Construction automatique de ressources lexicales pour la fouille d'opinion

Extension aux n-grammes

Yves Bestgen

Centre for English Corpus Linguistics
F.R.S-FNRS et Université catholique de Louvain,
10, Place Cardinal Mercier – 1348 Louvain-la-Neuve – Belgique
yves.bestgen@uclouvain.be

RÉSUMÉ. *De nombreuses méthodes automatiques de fouille d'opinion s'appuient sur un lexique dans lequel à chaque entrée est associé un degré de polarité. La construction de telles ressources linguistiques est donc devenue un champ de recherche important en linguistique computationnelle. Des techniques automatiques, basées sur les similarités sémantiques entre les mots dont on veut estimer la polarité et des mots dont la polarité est connue, ont été développées ces dix dernières années et leur efficacité a été confirmée. La présente recherche propose d'étendre de telles procédures à l'analyse de n-grammes de mots et de déterminer les éventuels bénéfices apportés par cette extension. Une expérience menée sur la base d'un vaste corpus de critiques de films indique non seulement que les n-grammes semblent être des indicateurs plus fiables de polarité que les mots simples, mais aussi que leur combinaison avec ces mêmes mots simples permet une meilleure prédiction de la polarité de textes. Soulignant le caractère exploratoire et les limitations de la présente étude, la conclusion propose quelques pistes pour des recherches futures.*

MOTS-CLÉS : *fouille d'opinion, lexique de polarité, n-grammes de mots, analyse sémantique latente, critiques de films.*

KEYWORDS: *opinion mining, polarity lexicon, word n-grams, Latent semantic analysis, movie reviews*

1. Introduction et état de l'art

La fouille d'opinion, qu'il s'agisse de détecter si un énoncé est subjectif ou de déterminer s'il exprime une opinion plus ou moins favorable, connaît depuis une dizaine d'années un développement important en raison de ces nombreuses applications (Abbasi *et al.*, 2008 ; Bestgen, 2002 ; Pang *et al.*, 2002 ; Pang et Lee, 2008 ; Popescu et Etzioni, 2005 ; Sun *et al.*, 2010 ; Taboada *et al.*, 2011 ; Turney, 2002 ; Wiebe *et al.*, 2004 ; Yu et Hatzivassiloglou, 2003). La plupart des techniques développées requièrent des lexiques dans lesquels à chaque entrée est associé un degré de polarité (p.e., *minable : ---, faible : -, correct : +, brillant :* +++). Ces lexiques sont employés tels quels par les systèmes qui attribuent une polarité globale aux textes selon des statistiques sur la présence de mots subjectifs (Bestgen, 2008 ; Chesley *et al.*, 2006 ; Ghorbel et Jacot, 2011 ; Turney, 2002). D'autres approches prennent en compte des phénomènes syntaxiques qui viennent modifier l'orientation sémantique de mots (Brooke, 2009 ; Gala et Brun, 2012 ; Vernier *et al.*, 2009 ; Wilson *et al.*, 2005). Pour ces approches aussi, "la qualité du lexique à la base du système reste cruciale" comme l'ont souligné encore tout récemment Gala et Brun (2012). La construction de telles ressources linguistiques est donc devenue un champ de recherche important en fouille d'opinion, renouvelant des travaux nombreux en analyse de contenu et, surtout, en psycholinguistique. De tels lexiques, appelés dictionnaires ou normes dans ces disciplines, y font l'objet de recherches intensives depuis plus de 40 ans (Proctor et Vu, 1999 ; Wilson, 1988).

La procédure la plus simple et la plus évidente pour construire ces lexiques consiste à demander à des personnes d'évaluer des listes de mots sur des échelles allant par exemple de *évoquant une idée très désagréable* à *évoquant une idée très agréable*. Cette procédure est extrêmement coûteuse en temps et en ressources, ce qui réduit fortement le nombre de mots qui composent les lexiques. Or, plus la couverture du lexique est large, plus le nombre de mots identifiés dans un texte est élevé, et meilleure est la prédiction de la polarité (Bestgen, 1994). Pour dépasser ces limitations, des techniques automatiques ou semi-automatiques ont été proposées en linguistique computationnelle, la majorité d'entre elles s'appuyant sur les similarités sémantiques entre les mots dont on veut estimer la polarité et des mots dont la polarité est connue.

Deux grands types d'approches peuvent être distingués : les approches basées sur des ressources linguistiques, comme WordNet, et celles basées sur des corpus de textes. Les premières calculent généralement la similarité entre les mots sur la base de leur relation de synonymie (par exemple : Esuli et Sebastiani, 2006 ; Hu et Liu, 2004 ; Kamps *et al.*, 2004 ; Kim *et* Hovy, 2004). Partant d'un ensemble de mots germes (*seed words*) dont la polarité est connue, elles emploient un algorithme de *bootstrapping* qui parcourt les liens synonymiques et antonymiques de la ressource linguistique et attribue la même orientation aux mots synonymes et vice-versa. La limitation principale de ces approches est qu'elles sont totalement dépendantes de la qualité des informations disponibles dans la ressource linguistique et de la quantité

de termes présents dans celle-ci. Les mots rares dans une langue ou ceux qui ont un emploi spécifique dans un domaine d'application (*abyssal*, *désopilant*, *navet* ou encore *nullard* et *chichiteux* dans des critiques de films) y sont rarement pris en compte.

Les approches qui s'appuient sur des corpus partent également d'un ensemble de mots germes dont la polarité est connue a priori, mais calculent les similarités au moyen d'une analyse statistique des contextes de cooccurrences des mots, avec ou sans recours à des modèles d'espaces vectoriels représentant le contenu sémantique des mots sous la forme de vecteurs. Un mot est considéré comme d'autant plus positif qu'il est plus proche des germes positifs et plus éloigné des germes négatifs. Pouvant être appliquées à des corpus de très grandes tailles (des milliards de mots récoltés sur le web comme l'ont fait Velikovich *et al.*, 2010) ou à des corpus spécifiques à une application (Bestgen, 2008), ces techniques répondent aux limitations des approches basées sur des ressources linguistiques, mais s'appuient sur des informations plus bruitées que celles qui sont encodées manuellement par des linguistes dans des ressources comme Wordnet ou comme un dictionnaire de synonymes. Récemment, Bestgen et Vincze (2012) ont montré que les approches basées sur un corpus permettaient de générer des normes lexicales ayant une fiabilité équivalente à celles obtenues en demandant à des juges d'évaluer les mots.

Une des limites de ces techniques automatiques d'estimation de la polarité est que le lexique produit ne contient le plus souvent que des mots simples et donc aucun groupe de mots, ou n-grammes de mots, alors que la polarité de celles-ci peut être très différente de celle des mots qui les composent (Polanyi et Zaenen, 2003). A titre d'exemple, on peut trouver dans un corpus de critiques de films : *courrez le voir*, *le plus beau navet*, *faussement subversif*, *d'habitude excellent* ou encore *jamais ridicule*. Si quelques études proposent des techniques d'estimation de la polarité qui s'appliquent aux n-grammes (Turney, 2002; Velikovich *et al.*, 2010; Vernier *et* Monceaux, 2010), aucune n'a essayé de déterminer le gain apporté par ceux-ci par rapport à un lexique composé exclusivement de mots. Ainsi, la technique développée par Velikovich *et al.* (2010), qui est basée sur la construction d'un graphe à partir des cooccurrences de mots et d'unités polylexicales dans un corpus de 4 milliards de pages web, a été évaluée dans une tâche de classification de phrases annotées manuellement, mais sans distinguer les mots simples des séquences composées de plusieurs mots. De même, Turney (2002) a proposé d'attribuer des scores de polarité à des bigrammes sur la base de leur proximité avec des mots germes positifs et négatifs, proximités mesurées à partir des réponses à des requêtes envoyées au moteur de recherche AltaVista. Il a montré que cette technique permet de prédire significativement la note attribuée à des critiques par leur auteur, mais, tout comme Velikovich *et al.* (2010), Turney ne compare pas les bigrammes aux unigrammes. Or, s'il n'est pas difficile de trouver des exemples d'expressions polylexicales dont la polarité est différente de celle des mots qui la composent, on peut se demander si leur prise en compte est susceptible d'apporter un gain significatif. Afin d'essayer d'apporter une première réponse à cette question, la

présente recherche propose d'étendre à l'analyse de n-grammes deux procédures efficaces pour estimer la polarité de mots et de déterminer les éventuels bénéfices apportés par cette extension. La section suivante présente l'approche proposée. Ensuite, une expérience évaluant son efficacité est rapportée.

Dans la suite, *n-gramme* est employé pour faire référence aux séquences d'au moins deux mots (n>1) et s'oppose ainsi à *mot*. *Terme*, conformément à une de ses définitions, est employé pour désigner un mot ou un groupe de mots et désigne donc tant les mots que les n-grammes.

2. Extension de techniques d'estimation de la polarité à des n-grammes

L'approche proposée dans cette étude pour estimer la polarité de n-grammes repose sur des techniques développées indépendamment par Turney et Littman (2003) et Bestgen (2002, 2008) pour estimer la polarité de mots. Elle s'inspire aussi de la technique développée par Turney (2002) pour estimer la polarité de bigrammes. Telle quelle, celle-ci n'est toutefois plus utilisable parce qu'elle repose sur des requêtes adressées au moteur de recherche AltaVista sur la base de l'opérateur "near", un service qui n'est plus disponible depuis plusieurs années.

2.1 SO-LSA

La technique proposée par Turney et Littman pour estimer la polarité de mots, SO-LSA pour *Semantic orientation* - *Latent Semantic Analysis*, se base sur la proximité sémantique entre le mot cible et 14 points de repère : 7 à polarité positive (*good*, *excellent*, ...) et 7 à polarité négative (*bad*, *poor*, ...). Un mot est d'autant plus positif qu'il est plus proche des points de repère positifs et plus éloigné des points de repère négatifs.

Pour mesurer la proximité sémantique sur la base d'un corpus, Turney et Littman ont recours à l'analyse sémantique latente (*Latent Semantic Analysis — LSA*), une technique mathématique qui vise à extraire un espace sémantique de très grande dimension à partir de l'analyse statistique des cooccurrences dans un corpus de textes (Deerwester *et al.*, 1990). Le point de départ de l'analyse est un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chaque segment de textes. Ce tableau fait l'objet d'une décomposition en valeurs singulières qui en extrait les dimensions orthogonales les plus importantes. Dans cet espace, le sens de chaque mot est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, ou dans le cas présent entre un mot et un point de repère, on calcule le cosinus entre les vecteurs qui les représentent. Plus un mot est sémantiquement proche d'un point de repère, plus les deux vecteurs pointent dans la même direction et donc plus leur cosinus se rapproche de 1. Pour estimer la polarité d'un mot, on calcule les cosinus entre ce mot et chaque point de repère positif ainsi que les cosinus entre ce mot et chaque point de repère négatif. La différence entre ces deux ensembles de cosinus indique la polarité du mot.

Turney et Littman ont évalué l'efficacité de leur technique en comparant l'orientation prédite à celle définie dans le *General Inquirer Lexicon* (Stone *et al.*, 1966) qui contient une liste de 3596 mots étiquetés comme positifs ou négatifs. Calculée sur la base de corpus de 10 000 000 de mots, SO-ASL étiquette correctement 65 % des mots.

2.2 DIC-LSA

La technique proposée par Bestgen (2002, 2008; Bestgen et Vincze, 2012), DIC-LSA pour *Dictionnary — Latent Semantic Analysis*, est très similaire à celles de Turney et Littman. La principale différence est que SO-LSA emploie comme points de repère quelques mots sélectionnés a priori alors que DIC-LSA sélectionne un ensemble spécifique de points de repère pour chaque mot à évaluer parmi plusieurs milliers de mots dont la polarité est connue. Plus précisément, DIC-LSA s'appuie sur un dictionnaire de 3000 mots dont la polarité a été évaluée par une trentaine de juges (Hogenraad *et al.*, 1995). La polarité inconnue d'un mot correspond à la polarité moyenne de ses 30 plus proches voisins dont la polarité est connue. Ici aussi, les plus proches voisins sont identifiés sur la base d'une analyse sémantique latente et correspondent aux 30 mots ayant le plus grand cosinus avec le mot cible. Pour évaluer cet indice, Bestgen et Vincze (2012) ont comparé les valeurs prédites par la technique aux valeurs réelles telles que renseignées dans le dictionnaire et ont obtenu une corrélation de 0.71.

2.3 Extension aux n-grammes

Telles quelles, ni SO-LSA, ni DIC-LSA ne peuvent estimer la polarité de n-grammes parce que le tableau lexical qu'elles analysent ne contient que des mots. Il n'y a en soi aucun problème à construire un tableau lexical incluant des bigrammes (ou des trigrammes...), mais ce n'est pas utile parce qu'on ne dispose pas de bigrammes points de repère pour la polarité. Pour contourner ce problème, la procédure proposée consiste à employer comme termes lors de la construction de l'espace sémantique tant les mots que les bigrammes. De cette manière, le cosinus entre chaque bigramme et les points de repère peut être aisément calculé et, donc, la procédure habituelle est applicable pour estimer leur polarité. Cette procédure est également applicable aux trigrammes, quadrigrammes... Cette approche n'a toutefois de sens que si la polarité des mots n'est que très peu modifiée lorsque ceux-ci sont soumis à une analyse sémantique latente en même temps que des n-grammes. Vérifier cette condition est un des objectifs de l'expérience rapportée ci-dessous. Plus généralement, celle-ci vise à évaluer l'efficacité de SO-LSA et de DIC-LSA pour estimer la polarité de n-grammes en la comparant à celle de ces mêmes techniques pour des mots.

3. Expérience

3.1 Objectifs

Afin de se faire une première idée de l'intérêt qu'il y a d'étendre l'estimation automatique de la polarité de termes à des n-grammes, la présente expérience tente d'abord de répondre à la question suivante : lorsque la technique prétend qu'un n-gramme est positif (ou négatif), l'est-il vraiment et surtout l'est-il plus souvent que lorsque cette technique fait la même prédiction pour un mot? Pour répondre à cette question, il n'est pas possible d'employer la procédure classiquement utilisée pour évaluer les mots. En effet, celle-ci consiste à comparer les polarités prédites à celles obtenues en demandant à des juges d'évaluer des mots sur cette dimension, comme c'est le cas dans les normes Valemo, F-POL ou encore le General Inquirer dans sa version anglaise ou française (Turney et Littman, 2003; Vincze et Bestgen, 2011). Or, à ma connaissance, aucune norme de ce type n'est disponible pour les n-grammes. Pour pouvoir néanmoins évaluer la qualité de ceux-ci, la solution adoptée consiste à employer des textes dont la polarité est connue parce que les auteurs de ceux-ci leur ont attribué une note sur une échelle de polarité, comme c'est fréquemment le cas pour les critiques de films. Si une procédure d'estimation de la polarité est efficace, on doit s'attendre à observer plus fréquemment les termes (mots ou n-grammes) les plus négatifs dans les critiques négatives et les termes les plus positifs dans les critiques positives.

Cette première manière d'évaluer l'extension de SO-LSA et DIC-LSA aux n-grammes peut être qualifiée d'intrinsèque en ce sens qu'elle vise à déterminer la qualité des scores de polarité attribués automatiquement à des n-grammes en dehors de toute application concrète. La deuxième analyse vise à déterminer si la prise en compte de n-grammes en plus des mots permet d'améliorer la prédiction de la polarité de textes.

3.2 Méthode

3.2.1 Corpus pour l'extraction des espaces sémantiques et pour la phase de test

Bestgen (2008) ayant montré que tant SO-LSA que DIC-LSA étaient plus efficaces pour estimer des polarités lorsque le corpus à partir duquel l'espace sémantique est extrait contient des textes similaires à ceux qui doivent être évalués lors de la phase de test, les textes employés tant pour extraire les espaces sémantiques que pour la phase de test ont été extraits d'un même site Web, le site Allocine.fr. Ce site rassemble de nombreuses informations sur le cinéma, dont une ample section de partage social dans laquelle les spectateurs peuvent donner leur avis à propos de films.

Un programme Perl a été employé pour sélectionner dans la base de données tous les films produits entre 2006 et septembre 2012 qui ne relevaient pas des catégories *Divers* ou *Documentaires*, pour lesquels on disposait d'au moins une

critique de presse (de sorte d'éliminer des séries télévisées) et qui avaient été commentés au moins 50 fois. Ce programme extrayait toutes les critiques de spectateurs disponibles en enregistrant le titre du film, le rédacteur de la critique, la date de diffusion de la critique, le texte de celle-ci et la note attribuée par l'auteur. Cette note, au moment de la construction du corpus, était présentée sur une échelle allant de 0 à 5 par pas d'un demi-point, 0 correspondant à un film *nul* et 5 à un *chef-d'oeuvre*.

Plusieurs post-traitements ont été effectués sur les textes des critiques afin de transcoder ou de recoder des caractères spéciaux et de supprimer quelques balises html. Ensuite, chaque critique a été traitée par le programme TreeTagger (Schmid, 1994) afin d'être segmentée en *tokens* (mots, mais aussi signes de ponctuation, symboles...) et lemmatisée. En moyenne, la longueur des critiques est de 86 tokens avec un écart-type très élevé de 92 et une étendue allant de 1 à 4434 tokens. Afin de réduire cette étendue, les critiques de moins de 30 tokens et de plus de 418 ont été supprimées, ces valeurs éliminant les critiques considérées arbitrairement comme trop courtes pour les inclure dans le tableau lexical et le pour-cent (1%) des critiques les plus longues.

Le corpus intégral est composé de 648 619 critiques à propos de 1837 films différents et compte un peu plus de 65 millions de tokens. Comme indiqué ci-dessus, ce corpus a été employé pour l'extraction des espaces sémantiques et pour la phase de test. Afin de garantir autant d'indépendance que possible entre l'étape d'estimation des polarités et l'étape d'évaluation, le corpus intégral a été divisé aléatoirement en deux sous-corpus (A et B) contenant chacun approximativement 324 000 critiques. Le corpus B a été employé pour l'évaluation des polarités estimée sur la base d'une analyse sémantique latente du corpus A et l'inverse a été effectué lorsque le corpus A est utilisé pour la phase de test. Les résultats donnés dans la suite correspondent toujours à la moyenne des scores obtenus sur chaque demi-corpus.

3.2.2 Sélection des mots et des n-grammes

Pour la construction des tableaux lexicaux et la phase de test, on n'a conservé dans chaque critique que les mots apparaissant au moins cinq fois dans l'ensemble du corpus et appartenant à une des catégories grammaticales suivantes : nom, adjectif, verbe et conjonction. Pour les termes composés de plus d'un mot, les mêmes conditions s'appliquent à chacun des mots qui les composent ainsi que la nécessité que tous les mots se suivent directement et fassent partie de la même phrase. On a choisi dans la présente étude de se limiter aux n-grammes composés au maximum de quatre mots.

3.2.3 Construction des espaces sémantiques et estimation de la polarité

Les matrices de cooccurrences *critiques* × *termes* ont été soumises à une décomposition en valeurs singulières réalisée par le programme SvdpackC (Berry *et al.*, 1993) et les 300 premiers vecteurs propres ont été conservés, ce nombre étant

généralement considéré comme un optimum (Landauer *et al.*, 2004). Ces vecteurs ont été employés pour mesurer les proximités entre les termes et les points de repère.

- Pour SO-LSA, les 14 points de repère de Turney et Littman (2003) ont été traduits en français. Il s'agit de *bon, gentil, excellent, positif, heureux, correct, supérieur* et de *mauvais, méchant, médiocre, négatif, malheureux, faux, inférieur*.

- Pour DICLSA, le dictionnaire de polarité est composé de 3 252 mots, chacun évalué par une trentaine de personnes (étudiants dans des établissements d'enseignement supérieur) sur une échelle à 7 points selon que le mot évoque une idée allant de *très désagréable* (1) à t*rès agréable* (7) (Hogenraad *et al.*, 1995). Le tableau 1 donne les valeurs attribuées à quelques mots extraits aléatoirement de ce dictionnaire.

Mot	Polarité	Mot	Polarité
détresse	1.4	contrôlable	3.5
imbécile	1.4	outil	4.3
tristesse	1.6	risquer	4.5
hostilité	2.2	entier	4.9
impassible	2.6	revenir	5.0
superstitieux	2.8	admiratif	5.7
hâtes	3.1	doux	6.0
ambigus	3.2	sincérité	6.1

Tableau 1 : *Polarité de mots sur une échelle allant de très désagréable (1.0) à très agréable (7.0)*.

3.3 Analyses et résultats

Au total, les techniques ont attribué une polarité à un peu plus de 18000 mots, 105 000 bigrammes, 59 000 trigrammes et 17 000 quadrigrammes. A titre d'illustration, quelques termes ayant reçu les polarités les plus négatives et le plus positives sont présentés ci-dessous.

- Mots : (-) *consterner, nullité, pitoyablement, ratage* ; (+) *bijou, épatant, envoûter, fabuleux.*

- Bigrammes (2G) : (-) *absolument affligeant, bouse intersidéral, cliché pleuvoir, navet total* ; (+) *air frais, aussi rocambolesque, découvrir absolument, dépaysement total.*

- Trigrammes (3G) : (-) *encore plus bas, faire jamais peur, seul aspect positif, tout simplement rater* ; (+) *avoir vraiment bouleverser, beau et attachant, pas hilarant mais, vrai petit bijou.*

- Quadrigrammes (4G) : (-) *absolument rien ne être, avoir rien avoir sauver, ne faire jamais peur, pas trop mauvais mais* ; (+) *drôle mais aussi émouvant, paraître plus vrai que, personnage être terriblement attachant, simple mais pas simpliste.*

3.3.1 Analyse 1 : Impact de la présence conjointe de mots et de n-grammes sur l'estimation de la polarité

La première analyse a pour objectif de vérifier que la présence dans un espace sémantique des n-grammes en plus des mots ne modifie pas les scores de polarité attribués aux mots. Pour répondre à cette question, on a corrélé les scores de polarité des mots obtenus sur la base des quatre espaces sémantiques : celui n'incluant que des mots et ceux incluant en plus des mots un type de n-grammes. Ces corrélations sont très élevées puisque toutes sont supérieures à 0.93 tant pour SO-LSA que pour DIC-LSA.

3.3.2 Analyse 2 : Efficacité de la procédure pour estimer la polarité de n-grammes

La deuxième analyse vise à déterminer si les n-grammes les plus négatifs s'observent dans les critiques les plus négatives et les n-grammes les plus positifs dans les critiques les plus positives. Pour répondre à cette question, les critiques de chaque corpus de test ont été divisées en trois tranches selon que la note attribuée par l'auteur est inférieure ou égale à 1.5, supérieure ou égale à 3.5 ou intermédiaire, cette dernière tranche n'étant pas employée dans les analyses. Les mots et les n-grammes de chaque longueur ont été indépendamment rangés du plus négatif au plus positif et divisés en tranches en fonction des pourcentiles[1] suivants : 1 %, 2.5 %, 5 %, 10 %, 25 %, 33.33 %, 66.67 %, 75 %, 90 %, 95 %, 97.5 % et 99 %. Pour les termes identifiés comme négatifs, la première tranche (1 %) inclut le pour cent des valeurs les plus négatives, la deuxième (2.5 %) le pour cent et demi suivant, la troisième (5 %) les 2.5 % qui suivent et ainsi de suite. Pour les termes identifiés comme positifs, la tranche 99 % inclut le pour cent des valeurs les plus positives, la tranche 97.5 % le pour cent et demi suivant et ainsi de suite. De cette manière, chaque terme est inclus dans une et une seule tranche. Les termes appartenant au tiers le plus neutre (de 33.34 % à 66.66 %) n'ont pas été analysés.

On a ensuite déterminé pour chaque terme de chaque tranche dans combien de critiques négatives et dans combien de critiques positives il apparaît. Si la procédure est efficace, on devrait observer plus de termes négatifs dans les critiques négatives et plus de termes positifs dans les critiques positives. L'indice d'efficacité pour un terme est donc la proportion de cas dans lesquels il apparaît dans une critique de polarité attendue. Pour les termes négatifs, il correspond au rapport entre le nombre d'occurrences dans des critiques négatives divisé par le nombre total d'occurrences du terme. Pour les critiques positives, le numérateur est le nombre d'occurrences du terme dans des critiques positives.

[1] Le pourcentile est défini comme la valeur sous lequelle on trouve ce pourcentage de valeurs.

Les tableaux 2 et 3 présentent les scores d'exactitude moyens pour les 4 types de termes et les 12 tranches calculés sur les deux sous-corpus de test et ce pour les deux techniques d'estimation.

	Tranche	Mot	2G	3G	4G
---	1.0	0.73	0.76	0.76	0.74
	2.5	0.60	0.65	0.68	0.71
	5.0	0.53	0.55	0.60	0.61
	10.0	0.47	0.48	0.53	0.55
	25.0	0.41	0.39	0.42	0.45
-	33.0	0.37	0.33	0.35	0.37
+	67.0	0.71	0.77	0.79	0.80
	75.0	0.72	0.79	0.82	0.84
	90.0	0.75	0.82	0.86	0.88
	95.0	0.77	0.83	0.87	0.92
	97.5	0.79	0.85	0.90	0.94
+++	99.0	0.83	0.88	0.92	0.96

Tableau 2 : *Scores d'exactitude moyens pourSO-LSA*

	Tranche	Mot	2G	3G	4G
---	1.0	0.49	0.47	0.49	0.55
	2.5	0.52	0.53	0.58	0.62
	5.0	0.47	0.50	0.56	0.56
	10.0	0.43	0.44	0.49	0.51
	25.0	0.38	0.37	0.41	0.43
-	33.0	0.36	0.34	0.35	0.37
+	67.0	0.69	0.75	0.77	0.78
	75.0	0.70	0.78	0.81	0.83
	90.0	0.72	0.81	0.86	0.89
	95.0	0.75	0.83	0.88	0.91
	97.5	0.78	0.85	0.90	0.93
+++	99.0	0.81	0.89	0.92	0.95

Tableau 3 : *Scores d'exactitude moyens pour DIC-LSA*.

On observe logiquement que les scores d'exactitude sont plus élevés pour les termes ayant les polarités les plus extrêmes, mais aussi qu'ils sont plus élevés pour les n-grammes que pour les mots et également plus le paramètre *n* est grand. On note aussi que les scores sont nettement meilleurs pour les termes positifs que pour les termes négatifs. Pour les termes négatifs, l'exactitude est même souvent inférieure à la valeur que produirait une attribution de la polarité par une procédure

totalement aléatoire (0.50). Une performance aussi faible n'était pas attendue. Enfin, si SO-LSA est très nettement plus efficace (ou moins inefficace) que DICLSA pour les mots négatifs, les différences pour les n-grammes positifs sont très faibles et pas toujours à son avantage.

3.3.3 Analyse 3 : Utilité des n-grammes pour prédire la polarité de textes

La dernière analyse vise à déterminer si la prise en compte de n-grammes en plus des mots permet d'améliorer la prédiction de la polarité de textes. Pour ce faire, j'ai employé une procédure d'estimation de la polarité d'un texte très simple, mais relativement efficace (Turney, 2002; Bestgen, 1994, 2008), qui considère que la polarité d'un texte est égale à la moyenne de toutes les polarités connues. Comme dans l'analyse précédente, les polarités des termes ont été calculées sur un des deux sous-corpus et le test a porté sur l'autre sous-corpus. Le tableau 4 donne les corrélations entre la note attribuée aux critiques par les auteurs et la polarité estimée sur la base des mots, des unigrammes et des bigrammes (Mot-2G), des unigrammes, bigrammes et trigrammes (Mot-2G-3G) et de l'ensemble des termes disponibles (Mot-2G-3G-4G).

Ces valeurs peuvent être comparées à un niveau de base : la corrélation obtenue lorsqu'on estime la polarité des critiques au moyen du dictionnaire de polarité original, composé de 3253 mots, chacun évalué par une trentaine de personnes. Cette corrélation n'est que de 0.37. Il apparaît qu'étendre une norme de polarité même seulement à des mots améliore nettement la qualité de la prédiction puisqu'on passe d'une corrélation de 0.37 à 0.50. L'adjonction des bigrammes aux mots améliore aussi la prédiction, mais moins fortement. Le gain pour les n-grammes de taille supérieure est nettement plus limité. On observe aussi que SO-LSA est légèrement plus efficace que DIC-LSA.

	Mot	Mot-2G	Mot-2G-3G	Mot-2G-3G-4G
SO-LSA	0.52	0.58	0.58	0.58
DIC-LSA	0.50	0.54	0.55	0.55

Tableau 4 : *Corrélations entre les polarités des critiques estimées par les procédures automatiques et les notes attribuées par les auteurs*.

Une série d'analyses complémentaires ont été menées en faisant varier les termes employés pour prédire la polarité des critiques de manière à ne prendre en compte que les termes les plus extrêmes ou seulement les termes positifs. Cette dernière option a donné lieu à des performances très faibles. Se baser exclusivement sur les termes les plus extrêmes donne lieu à des corrélations très légèrement meilleures, mais a aussi pour conséquence de réduire le nombre de critiques auxquelles la

technique peut attribuer un score parce que certaines critiques ne contiennent plus aucun terme dont la polarité a été estimée. Une comparaison avec les résultats obtenus sur l'ensemble du matériel de test devient donc discutable.

4. Discussion et conclusion

Cette recherche a pour objectif principal d'évaluer l'intérêt qu'il y a d'étendre l'estimation automatique de la polarité de termes à des n-grammes. Deux procédures efficaces pour estimer la polarité de mots ont été modifiées de manière à pouvoir attribuer un score de polarité à ce type d'expressions. L'expérience rapportée indique que plus un n-gramme contient de mots, plus fréquemment il est observé dans une critique de même polarité que lui. Les valeurs obtenues pour les n-grammes positifs, surtout pour les expressions les plus extrêmes, peuvent être considérées comme très élevées étant donné qu'il n'est pas rare que des critiques globalement négatives mentionnent néanmoins quelques qualités et que, comme d'autres chercheurs l'ont déjà noté (Ghorbel et Jacot, 2011), un matériel de test composé de critiques de films n'est pas exempt d'erreurs. La critique reprise ci-dessous, trouvée par hasard sur le site allociné.fr et ne faisant pas partie du matériel expérimental, en est un exemple particulièrement manifeste : *1 - Très mauvais : Du grand Claude Chabrol. Une merveille ce film, une bouffée de bonheur ! Les acteurs s'amusent et le résultat est convaincant; Philippe Noiret atteint la quasi-perfection dans son rôle d'animateur-télé-hypocrite, Robin Renucci est littéralement fondant, Bernadette Lafont est parfaitement horripilante dans son rôle de masseuse et Anne Brochet s'en tire bien pour un rôle difficile à jouer. Une trame qui tient la route et un décor parfait viennent s'ajouter à un film hors du commun !*

Cette analyse a aussi mis en évidence un résultat inattendu. Les termes identifiés comme négatifs par la procédure automatique sont nettement moins souvent associés à des critiques négatives que les termes positifs à des critiques positives et ceci est vrai tant pour les mots que les n-grammes. On peut se demander si ce résultat ne devrait pas être interprété, en partie au moins, comme une faiblesse de la procédure d'évaluation et non de la procédure d'estimation des polarités. En effet, on trouve parmi les n-grammes négatifs fréquents dans des critiques positives des expressions comme *pas vraiment mauvais mais*, *pouvoir être meilleur si* ou encore *être trop lisse*. Ces n-grammes sont clairement négatifs, mais ils ont leur place dans une critique globalement positive qui souligne quelques éléments plus négatifs. Ce résultat s'explique sans doute aussi par le type de textes employés dans l'expérience. Comparant différents types de critiques (film, banque, voiture...), Turney (2002) a observé que, contrairement aux autres types, le contenu du film (film d'horreur, drame social) peut aussi avoir une forte polarité sans que celle-ci soit liée au point de vue de l'auteur de la critique à propos de celui-ci : on peut apprécier fortement un film qui vous remplit d'effroi et on peut être heureux d'avoir versé de nombreuses larmes sur les malheurs des personnages. Il serait donc intéressant de reproduire l'expérience avec un matériel composé de critiques d'un tout autre type afin de

déterminer si, dans celui-ci aussi, les termes identifiés comme négatifs sont peu souvent associés à des critiques négatives. Quoiqu'il en soit, une étude approfondie de l'estimation de la polarité de termes négatifs est indispensable.

La dernière analyse indique que la prise en compte des n-grammes permet d'améliorer la prédiction de la polarité de critique de films. Dans cette analyse aussi, les résultats peuvent être vus comme très bons étant donné la nature particulièrement simpliste de la procédure d'assignation d'une polarité aux critiques : la simple moyenne de toutes les polarités connues dans le texte en question. Il n'en reste pas moins que le bénéfice le plus important s'observe lors de l'adjonction des bigrammes aux mots et que ni les trigrammes, ni les quadrigrammes n'ont un effet supplémentaire important.

Enfin, les différentes analyses effectuées mettent en évidence une légère supériorité de SO-LSA sur DIC-LSA. Cette observation s'explique probablement par le fait que les points de repère employés dans SO-LSA ont, en partie au moins, été sélectionnés par Turney et Littman (2003, Turney, 2002) parce qu'ils étaient justement associés à des critiques positives (*excellent*, *bon*) ou à des critiques négatives (*mauvais*, *médiocre*).

Globalement, les résultats obtenus plaident en faveur de l'intégration de n-grammes dans les lexiques de polarité construits par une procédure automatique à partir d'un corpus. Non seulement les n-grammes semblent être des indicateurs plus fiables de polarité que les mots (analyse 2), mais leur combinaison avec ces mêmes mots permet une meilleure prédiction de la polarité de textes (analyse 3). Il faut toutefois rester prudent en raison du caractère exploratoire de la présente étude et de ses limitations. En premier lieu, les résultats ont été obtenus par l'analyse d'un seul type de textes : des critiques déposées sur des sites de partage social. Les généraliser à d'autres types de textes est indispensable. L'intérêt majeur des techniques employées est qu'elles ne prennent pas en compte les notes attribuées aux critiques lors de l'estimation de la polarité. Ces techniques peuvent donc être appliquées à tout corpus, que celui-ci inclue ou non une évaluation de l'opinion émise dans le texte. Des billets de blogs, des commentaires de lecteurs et même des articles de presse peuvent être soumis à ces techniques pour étendre un lexique de polarité. Il sera seulement nécessaire d'utiliser une autre procédure d'évaluation des performances que celle employée dans la présente recherche.

Ensuite, les espaces sémantiques utilisés pour estimer les polarités sont uniquement extraits d'un corpus spécifique au matériel de test. Ce choix a permis par exemple de détecter comme très négatifs des termes tels que *truelle* que l'on trouve employée dans des expressions comme *Les blagues sont torchées à la truelle* ou *soulignant à la truelle*. Il serait néanmoins intéressant de comparer l'efficacité des procédures d'estimation sur la base d'un corpus spécifique ou d'un corpus générique afin de déterminer si les observations de Bestgen (2008), obtenues dans une analyse de phrases extraites d'articles de presse, peuvent être reproduites avec des critiques.

Une autre question, plus fondamentale, à laquelle cette recherche ne répond pas est de déterminer si la construction des lexiques de polarité est le lieu où prendre en compte les facteurs contextuels qui affectent la polarité des mots. L'autre option consiste à laisser le traitement des effets contextuels pour d'autres modules du système d'analyse de l'opinion comme c'est fréquemment le cas pour la négation. L'intérêt de l'approche basée sur les n-grammes est que les effets contextuels sont (potentiellement) découverts de manière automatique et non au travers de l'analyse manuelle et fine d'un corpus de textes. Il reste toutefois à prouver que cette manière de procéder est aussi efficace que le recours à des modules spécifiques.

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS. Il tient à remercier Nadja Vincze, auteure de la version préliminaire du programme Perl employé pour collecter les critiques.

5. Bibliographie

Abbsasi A., Chen H., Salem A. « Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums », *ACM Transactions on Information Systems* 26, 2008.

Berry M., Do, T., O'brien G., Krishna V., Varadhan S. SVDPACKC : Version 1.0 user's guide, Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

Bestgen Y. « Can emotional valence in stories be determined from words? », *Cognition and Emotion*, *7*, 1994., p. 21-36.

Bestgen Y. « Détermination de la valence affective de termes dans de grands corpus de texte », *Actes de CIFT*'02, 2002, p. 81-94.

Bestgen Y. « Building affective lexicons from specific corpora for automatic sentiment analysis », Proceedings of *LREC 2008*, 496-500.

Bestgen Y.,Vincze N. « Checking and bootstrapping lexical norms by means of word similarity indexes », *Behavior Research Methods*, *44*, 2012, p. 998-1006.

Brooke J. *A Semantic Approach to Automated Text Sentiment Analysis*. Simon Fraser University, 2009.

Chesley P., Vincent B., Xu L., Srihari R.K. « Using verbs and adjectives to automatically classify blog sentiment », Proceedings of *AAAI-CAAW-06*, 2006, p. 27-29.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R.. « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, 1990, p. 391-407.

Esuli A., Sebastiani F., « Sentiwordnet: A publicly available lexical resource for opinion mining », *Proceedings of LREC'06*, 2006, p. 417-422.

Gala N., Brun C. « Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions », Actes de *Traitement Automatique des Langues Naturelles,* 2012. Grenoble, juin 2012.

Ghorbel H., Jacot D. « Sentiment Analysis of French Movie Reviews », In V. Pallotta, A. Soro, and E. Vargiu (Eds.): *Advances in DART*, SCI 361, 2011, p. 97–108.

Hogenraad R., Bestgen Y., Nysten J. L., « Terrorist rhetoric: Texture and architecture », In E. Nissan *et* K.M. Schmidt (Eds.) *From information to knowledge*. *Conceptual and content analysis by computer*, Oxford, England: Intellect Books, 1995, p. 54-67.

Hu M., Liu B.. « Mining Opinion Features in Customer Reviews », Proceedings of *AAAI*, 2004, p. 755-760.

Kamps J., Marx M., « Words with Attitude », *Proceedings of the 1st Interational Conference on Global WordNet,* 2002, p. 332-341.

Kim S. M., Hovy E., « Determining the sentiment of opinions », *Proceedings of COLING*, 2004, p. 1367-1373.

Landauer T. K., Laham D., Derr M., « From paragraph to graph: Latent Semantic Analysis for information visualization », *Proceedings of the National Academy of Science* 101, 2004, p. 5214-5219.

Pang B., Lee L. « Opinion Mining and Sentiment Analysis », *Foundations and Trends in Information Retrieval* , vol. 2, 2008, p. 1-135.

Pang B., Lee L., Vaithyanathan S., « Thumbs up? Sentiment classification using machine learning techniques », *Proceedings of the ACL-02*, 2002, 79-86.

Polanyi L., Zaenen A. « Shifting attitudes ». In L. Lagerwerf, W. Spooren, *et* L. Degand (Eds.), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse 2003* (pp. 61-69). Münster: Nodus Publikationen.

Proctor R. W., Vu, K. L., « Index of norms and ratings published in the Psychonomic Society journals », *Behavior Research Methods, Instruments, and Computers,* vol. 31, 1999, p. 659-667.

Popescu A.-M., Etzioni O., « Extracting Product Features and Opinions from Reviews », *Proceedings of the HLT/EMNLP*, 2005, p. 339-346.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », Proceedings of *the International Conference on New Methods in Language Processing*, 1994, p. 44-49.

Stone P.J., Dunphy D.C., Smith M.S., Ogilvie D.M., *The General Inquirer: A Computer Approach to Content Analysis, MIT Press*, 1966.

Taboada M., Brooke J., Tofiloskiy M. Vollz K., Stede M. « Lexicon-based methods for sentiment analysis », *Computational Linguistics*, 37, 2011, p. 267–307.

Turney P. « Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews », In Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL), 2002.

Turney P., Littman M. « Measuring Praise and Criticism: Inference of Semantic Orientation from Association », *ACM Transactions on Information Systems*, vol. 21, 2003, p. 315-346.

Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R., « The Viability of Web-derived Polarity Lexicons ». *Proceedings of NAACL*, 2010, p. 777-785.

Vernier M., Monceaux L. « Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques », *Traitement automatique des langues* 51, 2010, 125-149.

Vernier M., Monceaux L., Daille B., Dubreil E. « Catégorisation sémantico-discursives des évaluations exprimées dans la blogosphère », Actes de *TALN 2009*.

Vincze N., Bestgen Y., « Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée », *Actes de TALN11*, vol. 1, 2011, p. 223-234.

Wiebe J., Wilson T., Bruce R., Bell M., Martin M., « Learning subjective language », *Computational Linguistics*, vol. 30, 2004, p. 277-308.

Wilson T., Wiebe J., Hoffmann P.. « Rcognizing contextual polarity in phrase-level sentiment analysis », *Proceedings of the 2005 Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 2005, p. 347-354.

Wilson M. D., « The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2 », *Behavioral Research Methods, Instruments and Computers*, vol. 20, 1988, p. 6-11.

Yu H., Hatzivassiloglou V. « Toward answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003, p. 129-136.