Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée

Nadja Vincze[1] Yves Bestgen[2]

(1) UCLouvain, Cental, Place Blaise Pascal, 1, B-1348 Louvain-la-Neuve, Belgique
(2) UCLouvain, CECL, B-1348 Louvain-la-Neuve, Belgique
nadja.vincze@uclouvain.be, yves.bestgen@uclouvain.be

Résumé

De nombreuses méthodes automatiques de classification de textes selon les sentiments qui y sont exprimés s'appuient sur un lexique dans lequel à chaque entrée est associée une valence. Le plus souvent, ce lexique est construit à partir d'un petit nombre de mots, choisis arbitrairement, qui servent de germes pour déterminer automatiquement la valence d'autres mots. La question de l'optimalité de ces mots germes a bien peu retenu l'attention. Sur la base de la comparaison de cinq méthodes automatiques de construction de lexiques de valence, dont une qui, à notre connaissance, n'a jamais été adaptée au français et une autre développée spécifiquement pour la présente étude, nous montrons l'importance du choix de ces mots germes et l'intérêt de les identifier au moyen d'une procédure d'apprentissage supervisée.

Abstract

Many methods of automatic sentiment classification of texts are based on a lexicon in which each entry is associated with a semantic orientation. These entries serve as seeds for automatically determining the semantic orientation of other words. Most often, this lexicon is built from a small number of words, chosen arbitrarily. The optimality of these seed words has received little attention. In this study, we compare five automatic methods to build a semantic orientation lexicon. One among them, to our knowledge, has never been adapted to French and another was developed specifically for this study. Based on them, we show that choosing good seed words is very important and identifying them with a supervised learning procedure brings a benefit.

Mots-clés : Analyse de sentiments, lexique de valence, apprentissage supervisé, analyse sémantique latente
Keywords: Sentiment analysis, semantic orientation lexicon, supervised learning, latent semantic analysis

1 Introduction

La classification de textes consiste à classer automatiquement les textes dans un ensemble prédéfini de catégories. Ce sont initialement les classifications thématiques et par genre qui ont motivé les recherches, mais, depuis une dizaine d'années, ce champ d'études s'est élargi et intègre la classification de textes en fonction des sentiments qui y sont exprimés : détection de la subjectivité, avec une classification objectif / subjectif (Wiebe et al., 2004 ;Yu, Hatzivassiloglou, 2003) et détermination de la valence des documents, avec une classification binaire positif / négatif, parfois multi-classes selon le degré de polarité (Abbasi et al., 2008 ; Pang et al., 2002). La plupart des implémentations de ces classifieurs requièrent des lexiques porteurs de valence, c'est-à-dire des lexiques où à chaque entrée est associée une polarité ou un degré de polarité. Une série d'approches attribuent une valence globale aux textes selon des statistiques sur la présence de mots subjectifs (Bestgen, 2006 ; Turney, 2002). Les approches dites symboliques intègrent la prise en compte de phénomènes syntaxiques qui viennent modifier l'orientation sémantique de mots ou de groupes de mots (Harb et al., 2008 ; Vernier et al., 2009 ; Wilson et al., 2005). Enfin, quelques tentatives d'apprentissages supervisés ont également pris en compte des mots, dont la valence est connue, comme caractéristiques de leurs vecteurs (Chesley et al., 2006). Ces lexiques constituent donc des ressources sémantiques capitales au développement de classifieurs efficaces.

Dans un premier temps, ces lexiques ont été construits manuellement par des juges (Nasukawa, Yi, 2003 ; Wiebe et al., 2005), mais le travail étant lent et couteux, des procédures automatiques ou semi-automatiques ont vu le jour et constituent aujourd'hui un sous-domaine de recherche important. Comme le souligne la présentation des travaux antérieurs (section 2), une spécificité des recherches menées dans ce champ est qu'elles portent presque exclusivement sur l'anglais, langue pour laquelle de nombreuses ressources linguistiques ont été développées comme WordNet (Miller, 1990). Un des deux objectifs principaux de notre étude est de déterminer dans quelle mesure ces méthodes sont applicables au français. Une autre spécificité des recherches menées dans ce champ est que la quasi-totalité des méthodes proposées utilise un petit nombre de mots, comme *bon*, *mauvais*, *gentil*, afin de servir de germes (*seed*) pour déterminer automatiquement la valence d'autres mots (voir par exemple Hu, Liu, 2004; Kamps, Marx, 2002 ; Turney, Littman, 2003). La question de l'optimalité de ces mots germes a bien peu retenu l'attention, le plus souvent les chercheurs reprenant ceux proposés dans des travaux antérieurs (Esuli, Sebastiani, 2006 ; Harb et al., 2008). Notre second objectif est de proposer une méthode permettant d'identifier automatiquement ces germes au moyen d'une technique d'apprentissage supervisée.

Après une brève présentation des travaux antérieurs, la section 3 décrit les différentes méthodes comparées dans le cadre de cette étude. Une série d'expériences visant à évaluer leur efficacité sont présentées dans la section 4. La section 5 rapporte les principaux résultats, dont les implications et les développements possibles sont discutés dans la conclusion.

2 Travaux antérieurs

Parmi les méthodes automatiques ou semi-automatiques proposées pour construire des lexiques porteurs de valences, on peut distinguer deux types d'approches : celles basées sur des ressources linguistiques comme WordNet et celles basées sur des corpus de textes.

Les approches qui s'appuient sur des bases de connaissances linguistiques calculent généralement la similarité entre les mots à partir de leur relation de synonymie. Une méthode de base consiste à partir de quelques mots dont la valence est connue et à lancer un algorithme d'amorçage (*bootstrapping*) qui parcourt les liens synonymiques et antonymiques de la base, en attribuant la même orientation aux mots synonymes et vice-versa (Hu, Liu, 2004 ; Kim, Hovy, 2004). Kamps et Marx (2002) ont probablement été les premiers à proposer une telle procédure en dérivant de WordNet un graphe dans lequel chaque nœud représente un terme et un lien est présent entre deux nœuds s'ils sont synonymes. À partir de ce graphe, ils calculent une valeur normalisée pour les nœuds liés aux mots g*ood* et *bad*. Esuli et Sebastiani (2006) ont étendu cette approche pour développer SentiWordNet, une ressource basée sur WordNet, qui assigne à chaque *synset* trois valeurs normalisées : une positive, une négative et une objective. La spécificité principale de leur approche est qu'elle s'appuie sur un apprentissage semi-supervisé basé sur les définitions de mots germes sélectionnés manuellement.

Ne disposant pas d'informations sur les liens synonymiques, les approches qui s'appuient sur des corpus calculent les similarités différemment. Hatzivassiloglou et McKeown (1997) ont proposé un algorithme capable de déterminer l'orientation sémantique d'adjectifs à partir de l'analyse de leurs cooccurrences avec des conjonctions. Turney et Littman (2003 ; Turney, 2002) et Bestgen (2002, 2008) ont proposé des méthodes plus générales puisqu'elles permettent d'estimer la valence de n'importe quel terme présent dans un corpus. Ils utilisent l'analyse sémantique latente (ASL, *Latent Semantic Analysis*, Deerwester et al., 1990) pour construire un espace sémantique à partir d'informations statistiques sur les cooccurrences de termes dans des textes. Turney et Littman l'emploient pour estimer la distance sémantique entre des mots et 14 mots germes, 7 positifs (*good*, *nice*, *excellent*, *positive*, *fortunate*, *correct*, *superior*) et 7 négatifs (*bad*, *nasty*, *poor*, *negative*, *unfortunate*, *wrong*, *inferior*). Un mot est d'autant plus positif qu'il est plus proche des germes positifs et plus éloigné des germes négatifs. Pour sa part, Bestgen (2002) a recours à l'ASL pour identifier les mots fréquemment associés aux mots dont il veut déterminer la valence affective. Il attribue à chaque mot la valence moyenne de ses plus proches voisins dont la valence est connue. Pour cela, il s'appuie sur un dictionnaire de 3000 mots dont la valence a été évaluée par des juges. On notera que les similarités peuvent être calculées sans passer par l'analyse sémantique latente, mais que, dans ce cas, des corpus de très grande taille semblent nécessaires (Turney, Littman, 2003; Velikovich et al., 2010), sauf si, à la manière de Harb et al. (2008), on emploie un corpus très spécifique et des règles d'associations.

Peu d'initiatives de construction automatique de lexiques ont eu lieu en français, comparé à l'effervescence dans le milieu anglophone. Nous pouvons citer Bestgen (2002) et Chardon (2010) qui a développé une méthode pour élaborer une ressource lexicale d'adjectifs d'opinion à partir d'une liste de mots germes et d'une taxinomie des mots du français. Pak et Paroubek (2010) ont proposé une méthode de construction automatique d'un lexique affectif à partir de messages disponibles sur Twitter. Leur procédure est basée sur la comparaison de la fréquence d'occurrence d'un mot dans les messages contenant une émoticône positive et dans ceux contenant une émoticône négative. Vernier et Monceaux (2010) ont proposé une méthode d'apprentissage pour enrichir automatiquement un lexique subjectif à partir d'un corpus annoté. L'apprentissage automatique se base sur des tests sémantiques, qui permettent de mesurer le degré de subjectivité des termes, ainsi que leur valence s'il s'agit d'adjectifs, et qui sont effectués à l'aide du moteur de recherche Yahoo!.

3 Méthodes évaluées pour estimer la valence de mots

Cinq méthodes pour estimer automatiquement la valence de mots ont été comparées, deux de celles-ci consistant en une transposition de méthodes efficaces pour la langue anglaise : celle de Turney et Littman (2002, 2003) et celle de Kamps et Marx (2002 ; Kamps et al., 2004). Nous avons également repris la méthode de Bestgen (2002, 2008). Ces trois méthodes serviront de référence pour évaluer deux nouvelles approches : une extension de la méthode de Kamps et Marx et une méthode d'apprentissage supervisé de mots germes. La présente section décrit les principes à la base de ces différentes méthodes. Des précisions à propos de leur implémentation et des ressources linguistiques qu'elles requièrent sont données dans la section suivante.

3.1 Niveaux de base : SO-ASL et DIC-ASL

Ces deux méthodes se basent sur l'analyse sémantique latente d'une collection de textes pour déterminer la proximité entre des mots et des germes dont la valence est connue.

- SO-ASL : il s'agit de la méthode proposée par Turney et Littman (2003) décrite ci-dessus. Elle est basée sur 14 mots germes choisis en raison de leur valence extrême sur la dimension positif-négatif. La valence d'un mot correspond à la somme des cosinus entre ce mot et les germes positifs dont on soustrait la somme des cosinus entre ce mot et les germes négatifs.

- DIC-ASL : il s'agit de la méthode proposée par Bestgen (2002) décrite ci-dessus. Pour chaque mot dont on veut déterminer la valence, on identifie les 30 plus proches voisins dont la valence est connue et on lui affecte la valence moyenne de ceux-ci.

3.2 Estimation sur la base de relations de synonymie : KA1 et KA7

Ces deux méthodes sont basées sur la fonction d'évaluation définie par Kamps et Marx (2002).

- KA1 : cette méthode est basée sur les liens synonymiques entre les adjectifs. Le principe consiste à mesurer la distance minimale, c'est-à-dire le plus court chemin, entre le mot auquel on veut attribuer une valeur et les mots germes *good* et *bad*. La valence d'un terme *t* est alors égale à sa distance relative avec les deux germes :

$$KA1(t) = \frac{d(t, mauvais) - d(t, bon)}{d(bon, mauvais)}$$

où *d (i, j)* représente la distance du plus court chemin synonymique entre les mots i et j.

- KA7 est une adaptation de KA1 dans laquelle le nombre de paires d'adjectifs de référence est multiplié par 7. Nous avons repris les 7 paires de référence de Turney et Littman (2003), que nous avons traduites comme suit : *bon, gentil, excellent, positif, heureux, correct, supérieur* et *mauvais, méchant, médiocre, négatif, malheureux, faux, inférieur*. La fonction d'évaluation adaptée reprend alors la somme des évaluations pour chaque paire :

$$KA7(t) = \frac{\sum_{k=1}^{n} d(t, i_k) - \sum_{k=1}^{n} d(t, j_k)}{\prod_{k=1}^{n} d(i_k, j_k)}$$

où i_k et j_k forment une paire d'adjectifs positif et négatif des n paires prises en compte.

3.3 Apprentissage supervisé de mots germes : ASG

Un des objectifs de cette recherche est de proposer et d'évaluer une méthode dérivée de celles de Turney et Littman (2003) et de Bestgen (2002) dans laquelle les mots germes originaux, sélectionnés arbitrairement, sont remplacés par des germes optimaux obtenus par une procédure d'apprentissage supervisée basée sur la régression. Pour ce faire, nous employons comme matériel d'apprentissage une norme lexicale pour la dimension évaluative obtenue en demandant à des juges d'évaluer un grand nombre de mots sur cette dimension. À la suite de Heise (1965), une série de normes de ce type ont été développées, principalement en psycholinguistique (Syssau, Font, 2005). La méthode proposée est composée des quatre étapes suivantes :

1. Sélectionner comme germes potentiels les mots qui sont les plus extrêmes sur la dimension positif-négatif selon une norme évaluative comme celle employée dans DIC-ASL.

2. Sur la base d'un espace sémantique obtenu par l'ASL d'une collection de textes, calculer le cosinus entre chacun de ces germes potentiels et tous les mots qui se trouvent dans la norme.

3. Utiliser une procédure de régression afin de construire un modèle prédictif basé sur les germes les plus efficaces pour prédire la valence.

4. Employer le modèle construit à l'étape précédente pour estimer la valence de termes présents dans l'espace sémantique, mais non dans la norme initiale.

Le critère de sélection des germes potentiels proposé à la première étape devrait permettre l'identification de mots germes similaires à ceux originellement choisis par Turney et Littman (2003). Toutefois, lorsqu'on considère le fait que le seuil pour sélectionner les mots les plus extrêmes est par définition arbitraire, il devient immédiatement évident que la procédure proposée n'est qu'un cas particulier d'une procédure plus générale dans laquelle les germes potentiels sont composés de l'ensemble des mots présents dans la norme. Et, d'une manière tout aussi évidente, cette première généralisation n'est, elle-même, qu'un cas particulier d'une seconde généralisation, qui emploie comme germes potentiels tous les mots pour lesquels il est possible de calculer un cosinus avec les mots qui se trouvent dans la norme, soit tous les mots présents dans l'espace sémantique, que leur valence soit connue ou non. Étant donné que les candidats germes pour l'approche la plus restrictive forment un sous-ensemble des candidats germes employés dans les approches plus générales, on doit s'attendre à ce que la qualité de la prédiction de la valence des mots du dictionnaire

initial soit d'autant meilleure que l'approche est la plus générale. Par contre, les capacités de généralisation des différents modèles pourraient être équivalentes si ceux basés sur le plus grand nombre de germes potentiels présentent un défaut de surapprentissage.

4 Expériences

4.1 Ressources linguistiques pour l'implémentation des méthodes

Les différentes méthodes proposées ci-dessus nécessitent des ressources linguistiques spécifiques comme un dictionnaire de synonymes ou une collection de textes pour extraire l'espace sémantique. Les ressources que nous avons employées sont décrites dans la présente section.

4.1.1 Dictionnaire de synonymes

L'adaptation de la méthode de Kamps et Marx (2002) au français nécessite une ressource plus ou moins équivalente au WordNet anglais. En raison de la trop faible couverture de WOLF (WordNet Libre du Français) et du WordNet français développé dans le cadre du projet EuroWordNet[1], nous avons employé le dictionnaire de synonymes développé par le laboratoire CRISCO de l'université de Caen (Manquin et al., 2004)[2]. Celui-ci a été constitué à partir de sept dictionnaires français et comprend plus de 49 000 entrées et 396 000 relations synonymiques. De manière similaire à Kamps et Marx (2002), nous avons récupéré récursivement tous les mots liés à la paire d'adjectifs *bon* et *mauvais*, avec des restrictions sur la catégorie grammaticale pour éviter de générer trop de bruit. Une petite adaptation a dû être faite pour rendre la liste des synonymes récupérés symétrique (Kamps et al., 2004 : 1115).

4.1.2 Norme de valence : Nev

La norme de valence employée pour les méthodes DIC-ASL et ASG est composée de 3252 mots évalués sur une échelle à 7 points allant de *très désagréable* (1) à *très agréable* (7) par un minimum de 30 juges (Hogenraad et al., 1995). À titre d'exemple, la liste suivante donne les valeurs attribuées à quelques mots extraits aléatoirement de ce dictionnaire : détresse = 1.4, impassible = 2.6, ambigu = 3.2, outil = 4.3, revenir = 5.0, admiratif = 5.7, doux = 6.0.

4.1.3 Constitution de l'espace sémantique

L'espace sémantique, utilisé pour calculer les cosinus entre les mots nécessaires pour SO-ASL, DIC-ASL et ASG, a été construit sur la base d'une collection de textes littéraires composée de romans, nouvelles et contes disponibles sur le Web (principalement dans les bases littéraires ABU et Frantext). Elle contient approximativement 5 300 000 mots. Chaque texte a été subdivisé en segments de 125 mots. Pour construire le tableau lexical, les prétraitements suivants ont été effectués : lemmatisation par le logiciel TreeTagger (Schmid, 1994), suppression de mots outils et suppression des mots de fréquence totale inférieure à 10. La matrice de cooccurrences des 12 285 termes dans les 40 635 segments a été soumise à une décomposition en valeurs singulières et les 300 premiers vecteurs propres ont été conservés.

4.2 Méthode pour l'évaluation

Pour évaluer l'efficacité de méthodes visant à déterminer automatiquement la valence de mots, le test classique, lorsque l'étude est réalisée en anglais, se base sur les listes de mots positifs et négatifs incluses dans le *General Inquirer* (p.e., Dragut et al., 2010 ; Kamps et al., 2004 ; Turney, Littman, 2003). Ces listes n'étant pas, à notre connaissance, disponibles en français, nous avons recherché un matériel équivalent dans

[1] Le WOLF couvre 30 % du WordNet de Princeton (Mouton & Chalendar, 2010) et, selon nos calculs, le WordNet français couvre environ 25 % des synsets de la version 1.5 de WordNet.

[2] www.crisco.unicaen.fr/cgi-bin/cherches.cgi

cette langue. La section 4.2.1 décrit les normes de valence de Syssau et Font (2005). Ces normes présentent l'avantage d'avoir été récoltées dans des conditions rigoureuses et bien documentées, alors qu'on ne dispose de pratiquement aucune information sur la procédure suivie pour constituer les deux listes du *General Inquirer*. Cependant, elles ne portent que sur 735 mots alors que les listes originales du *General Inquirer* en contiennent plusieurs milliers. À titre comparatif, nous avons réalisé une première adaptation française des listes du *General Inquirer*.

4.2.1 Valemo : V80, V50 et Vscore

Syssau et Font (2005) ont demandé à 600 juges d'évaluer 735 mots[3] sur deux échelles : une échelle nominale à trois modalités (négatif, neutre et positif) et une échelle bipolaire en 11 points allant de très négatif (-5) à très positif (+5) (voir Syssau et Font pour une discussion des avantages et inconvénients de ces deux types d'évaluation). Chaque mot a été évalué par 100 juges et un même juge n'a effectué qu'un seul des deux types d'évaluation. Les mots ont été sélectionnés sur la base de deux normes d'associations verbales de manière à constituer "un ensemble de mots suffisamment diversifié pour être représentatif de la langue française" (Syssau, Font, 2005). De la première évaluation, Syssau et Font ont dérivé deux normes catégorielles : les mots "indubitablement" positifs ou négatifs qui ont été classés dans la catégorie correspondante par au moins 80% des juges (V80) et les mots "majoritairement" positifs ou négatifs qui ont été classés ainsi par au moins 50% des juges (V50). La seconde évaluation a produit une norme valencée (Vscore) avec pour chaque entrée un score compris entre -5 et +5.

4.2.2 General Inquirer (version francisée) : GI

Le *General Inquirer* est un projet né en 1961 qui visait à développer un programme d'analyse objective de contenu (Stone et al., 1966) basé sur un dictionnaire composé de 182 catégories sémantiques. Les deux dernières catégories ajoutées sont les catégories positive et négative, qui répertorient respectivement 1915 et 2291 mots. Ces listes n'étant pas, à notre connaissance, disponibles en français, nous les avons traduites automatiquement à l'aide du traducteur en ligne Systran. Après avoir été lemmatisées avec TreeTagger, ces deux listes ont été contrôlées par deux juges. Après suppression des doublons et des mots présents dans les deux listes – problèmes présents dans la version originale, mais également dus à la traduction – , nous avons obtenu 1246 mots positifs et 1527 mots négatifs.

5 Résultats

Cinq normes ont été employées pour comparer l'efficacité des méthodes de construction automatique de lexiques dans l'estimation de la valence de mots : la norme Nev, les trois normes issues du projet Valemo (Vscore, V50 et V80) et notre traduction des listes positive et négative du *General Inquirer* (GI). Pour les deux normes qui définissent la valence comme une variable continue (Nev et Vscore), nous avons évalué la qualité de la prédiction en calculant le coefficient de corrélation de Pearson entre les valences prédites par les méthodes automatiques et les valeurs moyennes attribuées par les juges. Lorsque la variable à prédire est dichotomique (positif versus négatif : V50, V80 et GI), nous avons employé comme mesure d'efficacité le pourcentage de mots classés par les procédures automatiques dans la catégorie déterminée par la norme. Pour chacune des méthodes évaluées, un mot est considéré comme négatif lorsque sa valence prédite est inférieure à la moyenne et comme positif dans le cas contraire[4].

La principale difficulté que nous avons rencontrée lors de ces analyses trouve son origine dans le fait que les différentes méthodes testées ne donnent pas des valeurs de valence aux mêmes mots : celles dérivées de Kamps et Marx (2002) en proposent un nombre nettement plus restreint que celles qui s'appuient sur l'ASL. Ceci nous a conduits à présenter séparément les résultats de ces deux groupes de méthodes.

---

[3] La norme initiale portait sur 605 mots, mais elle a été ultérieurement étendue à 735 mots. Elle est disponible à l'adresse : http://www.lexique.org/

[4] Des analyses complémentaires ont montré que ce seuil était proche de la valeur optimale obtenue par régression logistique.

5.1 Approche basée sur le dictionnaire de synonymes : KA1 et KA7

Le tableau 1 présente les performances des méthodes KA1 et KA7 pour les différentes normes. Pour tous les tests, KA7, la version basée sur les 7 paires de mots germes de Turney et Litman (2003), est supérieure à KA1 qui n'emploie qu'une seule de ces paires, celle sélectionnée par Kamps et Marx (2004). Les corrélations entre la valence prédite par les méthodes et la valence moyenne selon les juges sont élevées et même très élevées pour Vscore. Pour la prédiction de la catégorie des mots, les performances sont également impressionnantes pour les trois tests. Dans leur étude sur l'anglais, Kamps et al. (2004) rapportent un pourcentage de mots bien classés par leur procédure de 67 % pour les 667 adjectifs pour lesquels ils ont pu calculer un score d'évaluation à partir de WordNet et qui se trouvent dans la liste du *General Inquirer* (Evaluation II, Table 1 dans Kamps et al., 2004). Cette valeur est nettement inférieure à celle que nous avons obtenue. S'il est difficile d'identifier précisément l'origine de l'amélioration, force est de constater que l'implémentation de la technique de Kamps et Marx sur la base d'un dictionnaire de synonymes plutôt que de WordNet est une alternative viable.

	Nev	Vscore	V80	V50	GI
N	663	76	20	43	688
KA1	0.55	0.64	90%	84%	80%
KA7	0.61	0.72	100%	88%	84%

Tableau 1 : Performances (corrélation et pourcentage de classification correcte)

5.2 Approches basées sur l'ASL

Dans cette section, nous comparons la nouvelle méthode ASG à celles de Turney et Littman (2003) et de Bestgen (2002). Quatre versions différentes de ASG ont été testées. Elles se distinguent par l'étendue des germes potentiels pris en compte : ASG0.5 limite ceux-ci aux valeurs les plus extrêmes de la norme (de 1 à 1.5 et de 6.5 à 7), ASG1.0 est moins stricte et prend en compte celles comprises entre 1 et 2 et entre 6 et 7, ASGnorme prend en compte l'ensemble des mots repris dans la norme Nev et ASGtout sélectionne les germes parmi l'ensemble des termes présents dans l'espace sémantique. Pour construire le modèle prédictif sur la base de ces ensembles de germes potentiels, nous avons employé une régression linéaire multiple[5] avec sélection des prédicteurs par la technique ascendante (*forward*) et un seuil de probabilité pour la sélection fixé à 0.01.

5.2.1 Performances pour le matériel d'apprentissage : Nev

La première ligne du tableau 2 présente les corrélations entre les valeurs données dans la norme Nev, qui a servi pour l'apprentissage, et les valeurs prédites par les différentes méthodes. Comme on pouvait s'y attendre, SO-ASL, la seule des méthodes qui ne s'appuie pas sur la norme, obtient le moins bon résultat. Tout aussi attendus sont les bénéfices apportés par l'apprentissage supervisé (ASG versus DIC-ASL) et par la possibilité de choisir les germes parmi un nombre plus important de candidats. On note néanmoins que la différence principale se situe entre ASG0.5 et ASG1.0.

5.2.2 Performances pour Vscore

L'analyse de Vscore, deuxième ligne du tableau 2, donne comme attendu, des valeurs inférieures à celles obtenues pour la norme ayant servi à l'apprentissage, mais la différence est assez faible. On note tout particulièrement que les méthodes ASG sont nettement plus performantes que SO-ASL, ce qui confirme l'hypothèse que les mots germes employés par cette dernière sont loin d'être optimaux.

[5] Toutes les analyses ont également été effectuées en employant la SVR (SVM appliqué à la régression), mais ils ne sont pas présentés, car les deux techniques ont produit des résultats très similaires.

Normes	N	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
Nev	2685	0.38	0.60	0.60	0.65	0.66	0.67
Vscore	631	0.32	0.60	0.56	0.61	0.61	0.60

Tableau 2 : Corrélation entre les valeurs prédites par les méthodes et les normes

5.2.3 Performances pour les catégories : V80, V50 et GI

Le tableau 3 présente le pourcentage de mots bien classés pour les différentes normes catégorielles. Les performances pour V80 et V50 sont très élevées, mais il faut prendre en compte le fait que ces deux normes ne contiennent qu'un nombre réduit de mots. Pour l'adaptation française du *General Inquirer*, les performances sont moins bonnes. Elles dépassent toutefois largement la performance de SO-ASL rapportée par Turney et Littman (2003) pour le *General Inquirer* en version anglaise (65%), valeur très proche de celle que nous avons obtenue pour l'adaptation française (64%). On observe aussi que DIC-ASL fait presque aussi bien que les méthodes basées sur une procédure d'apprentissage automatique.

Test	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
V80 (N=128)	73%	88%	83%	87%	88%	91%
V50 (N=280)	63%	82%	78%	82%	84%	82%
GI (N=1992)	64%	71%	70%	72%	73%	72%

Tableau 3 : Pourcentage de classification correcte

Dans le tableau 3, tous les mots mentionnés dans les normes sont pris en compte, même ceux qui sont présents dans la norme Nev qui a servi à l'apprentissage supervisé. Il s'ensuit qu'il est problématique de se baser sur ces données pour évaluer les capacités de généralisation de la méthode ASG à des mots qui ne font pas partie du matériel d'apprentissage. Pour cette raison, les mêmes analyses que celles rapportées ci-dessus ont été effectuées après suppression dans les normes catégorielles de tous les mots présents dans Nev. Les résultats sont présentés dans le tableau 4. Pour GI, on observe une diminution assez faible et relativement égale des performances pour toutes les méthodes, y compris celles qui n'ont pas recours à l'apprentissage supervisé. Pour V50 et surtout V80, les différences sont plus nettes et s'observent même pour SO-ASL, alors que cette méthode ne s'appuie pas sur la norme Nev. L'explication la plus probable est que les mots qui ont été supprimés sont particulièrement faciles à classer par toutes les méthodes.

Test	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
V80 (N=25)	60%	80%	68%	72%	72%	76%
V50 (N=82)	60%	82%	72%	73%	78%	74%
GI (N=1130)	62%	68%	68%	71%	71%	70%

Tableau 4 : Pourcentage de classification correcte pour les mots non inclus dans Nev

D'une manière générale, ces tests confirment le caractère non optimal des mots germes employés dans l'approche SO-ASL, cette méthode atteignant un niveau de performance nettement inférieur à celui atteint par toutes celles basées sur l'apprentissage supervisé de germes.

5.3 Comparaison globale

Une dernière série d'analyses visent à comparer le plus rigoureusement possible les performances de toutes les procédures testées, y compris KA1 et KA7, sur une même tâche afin de les rendre comparables. On a donc calculé le pourcentage de termes bien classés pour les mots de GI traités par toutes les méthodes. Le tableau 5, qui présente ces résultats, souligne la supériorité de KA7 sur toutes les autres méthodes. Il faut toutefois garder à l'esprit que KA7 propose au maximum des valeurs pour 688 mots du GI alors que les méthodes basées sur l'ASL traitent 1992 mots de cette même liste. De plus, nous n'avons employé qu'un seul espace sémantique d'un genre très spécifique (voir discussion). Les mêmes analyses ont été réalisées en supprimant, en plus, les mots qui sont dans la norme NEV, sans que les conclusions ne soient modifiées (différences plus petites ou égales à 2%).

N	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASGtout	KA1	KA7
550	64%	70%	75%	75%	76%	75%	80%	83%

Tableau 5 : Pourcentage de classification correcte pour les mots de GI traités par toutes les méthodes

5.4 Mots germes les plus importants pour prédire la valence

Si la méthode ASG n'apparait pas comme nettement supérieure à DIC-ASL, elle présente un avantage potentiellement très important en termes d'identification de mots germes. Alors que DIC-ASL sélectionne les germes localement puisqu'un ensemble différent de germes est employé pour chaque mot, ASG sélectionne les germes globalement : un seul et même ensemble de germes est employé pour prédire la valence de tous les mots. Il reste cependant à montrer que les germes choisis par ASG sont bien pertinents.

Une première manière de répondre à cette question consiste à s'intéresser au modèle prédictif construit par la régression multiple. Faute de place, il n'est pas possible de reprendre ici tous les mots germes sélectionnés par les différentes versions de ASG. La liste suivante présente l'ensemble des germes sélectionnés par ASG1.0, suivant l'ordre dans lequel ils ont été introduits dans le modèle (chaque fois suivi par la valence selon la norme Nev) : *épouvantable* (1.8), *délicieux* (6.2), *irriter* (1.9), *admiration* (6.1), *affectueux* (6.2), *atroce* (1.5), *heureux* (6.5), *monstrueux* (1.4), *magnifique* (6.5), *embrasser* (6.4), *lugubre* (1.8), *rêver* (6.3), *libre* (6.3), *savourer* (6.0), *ennui* (1.7), *intéressant* (6.0), *indifférence* (2.0), *espoir* (6.1), *pire* (1.4), *fidèlement* (6.1), *gaieté* (6.4), *rat* (1.9), *insulte* (1.6), *maladie* (1.5), *laideur* (1.6), *enlacer* (6.4), *enfant* (6.3), *crasse* (1.8), *voyage* (6.2), *malchance* (1.6), *admirable* (6.1).

L'analyse qui précède repose sur le modèle prédictif construit par la régression multiple. Celui-ci correspond à la meilleure combinaison possible de mots germes pour prédire la norme et non aux mots germes qui apportent individuellement la contribution la plus importante à la prédiction de celle-ci. Tout particulièrement, la régression multiple ne sélectionnera qu'un seul de deux mots sémantiquement très liés, même si tous les deux sont d'excellents prédicteurs (cf. *rage* et *colère* dans le tableau 6). Or, comme notre objectif prioritaire est d'identifier des mots germes spécifiques qui pourraient être ensuite employés dans d'autres méthodes, comme celle de Kamps et Marx (2002), il semble préférable de s'intéresser à ces derniers et donc à ceux dont le vecteur de cosinus (avec les mots présents dans la norme) est le plus corrélé avec la valence de ces mots. Le tableau 6 présente, à titre d'exemple, une petite fraction des germes les plus importants pour prédire la valence, classés par ordre d'efficacité, lorsqu'on prend en compte l'ensemble des mots présents dans l'espace sémantique. La partie gauche reprend les 30 germes les plus corrélés négativement avec la valence et la partie droite les 30 germes les plus corrélés positivement. La quasi-totalité des germes négatifs mentionnés dans ce tableau correspond à ce qu'on entend habituellement par mots germes pour la valence[6]. La grande majorité des germes positifs sont aussi pertinents et plus de la moitié d'entre eux ne se trouve pas dans la norme ayant servi à l'apprentissage (signalé par un "-" à la place du score de valence). Cette observation souligne la valeur heuristique de la méthode proposée. On y trouve néanmoins quelques mots spécifiques à la collection de textes employée pour l'ASL (*mythologique*, *nymphe*, *pampre*). Il est à noter que les germes qui suivent, par ordre d'importance, ceux présentés dans le

[6] Il n'est pas possible, à ce stade de l'analyse, de déterminer le nombre de cas dans lesquels *débattre* correspond à *se débattre*. Il s'agit là d'une limite évidente des prétraitements effectués avant l'extraction de l'espace sémantique

tableau semblent tout aussi pertinents. À titre d'exemple, on trouve de 10 en 10 pour l'orientation négative : 31. *brute*, 41. *monstrueux*, 51. *exécration*, 61. *exaspération*, 71. *désespérer*, 81. *sourd*, 91. *égorgement*, 101. *râle*.

	Négatif	Nev		Négatif	Nev		Positif	Nev		Positif	Nev
1	rage	2.1	16	infamie	-	1	charmant	5.7	16	charme	6.1
2	colère	2.2	17	imprécation	-	2	charmer	5.8	17	description	-
3	épouvantable	1.8	18	tourmenteur	-	3	ravissant	6.4	18	modeste	-
4	fureur	2.8	19	lâche	1.1	4	délicieux	6.2	19	ravir	5.6
5	atroce	1.5	20	menaçant	-	5	gracieux	5.9	20	admirable	6.1
6	horrible	1.8	21	menace	-	6	merveille	6.1	21	romance	4.4
7	abominable	1.9	22	épouvanter	2.0	7	magnifique	6.5	22	nymphe	-
8	écraser	1.9	23	saigner	-	8	brillant	-	23	exquis	-
9	horreur	2.1	24	cracher	-	9	aimable	5.9	24	distingué	-
10	crachat	1.3	25	débattre	-	10	harmoniser	-	25	pampre	-
11	exaspérer	2.4	26	effrayant	2.4	11	élégant	-	26	enchanter	-
12	misérable	1.8	27	plainte	-	12	riant	-	27	exotique	-
13	étrangler	-	28	crever	1.7	13	splendide	-	28	raffoler	-
14	affreux	1.9	29	meurtre	1.4	14	mythologique	-	29	modestement	-
15	assassin	-	30	injurier	1.9	15	composer	-	30	fraîcheur	-

Tableau 6 : Mots germes sélectionnés par la méthode ASG

6 Conclusion

Pour conclure, nous avons transposé au français deux méthodes de construction automatique de lexiques porteurs de valences bien établies dans le monde anglo-saxon : celles de Turney et Littman (2003) et de Kamps et Marx (2002). Cette dernière montrant des résultats encourageants, nous l'avons étendue en augmentant le nombre de paires de mots germes. Cette modification nous a permis d'obtenir les meilleurs résultats, avec plus de 80 % de termes bien classés. Ce pourcentage doit cependant être relativisé dans la mesure où il est calculé sur un nombre restreint de mots. Nous avons également développé une méthode qui sélectionne les mots germes par apprentissage supervisé. Avec une efficacité d'environ 75 %, elle surpasse nettement la méthode SO-ASL dont elle est dérivée. Il est, hélas, impossible de déterminer si les valeurs obtenues reflètent un niveau de performance proche de celui atteint par des annotateurs parce qu'on ne dispose pas d'information à propos du degré d'accord entre ceux-ci. L'analyse des mots apportant la plus grande contribution individuelle à la prédiction de la valence souligne l'intérêt de cette méthode pour l'identification de mots germes. Un des principaux développements envisagés est d'utiliser ces mots germes dans des méthodes comme celles de Kamps et Marx (2002) ou d'Esuli et Sebastiani (2006). Des adaptations seront nécessaires puisque, dans la version actuelle, les mots germes identifiés ne forment pas des couples comme requis par la méthode de Kamps et Marx. Il sera tout particulièrement intéressant de déterminer si la méthode proposée, qui ne requiert pas WordNet, est plus efficace que celle développée par Esuli et Sebastiani et, surtout, si l'emploi dans leur méthode des mots germes identifiés par ASG améliore encore les performances. Enfin, il sera nécessaire d'évaluer les bénéfices apportés par l'apprentissage supervisé de germes pour l'objectif principal de ce genre d'études : déterminer l'orientation de textes (Harb et al., 2008).

Cette étude comporte plusieurs limitations qui sont autant de pistes pour des recherches futures. Tout d'abord, un seul espace sémantique, extrait de textes littéraires, a été exploité. Les implications de cette limitation sont particulièrement mises en évidence par la sélection de mots germes spécifiques à ce genre de textes. Il serait intéressant d'effectuer ces analyses sur un corpus plus diversifié ou, séparément, sur des corpus de genres différents. Dans ce dernier cas, il devrait être possible d'attribuer aux mots germes un indice qui traduit leur degré de généralité. Ensuite, les germes identifiés par la méthode ASG consistent en des formes (lemmes) isolées, ce qui réduit fortement la qualité linguistique de l'analyse (voir *débattre*). La prise en compte de mots composés ou d'expressions figées serait également un développement intéressant (Vernier, Monceaux, 2010). D'autres méthodes pour mesurer les proximités sémantiques devraient également être testées. Il est en effet loin d'être évident que le passage par l'ASL améliore l'efficacité (Bestgen, 2006). Enfin, notre traduction des listes du *General Inquirer* pourrait sans aucun doute être améliorée afin de récupérer un certain nombre de mots perdus. Cependant, on peut s'interroger sur l'utilité d'un tel travail, étant donné le peu d'information disponible sur la procédure de construction de ces listes. Il

nous semble plus intéressant pour la communauté scientifique d'étendre les normes V50 et V80, dont la rigueur et les procédés de construction sont bien établis.

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS. Les auteurs remercient vivement A. Syssau pour les explications complémentaires à propos de la norme *Valemo* et l'équipe du CRISCO pour l'autorisation d'extraction des informations incluses dans le dictionnaire de synonymes.

Références

ABBSASI, A., CHEN, H., SALEM, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems* 26.

BESTGEN, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. Actes de *CIFT*'02, 81-94.

BESTGEN, Y. (2006). Déterminer automatiquement la valence affective de phrases : Amélioration de l'approche lexicale. Actes des *JADT 2006*, 179-188.

BESTGEN, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. Proceedings of *LREC 2008*, 496-500.

CHARDON, B. (2010). Catégorisation automatique d'adjectifs d'opinion à partir d'une ressource linguistique générique, Actes de *RECITAL 2010*.

CHESLEY, P., VINCENT, B., XU, L., SRIHARI, R.K. (2006). Using verbs and adjectives to automatically classify blog sentiment. Proceedings of *AAAI-CAAW-06*, 27-29.

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K., HARSHMAN, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science* 41, 391-407.

DRAGUT, E.C., YU, C., SISTLA, P., MENG, W. (2010). Construction of a sentimental word dictionary. Proceedings of *ACM ICIKM*, 1761-1764**.**

ESULI, A., SEBASTIANI, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. Proceedings of *LREC'06*, 417–422,.

HARB, A., PLANTIE, M., ROCHE, M., DRAY, G., TROUSSET, F., PONCELET, P. (2008). Détection d'opinion. Comment déterminer les adjectifs d'opinion d'un domaine donné? *Document numérique* 11, 37-61.

HATZIVASSILOGLOU, V., MCKEOWN, K.R. (1997). Predicting the semantic orientation of adjectives. Proceedings of *EACL 1997*, 174-181.

HEISE, D.R. (1965). Semantic differential profiles for 1000 most frequent english words. *Psychological Monographs* 79, 1-31.

HOGENRAAD, R., BESTGEN, Y., NYSTEN, J.L. (1995). Terrorist Rhetoric : Texture and Architecture, In Nissan et Schmidt (Eds.), *From Information to Knowledge*,48-59, Intellect

HU, M., LIU, B. (2004). Mining Opinion Features in Customer Reviews. Proceedings of *AAAI*, 755-760.

KAMPS, J., MARX, M. (2002). Words with Attitude. Proceedings of *the 1st Interational Conference on Global WordNet,* 332-341.

KAMPS, J., MARX, M., MOKKEN, R.J., DE RIJKE, M. (2004). Using WordNet To Measure Semantic Orientations Of Adjectives. Proceedings of *LREC 2004*, 1115-1118.

KIM, S.M., HOVY, E. (2004). Determining the sentiment of opinions. Proceedings of *COLING*, 1367-1373.

MANQUIN, J.L., FRANÇOIS, J., EUFE, R., FESENMEIER, L., OZOUF, C., SENECHAL, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *Les Cahiers du CRISCO* 17, 1-64.

MILLER, G.A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography* 3, 235–312.

MOUTON, C., CHALENDAR, G. (2010). JAWS : Just AnotherWordNet Subset. Actes de *TALN 2010*.

NASUKAWA, T., YI, J. (2003). Sentiment analysis: capturing favorability using natural language processing. Proceedings of *the 2nd international conference on Knowledge capture (K-CAP)*, 70-77.

PAK, A., PAROUBEK, P. (2010). Construction d'un lexique affectif pour le français à partir de Twitter. *Actes de TALN 2010*.

PANG, B., LEE, L., VAITHYANATHAN, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 79-86.

SCHMID, H., (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of *the International Conference on New Methods in Language Processing*, 44-49.

STONE, P.J., DUNPHY, D.C., SMITH, M.S., OGILVIE, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge : MIT Press.

SYSSAU, A., FONT, N. (2005). Evaluations des caractéristiques émotionnelles d'un corpus de 604 mots. *Bulletin de Psychologie* 58, 361-367.

TURNEY, P.D., (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of *the 40th Annual ACL Meeting*, 417-424.

TURNEY, P.D., LITTMAN, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Technical Report*, National Research Council Canada.

TURNEY, P.D., LITTMAN, M. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems* 21, pp. 315--346

VELIKOVICH, L., BLAIR-GOLDENSOHN, S., HANNAN, K., MCDONALD, R. (2010). The Viability of Web-derived Polarity Lexicons. Proceedings of *NAACL 2010*, 777-785.

VERNIER, M., MONCEAUX, L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement automatique des langues* 51, 125-149.

VERNIER, M., MONCEAUX, L., DAILLE, B., DUBREIL, E. (2009). Catégorisation sémantico-discursives des évaluations exprimées dans la blogosphère. Actes de *TALN 2009*.

WIEBE, J., WILSON, T., BRUCE, R., BELL, M., MARTIN, M. (2004). Learning subjective language. *Computational Linguistics* 30, 277-308.

WIEBE, J., WILSON, T., CARDIE, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39, 165-210.

WILSON, T., WIEBE, J., HOFFMANN, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of *HLT-EMNLP 2005*, 347-354.

YU, H., HATZIVASSILOGLOU, V. (2003). Toward answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of *EMNLP 2003*, 129-136.