucoup les réseaux de neurones : (Glorot *et al.,* 2011) se base sur les auto-encodeurs, (Bespalov *et al.,* 2011) sur les réseaux à convolutions et (Maas *et al.,* 2011) sur les réseaux récursifs.

2.2. *Les cardes multi-domaines*

Deux cadres distincts existent en classification de sentiments multi-domaines : le plus simple consiste à prendre un seul domaine pour l'apprentissage (mono-source) et un pour le test (cible) (Blitzer *et al.,* 2007). Cependant, l'intérêt pratique de cette approche est discutable : en situation réelle, face à une cible inconnue, toutes les données d'apprentissage utiles peuvent être utilisées sans restriction.

(Mansour *et al.,* 2008) propose une étude théorique du gain associé à l'apprentissage multi-sources : les auteurs modélisent la base cible comme une mixture des différentes sources en quantifiant les contributions de chaque source. Une autre étude, empirique, est décrite dans (Whitehead *et al.,* 2009). Cet article introduit la notion de *leave-one-out* au niveau des sources d'apprentissage. Cependant, la conclusion des auteurs est négative : dans leur étude, le *leave-one-out* est toujours dépassé par (au moins) une source seule.

Dans (Dredze *et al.,* 2010b), les auteurs montrent au contraire que l'utilisation de plusieurs sources combinées est plus efficace qu'une source seule (même la meilleure).

Mais leur système est complexe et un passage à l'échelle semble délicat : ils modélisent la cible comme une mixture des sources mais se limitent à l'étude des petits corpus Amazon (2000 documents).

2.3. *Des corpus de plus en plus grands*

Les contributions des utilisateurs représentent une source quasi-infinie de données étiquetées en sentiments sur le Web participatif (2.0). Cependant, les premiers corpus mis en ligne étaient relativement petit : entre quelques centaines de documents (Riloff *et al.,* 2003, Whitehead *et al.,* 2009) et quelques milliers (Pang *et al.,* 2002). Les jeux de données plus récents sont nettement plus conséquents : 50k critiques de films dans (Maas *et al.,* 2011), 50k revues (hôtels et voitures) dans (Wang *et al.,* 2010), 300k revues Amazon dans (Blitzer *et al.,* 2007) et 5.8 millions de revues (toujours Amazon) dans (Jindal *et al.,* 2008). Le fait que ces données soient mises à disposition de la communauté scientifique est une vraie motivation pour étudier en profondeur l'évolution de la performance des techniques actuelles en fonction de la taille des bases d'apprentissage.

3. Motivation et Approche

Les organisateurs de l'atelier (Blitzer *et al.,* 2011) constatent que malgré les avancées récentes en adaptation multi-domaines, la plupart des approches ne sont pas robustes. Nous estimons que cette conclusion est directement liée à la petite taille des ensembles d'apprentissage généralement utilisés et nous proposons une analyse de l'impact de la taille des corpus sur les performances multi-domaines des modèles appris. Nous avons aussi remarqué que plusieurs approches proposées sont assez complexes et passent difficilement à l'échelle (Blitzer *et al.,* 2007, Pan *et al.,* 2010, Dredze *et al.,* 2010b). Nous nous focalisons sur une approche simple (SVM linéaires) et nous montrons qu'il est possible d'atteindre des performances de premier ordre si suffisamment de données sont disponibles.

3.1. *Motivation : passage à l'échelle en classification de sentiments*

Un article récent démontre clairement le lien entre les performances en reconnaissance de sentiments et la taille des ensembles d'apprentissage (Bespalov *et al.,* 2011). Etudions une expérience simple, sans transfert. A partir d'un corpus Amazon (Blitzer *et al.,* 2007), où tous les domaines sont mélangés, nous apprenons un SVM sur différentes portions de la base d'apprentissage (allant de 10% à 100% de la base). Pour chaque sous-expériences, les performances sont évaluées sur le même ensemble de test qui résulte de la fusion de tous les sous-ensembles de test Amazon (cf Figure 1).

Une tendance très claire se dégage de la courbe : utiliser plus de données conduit à de meilleures performances et même en utilisant les 90k documents disponibles en

apprentissage, la courbe n'atteint pas de plateau : il semble nécessaire d'avoir plus de données pour obtenir des résultats optimaux.

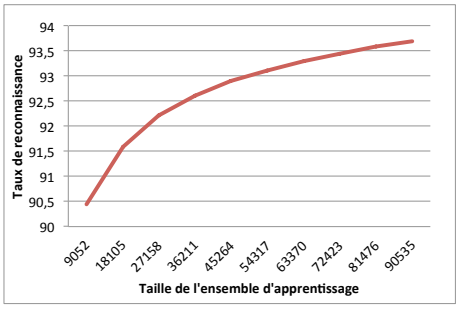


Figure 1 – Taux de reconnaissance en classification de sentiments par rapport à la taille de l'ensemble d'apprentissage (les détails concernant les corpus sont donnés dans le tableau 1).

Nous sommes convaincus que ce phénomène est critique pour l'adaptation de domaine. L'expérience précédente montre qu'un large ensemble d'apprentissage ne permet pas de modéliser complètement la distribution des mots pour la classification de sentiments alors que nous n'avons pas encore introduit le cadre multi-domaine. Nous pouvons estimer que les ensembles de 2k revues de (Blitzer *et al.,* 2007) ne permettent pas non plus d'obtenir un modèle fiable et général pour la distribution des mots. Les bonnes performances multi-domaines obtenues par (Glorot *et al.,* 2011) nous poussent également à utiliser des corpus plus larges (quelque soit le domaine de ces corpus et même s'ils ne sont pas étiquetés).

3.2. *Modèle et Représentation*

Pour toutes les expériences nous utilisons des SVM linéaires classiques. Nous souhaitons montrer qu'un modèle de base peut atteindre les performances de l'état de l'art en utilisant simplement plus de données. Pour cette raison, nous nous restreignons à une description classique des données (uni-grammes et bi-grammes) ainsi qu'au codage présentiel recommandé dans (Pang *et al.,* 2008).

Dans le cadre des expériences d'adaptation en contexte hétérogène, nous avons implémenté deux stratégies explicites de référence : l'algorithme *Structural Correspondence Learning* (SCL) (Blitzer *et al.,* 2007) et la méthode FEDA (*Frustratingly Easy Domain Adaptation*) (Daumé-III, 2007).

3.2.1. *Structural Correspondence Learning (SCL)*

SCL (Blitzer *et al.,* 2007) est un algorithme en trois étapes : np mots pivots sont d'abord extraits, ils doivent avoir le même comportement sur les données sources et cibles. np classifieurs linéaires $\mathbf{w} \in \mathbb{R}^d$ sont ensuite entrainés à déterminer si chaque

mot pivot est présent dans les documents ou pas (d désigne la taille du dictionnaire). Finalement les vecteurs $\mathbf{w}$ sont concaténés dans une matrice sur laquelle est appliquée une SVD. Il en résulte une matrice $\Theta \in \mathbb{R}^{np \times d}$ qui permet de construire np caractéristiques trans-domaines à ajouter à chaque document. Dans nos expériences, $np = 100$ et les mots pivots sont les plus fréquents dans les corpus sources et dans le corpus d'apprentissage de la cible (utilisé ici sans les étiquettes). Afin de garantir la teneur en sentiments des pivots, nous cherchons les mots fréquents en restreignant les possibilités à la liste proposée dans (Hu *et al.,* 2004). 100 nouvelles caractéristiques sont donc ajoutées à chaque document.

3.2.2. *Frustratingly Easy Domain Adaptation (FEDA)*

l'algorithme est décrit dans (Daumé-III, 2007) : l'idée est d'étendre la représentation des revues en dupliquant le dictionnaire $n + 1$ fois. Une représentation générale est utilisée pour tous les documents et n représentations pour les sources : chaque document est donc codé deux fois et $n - 1$ sous-vecteurs sont laissés à 0 (le codage dépend donc de la source). Afin d'améliorer les performances, un petit corpus cible étiqueté est introduit parmi les sources[1]. Cette approche nécessite un classifieur régularisé pour traiter correctement la grande dimensionnalité des données, nous avons utilisé SVMLight.

4. Expériences

Nous proposons dans cette partie deux séries d'expériences. Dans un premier temps, nous étudions le corpus Amazon qui est largement utilisé dans la littérature (Blitzer *et al.,* 2007, Pan *et al.,* 2010, Dredze *et al.,* 2010a). Nous testons plusieurs cadres d'apprentissage : le cas i.i.d, dit intra-domaine (apprentissage et test sur le même domaine, Section 4.2), les cas d'adaptation mono-source (Section 4.3) et multi-sources (Section 4.4). Une étude asymptotique basée sur le corpus Amazon large échelle conclut cette série en Section 4.5. Nous montrons que les performances en transfert peuvent dépasser les performances du cas i.i.d. à condition d'avoir suffisamment de données en apprentissage.

Dans un second temps, nous testons l'adaptation hétérogène dans une nouvelle série d'expériences, en utilisant les corpus 50k movie reviews (Maas *et al.,* 2011) et 50k TripAdvisor (Wang *et al.,* 2010).

4.1. *Données & Paramétrage des Expériences*

Tous les ensembles utilisés (Blitzer *et al.,* 2007, Maas *et al.,* 2011, Wang *et al.,* 2010) sont décrits dans le tableau Table 1. Les documents sont représentés en sacs de mots. Nous limitons le dictionnaire aux 5000 uni-grammes et bi-grammes les plus

1. Nous avons utilisé 1/3 des données d'apprentissage du domaine cible.

(a)

Amazon complet (Blitzer *et al.,* 2007) - 25 domaines			
Domaine	taille app.	taille test	% ex. neg.
Toys	6318	2527	19.63%
Software	1032	413	37.77%
Apparel	4470	1788	14.49%
Video	8694	3478	13.63%
Automotive	362	145	20.69%
Jewelry	982	393	15.01%
Grocery	1238	495	13.54%
Camera	2652	1061	16.31%
Baby	2046	818	21.39%
Magazines	1195	478	22.59%
Cell	464	186	37.10%
Outdoor	729	292	20.55%
Health	3254	1301	21.21%
Music	10625	24872	8.33%
Videogame	720	288	17.01%
Beauty	1314	526	15.78%
Sports	2679	1072	18.75%
Food	691	277	13.36%
Office	195	78	0.16%
Instruments	164	65	0.15%
Tools	32	11	0.03%

(b)

Ensembles de test (Blitzer *et al.,* 2007)			
Books	10625	10857	12.08%
DVDs	10625	9218	14.16%
Electronics	10196	4079	21.94%
Kitchen	9233	3693	20.96%

Amazon fusionné (union des précédents ensembles) (Blitzer *et al.,* 2007)			
Domaine	taille app.	taille test	% ex. neg.
Tous domaines	90535	68411	15.04%

Movie Reviews (version large) (Maas *et al.,* 2011)			
Domaine	taille app.	taille test	% ex. neg.
Movies	50000	-	50%

TripAdvisor (Wang *et al.,* 2010)			
Domaine	taille app.	taille test	% ex. neg.
Hotel & cars	50000	-	50%

Tableau 1 – Description des données. L'ensemble Amazon fusionné conserve deux ensembles séparés pour l'apprentissage et le test. Les deux derniers ensembles sont utilisés uniquement en apprentissage.

fréquents comme cela est généralement fait dans la littérature. Comme le préconnisent (Pang *et al.,* 2008), nous utilisons un codage présentiel[2]. L'apprentissage est réalisé avec SVMLight (Joachims, 2002) en conservant tous les régalages par défaut (y compris la régularisation).

L'ensemble de test est fixé une fois pour toutes : il s'agit des parties de test des sous-corpus d'Amazon *Books*, *Dvd*, *Electronics* et *Kitchen*. Cela représente un total de 27847 documents. Les expériences montrent deux comportements types : d'une part *Books* et *Dvd* qui sont assez proches et d'autre part, *Electronics* et *Kitchen*, qui réagissent de la même manière aux différents tests.

En fonction des expériences, nous utilisons les ensembles d'apprentissage suivants :

2. Les caractéristiques sont binaires, nous n'utilisons ni tf ni tf-idf.

– pour le cas i.i.d (intra-domaine), les parties d'apprentissage des domaines cibles sont utilisés (*Books*, *Dvd*, *Electronics* et *Kitchen*).

– Dans les expériences suivantes, nous utilisons les sous-domaines Amazon externes (différents de la cible). Amazon compte 25 domaines, il y a donc 24 domaines externes pour chaque cible. En adaptation mono et multi-sources, aucune donnée cible n'est utilisée lors de l'apprentissage.

– Amazon large échelle (Jindal *et al.,* 2008) regroupe 5,8 millions de revues, nous l'utilisons pour tester le comportement asymptotique des SVM en adaptation. Le découpage en sous-domaine est identique par rapport au corpus Amazon précédent.

– Pour le transfert hétérogène, l'apprentissage est réalisé sur les corpus 50k *movie reviews* (Maas *et al.,* 2011) et 50k *TripAdvisor* (Wang *et al.,* 2010) dont les critiques ne concernent pas des produits physiques mais respectivement des films, des hôtels et des voitures. Les revues ont également une forme différente : les critiques de films font 740 mots en moyenne contre 200 pour les revues Amazon.

4.2. *Intra-domaine (cas i.i.d)*

Présentons d'abord comme référence les résultats en intra-domaine, lorsque le domaine cible est présent dans l'ensemble d'apprentissage[3]. Les performances sur les 4 sous corpus cibles d'Amazon sont données en figure 2.

Nous proposons ensuite d'enrichir la base d'apprentissage avec les données provenant de n sources externes. Pour chaque valeur de n, plusieurs apprentissages sont effectués et les performances sont moyennées :

– pour $n = 1$, nous utilisons successivement 24 sources externes en plus de la base d'apprentissage de la cible. Pour $n = 2$, une combinaison de deux sources externes est ajoutée etc... Lorsque $n = 24$ une seule expérience est possible sur chaque cible (toutes les données disponibles pour l'apprentissage sont utilisées).

– Dès que la combinatoire dépasse 100 expériences, nous tirons aléatoirement 100 combinaisons de n sources externes.

– Le test est effectué sur les parties de test des 4 cibles.

Par soucis de clarté, la figure 2 présente une seule courbe de résultats, moyennée sur les 4 domaines *Books*, *Dvd*, *Electronics* et *Kitchen*. Les comportements des 4 base sont proches et les détails sont disponibles sur la figure 5 (ligne rouge pointillée). Notre analyse nous conforte dans notre hypothèse de la section précédente (Figure 1) : plus la base d'apprentissage est grande, meilleures sont les performances. Il est vrai que le gain n'est que de $1,3\%$ mais c'est significatif sur plus de 27800 documents. Une fois de plus, le plateau de performance n'est pas atteint dans cette expérience.

3. Chaque domaine cible comporte une partie d'apprentissage et une partie de test afin de ne pas biaiser les expériences (cf Tableau 1)

Domaine	Tx de reco.	♯ ens. test
Books	91.1%	10857
DVD	90.6%	9218
Electronics	90.6%	4079
Kitchen	91.7%	3693
Total	91%	27847

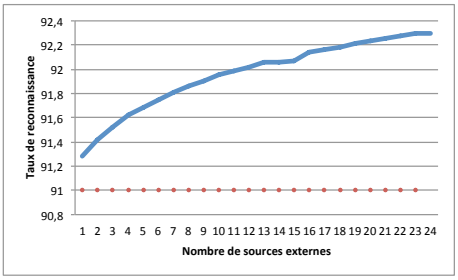


Figure 2 – Tableau de gauche : taux de reconnaissance intra-domaine sur *Books*, *DVD*, *Electronics* et *Kitchen*. Courbe de droite : évolution du taux (moyenné sur les 4 expériences) lorsque l'ensemble d'apprentissage est enrichi avec des sources externes.

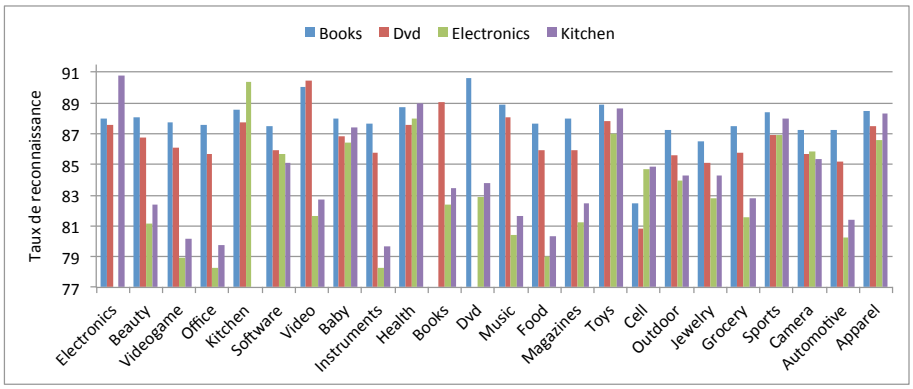


Figure 3 – Taux de reconnaissance sur les cibles *Books, Dvd, Electronics* et *Kitchen* en fonction des domaines sources. Les expériences intra-domaines ne sont pas reportées ici (une barre manque donc pour chaque domaine cible).

4.3. *Adaptation Mono-Source*

Le transfert mono-source est le cadre le plus répandu en classification de sentiments (Blitzer *et al.,* 2007). Une fois le modèle appris sur un domaine source (un sous-corpus d'Amazon), le test est effectué sur *Books*, *Dvd*, *Electronics* et *Kitchen*. Le principal constat réside dans l'instabilité chronique des résultats présentés en Figure 3 : les performances varient énormément en fonction de la source et de la cible. Des sources comme *Health, Toys, Sports*, ou *Apparel* proposent des performances correctes sur toutes les cibles. D'autres sources comme *Office, Musical Instruments, Video*,ou *Foods* donnent de bon taux de reconnaissance sur les cibles *Books* et *Dvd* mais pas sur *Kitchen* et *Electronics*.

Notre analyse de la Figure 3 est la suivante :

– les cibles *Books* et *Dvd* ont des comportements similaires par rapport aux différentes sources. La même conclusion s'impose pour le couple *Kitchen, Electronics*. Dans les deux cas, les performances croisées à l'intérieur du couple (apprentissage sur l'un des membres et test sur l'autre) sont particulièrement élevées.

– D'une manière générale, les cibles *Books* et *Dvd* sont plus simple à classer que les cibles *Kitchen* et *Electronics*.

– Les performances varient beaucoup. Les taux de reconnaissance sur *Books* et *Dvd* vont de 80.8% ào 90.6% en fonction de la source. C'est encore plus frappant pour *Kitchen, Electronics* (entre 78.3% et 90.7% de reconnaissance).

– Finalement ce cadre ne convient pas bien à l'adaptation : il requiert un oracle pour trouver la source optimale pour chaque cible.

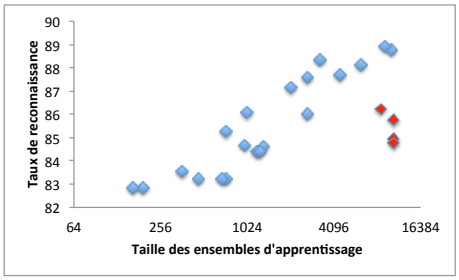


Figure 4 – Chaque point représente une source de données : l'abscisse de ces points donne la taille de la base utilisée, l'ordonnée la performance moyenne obtenue sur les 4 cibles. Les sources particulières (*Music, Video, Books* et *Dvd*) sont marquées en rouge.

Afin d'étudier l'impact de la taille des bases d'apprentissage, nous avons réalisé la Figure 4 dans laquelle chaque point représente une source de données : l'abscisse de ces points donne la taille de la base utilisée (en échelle log), l'ordonnée la performance moyenne obtenue sur les 4 cibles. Nous observons une relation claire entre la taille des sources et la performance en test sauf pour 4 bases : *Music*, *Video*, *Books* et *Dvd*. Ces quatre sources sont particulières : l'expression des sentiments y est particulièrement liée aux entités nommées, les critiques sont donc souvent implicites et ces revues utilisent largement la comparaison.

4.4. *Adaptation Multi-Sources*

Les premiers travaux en classification de sentiments multi-domaines se sont focalisés sur l'adaptation mono-source (Blitzer *et al.,* 2007). Comme nous l'avons déjà dit, le cas multi-sources est plus réaliste : nous cherchons à obtenir la meilleure performance sur une base inconnue à partir de tous les documents disponibles.

Nous reprenons les expériences de la section 4.2 en éliminant les domaines cibles de l'apprentissage. Les quatre ensembles de test ne change pas mais *Books*, *Dvd*, *Elec-*

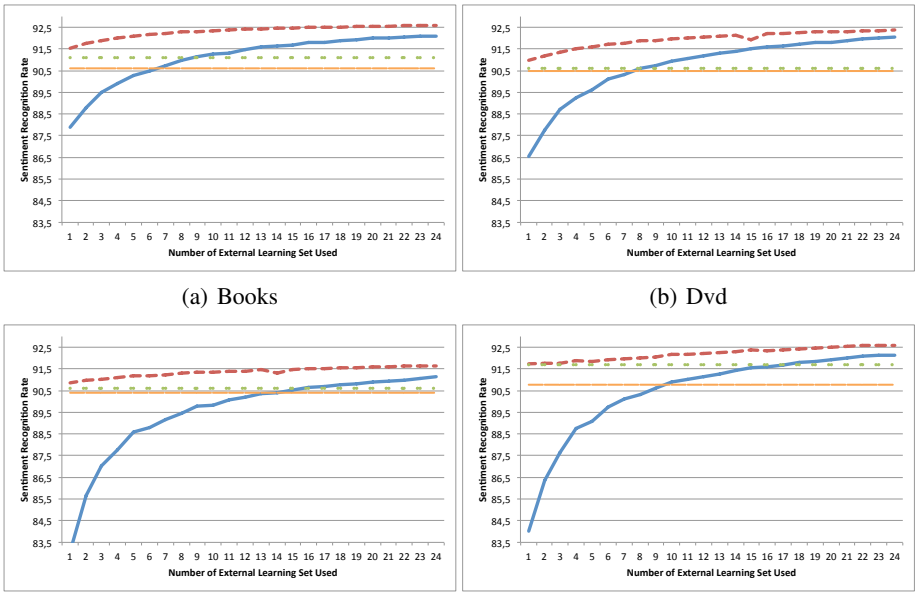
tronics et *Kitchen* ne sont plus utilisés en apprentissage. La figure 5 montre l'évolution des performances en fonction du nombre de sources externes utilisées, les taux de reconnaissance sont moyennés sur plusieurs expériences (cf détails en section 4.2). La figure compare quatre performances pour chaque cible :

1) ligne verte pointillée : expérience intra-domaine, chiffres issus du tableau 2,

2) ligne fine orange horizontale : transfert mono-source avec un oracle donnant la meilleure source pour chaque cible,

3) ligne bleue : transfert multi-sources, taux de reconnaissance en fonction du nombre de sources externes utilisées (le premier point de la courbe correspond au transfert mono-source),

4) ligne rouge pointillée : expériences intra-domaine + enrichissement de $n = 1$ à 24 sources externe (la cible est utilisée en apprentissage, cf section 4.2).



(a) Books (b) Dvd

(c) Electronics (d) Kitchen

Figure 5 – Taux de reconnaissance sur les 4 cibles en fonction du nombre de sources externes utilisées. Performance intra-domaine (pointillés verts), transfert mono-source+oracle (ligne fine orange), transfert multi-sources (bleu), intra-domaine+enrichissement (pointillés rouges). La dernière courbe donne les résultats détaillés de la section4.2.

Nous tirons plusieurs conclusions de ces expériences :

– le taux de reconnaissance est toujours lié à la taille de l'ensemble d'apprentissage (et le plateau de performance n'est pas atteint).

– Le cas du transfert mono-source est particulièrement défavorable : l'écart au début de la courbe par rapport à la performance intra-domaine va de 3,3% à 7,7% mais il diminue rapidement dès que de nouvelles sources sont utilisées.

– Le transfert multi-sources permet de dépasser systématiquement la performance mono-source+oracle (pour choisir la meilleure source). Nous reproduisons ici les résultats de (Dredze *et al.,* 2010b) mais avec un système nettement plus simple. Les gains en taux de reconnaissance vont de 0,7 à 1,6% comparé à l'oracle.

– Quelque soit la cible, la courbe multi-sources dépasse toujours la courbe intra-domaine (de 0,4 à 1,4%) : cette observation est nouvelle et à notre connaissance aucun article publié ne propose un tel résultat. Ces performances doivent être comparées notamment à (Glorot *et al.,* 2011) qui utilise le même jeu de données[4] sans aboutir à un gain systématique.

– Par rapport aux performances intra-domaine+enrichissement, l'écart reste toujours inférieur à 0,5% : la complexité et le coût des modèles de transfert deviennent des freins importants face aux maigres perspectives de gain.

4.5. *Adaptation multi-sources : cas asymptotique*

Dans les expériences précédentes, nous avons souvent mentionné le fait que le plateau de performance attendu n'était pas encore atteint. Nous avons donc utilisé le corpus Amazon large échelle (Jindal *et al.,* 2008) afin de chercher une asymptote empirique aux performances de notre approche (les ensembles de test restent inchangés -base Amazon d'origine- et le cadre est identique à celui décrit dans la section précédente). Les résultats proposés sont étonnants : le nouveau corpus apporte un faible gain sur *Books* et *Dvd* mais il pénalise la reconnaissance sur *Electronics* et *Kitchen*. Nous pensons que cela s'explique par le contexte différent des deux bases qui pénalise le transfert[5]. Une seconde conclusion était par contre attendue : l'écart entre l'intra-domaine enrichi et le multi-sources a pratiquement disparu.

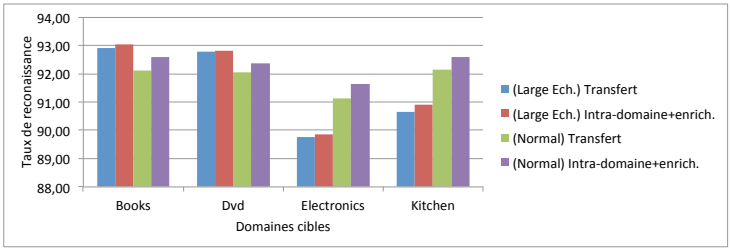


Figure 6 – Comparaison des taux de reconnaissance sur les 4 cibles en fonction du corpus d'origine (Amazon normal/Amazon large échelle)

4. sans les étiquettes des domaines externes
5. Les deux bases Amazon ont été collectées par différentes personnes à différents moments.

4.6. *Adaptation en contexte hétérogène*

Nos expériences montrent qu'Amazon est un corpus relativement homogène. Nous proposons une nouvelle série d'expérience en contexte hétérogène. La cible ne change toujours pas mais nous effectuons l'apprentissage sur différentes combinaisons de : Amazon, 50k movie reviews (IMDB) (Maas *et al.,* 2011) et 50k TripAdvisor (Wang *et al.,* 2010). Les nouvelles sources ne traitent pas de vente de produits physiques, les corpus sont équilibrés et les revues sont de tailles très différentes (cf section 4.1).

Les résultats présentés en figure 7 confirment plusieurs intuitions : Amazon est une source vraiment très efficace pour un test sur nos 4 cibles[6], les autres sources sont moins efficaces. Malheureusement, nous nous retrouvons dans le même cas qu'avec l'adaptation mono-source : les performances sont instables et il faudrait un oracle pour choisir les sources adaptées à chaque cible (le modèle utilisant toutes les données est toujours sous-optimal). Il faudrait sans doute beaucoup plus de revues pour stabiliser le processus hétérogène : passer à un niveau d'échelle supérieur pourrait apporter les mêmes résultats que ceux constatés sur Amazon en multi-sources.

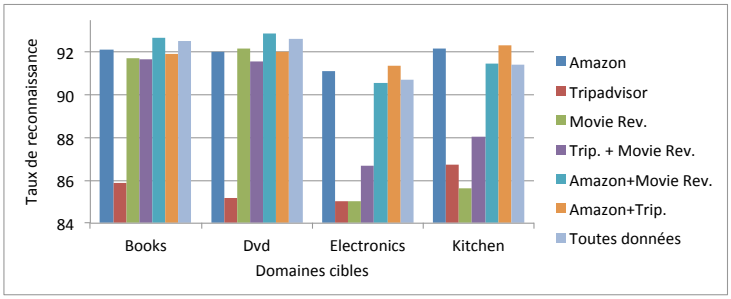


Figure 7 – Résultats obtenus sur les 4 cibles en fonction des sources utilisées.

Nous proposons ici une seconde série d'expériences plus proche d'un cas d'usage réel : Amazon est retiré des sources car nous considérons que le corpus est trop proche des cibles. Nous testons ensuite deux approches de référence en transfert explicite : SCL, la première approche en sentiments multi-domaines et FEDA, une approche simple et efficace ouverte sur les problèmes large-échelles (cf section 3.2).

La figure 8 compare 9 résultats pour chaque cible. Les 3 premiers sont des références : apprentissage sur Amazon, transfert hétérogène[7] dans le pire des cas, transfert hétérogène[7] avec un oracle. Les 3 suivants utilisent SCL (avec différentes sources), les 3 derniers FEDA (avec différentes sources également). Les techniques d'adaptation explicite fonctionnent très bien : les meilleures performances sont obtenues avec les bases complètes (plus besoin d'oracle). FEDA donne des résultats supérieurs simplement car il utilise des données cibles étiquetées (1/3 de l'ensemble d'apprentissage) : sans surprise, si ces données sont disponibles, il faut les utiliser.

6. Les domaines cibles ne sont pas utilisés en apprentissage.
7. sans Amazon

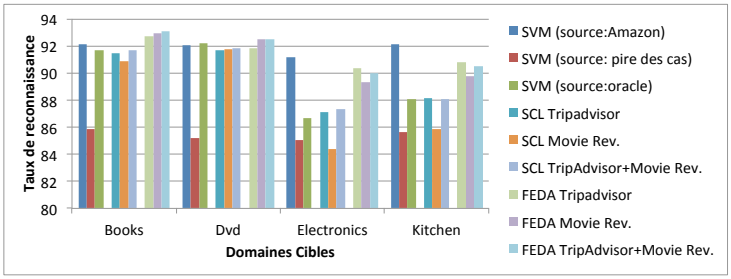


Figure 8 – Performances sur les 4 cibles. Pour chaque cible il y a 3 expériences de référence (Amazon, transfert hétérogène dans le pire des cas, transfert hétérogène avec un oracle), 3 expériences SCL et 3 expériences FEDA (correspondant à différentes sources).

5. Conclusion

Notre première conclusion est que les performances intra-domaines sur Amazon peuvent être dépassées à condition d'avoir suffisamment de données étiquetées (quelque soit la source dont elles viennent). Ce résultat constitue la nouveauté et l'apport de cet article.

Nous montrons aussi dans cet article que le transfert mono-source n'est pas un cadre intéressant sur les applications réelles : si l'intérêt académique et celui des algorithmes de (Blitzer *et al.,* 2007) sont indiscutables, le cas est trop instable et la limitation à une source est contre-productive.

Une fois démontré l'intérêt du passage à l'échelle sur la classification de sentiments, nous avons proposé un nouveau cadre d'adaptation hétérogène : nous montrons que le problème est difficile et requiert des techniques spécifiques. Les perspectives de ce travail concerne donc essentiellement un nouveau passage à l'échelle pour vérifier si le processus hétérogène est stabilisable avec plus de données.

6. Bibliographie

Bespalov D., Bai B., Qi Y., Shokoufandeh A., « Sentiment classification based on supervised latent n-gram analysis », *ACM CIKM*, p. 375-382, 2011.

Blitzer J., Cortes C., Rostamizadeh A., « Domain Adaptation Workshop : Theory and Application », *NIPS Workshop*, 2011.

Blitzer J., Dredze M., Pereira F., « Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification », *ACL*, 2007.

Crammer K., Kulesza A., Dredze M., « Adaptive Regularization Of Weight Vectors », *NIPS*, 2009.

Das S., Chen M., « Yahoo ! for Amazon : Extracting Market Sentiment from Stock Message Boards », *Asia Pacific Finance Association Annual Conference*, 2001.

Daumé-III H., « Frustratingly Easy Domain Adaptation », *ACL*, 2007.

Dave K., Lawrence S., Pennock D. M., « Mining the peanut gallery : opinion extraction and semantic classification of product reviews », *WWW*, ACM, p. 519-528, 2003.

Dredze M., Kulesza A., Crammer K., « Multi-domain learning by confidence-weighted parameter combination », *Machine Learning Jour.*, vol. 79, n° 1–2, p. 123-149, 2010a.

Dredze M., Kulesza A., Crammer K., « Multi-domain learning by confidence-weighted parameter combination », *Machine Learning*, vol. 79, p. 123-149, 2010b.

Gerrish S., Blei D., « Predicting Legislative Roll Calls from Text », *ICML*, p. 489-496, 2011.

Glorot X., Bordes A., Bengio Y., « Domain Adaptation for Large-Scale Sentiment Classification : A Deep Learning Approach », *ICML*, 2011.

Hu M., Liu B., « Mining and summarizing customer reviews », *ACM SIGKDD*, p. 168-177, 2004.

Jindal N., Liu B., « Opinion Spam and Analysis », *ACM WSDM*, 2008.

Joachims T., *Learning to Classify Text using Support Vector Machines*, Springer - Kluwer Academic Publishers, 2002.

Lin C., He Y., « Joint sentiment/topic model for sentiment analysis », *CIKM*, ACM, p. 375-384, 2009.

Liu Y., Huang X., An A., Yu X., « ARSA : a sentiment-aware model for predicting sales performance using blogs », *ACM SIGIR*, 2007.

Maas A. L., Daly R. E., Pham P. T., Huang D., Ng A. Y., Potts C., « Learning Word Vectors for Sentiment Analysis », *Association for Computational Linguistics (ACL)*, 2011.

Mansour Y., Mohri M., Rostamizadeh A., « Domain Adaptation with Multiple Sources », *NIPS*, 2008.

Matsumoto S., Takamura H., Okumura M., « Sentiment Classification using Word Sub-Sequences and Dependency Sub-Tree », *PAKDD*, 2005.

Pan S., Ni X., Sun J.-T., Yang Q., Chen Z., « Cross-Domain Sentiment Classification via Spectral Feature Alignment », *WWW*, 2010.

Pang B., Lee L., « Opinion mining and sentiment analysis », *Information Retrieval*, vol. 2, p. 1-135, 2008.

Pang B., Lee L., Vaithyanathan S., « Thumbs up ? : sentiment classification using machine learning techniques », *ACL-Empirical Methods in NLP*, vol. 10, p. 79-86, 2002.

Rafrafi A., Guigue V., Gallinari P., « Coping with the Document Frequency Bias in Sentiment Classification », *AAAI ICWSM*, 2012.

Riloff E., Wiebe J., « Learning Extraction Patterns for Subjective Expressions », *Empirical Methods in NLP*, ACL, 2003.

Wang H., Lu Y., Zhai C., « Latent Aspect Rating Analysis on Review Text Data : A Rating Regression Approach », *ACM SIGKDD*, p. 783-792, 2010.

Whitehead M., Yaeger L., « Building a General Purpose Cross-Domain Sentiment Mining Model », *IEEE Computer Science and Information Engineering*, p. 472-476, 2009.

ANNEXE POUR LE SERVICE FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :

CORIA 2013

2. AUTEURS :

Abdelhalim Rafrafi — Vincent Guigue — Patrick Gallinari

3. TITRE DE L'ARTICLE :

Classification de Sentiments Multi-Domaines en Contexte Hétérogène & Passage à l'Echelle

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :

Classification de Sentiments

5. DATE DE CETTE VERSION :

11 décembre 2012

6. COORDONNÉES DES AUTEURS :

– adresse postale :
Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie, Paris 6 - 4 place Jussieu F-75252 PARIS cedex 05
{abdelhalim.rafrafi, vincent.guigue, patrick.gallinari}@lip6.fr
– téléphone : 00 00 00 00 00
– télécopie : 00 00 00 00 00
– e-mail : CORIA

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :

LaTeX, avec le fichier de style `article-hermes.cls`,
version 1.2 du 2012/06/04.

8. FORMULAIRE DE COPYRIGHT :

Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
`http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com