Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions

Núria Gala[1] *Caroline Brun*[2]
(1) LIF-CNRS UMR 7279, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France
(2) Xerox Research Centre Europe, 6 chemin de Maupertuis 38240 Meylan, France
nuria.gala@lif.univ-mrs.fr, caroline.brun@xrce.xerox.com

RESUME

Les ressources lexicales sont cruciales pour de nombreuses applications de traitement automatique de la langue (par exemple, l'extraction d'opinions à partir de corpus). Cependant, leur construction pose des problèmes à différents niveaux (coût, couverture, etc.). Dans cet article, nous avons voulu vérifier si les informations morphologiques liées à la dérivation pouvaient être exploitées pour l'annotation automatique d'informations sémantiques. En partant d'une ressource regroupant les mots en familles morphologiques en français, nous avons construit un lexique de polarités pour 4 065 mots, à partir d'une liste initiale d'adjectifs annotés manuellement. Les résultats obtenus montrent que la propagation des polarités est correcte pour 78,89% des familles avec un seul adjectif. Le lexique ainsi obtenu améliore aussi les résultats du système d'extraction d'opinions.

ABSTRACT

Spreading Polarities among Word Families: Impact of Morphology on Building a Lexicon for Sentiment Analysis

Lexical resources are essential for many natural language applications (for example, opinion mining from corpora). However, building them entails different problems (cost, coverage, etc.). In this paper, we wanted to verify whether morphological information about derivation could be used to automatically annotate semantic information. Starting from a resource that groups words into morphological families in French, we have built a lexicon with polarities for 4 065 words from an initial seed set of manual annotated adjectives. The results obtained show that spreading polarities is accurate for 78.89% of the families with a unique adjective. The lexicon obtained also improves the results of the opinion mining system on different corpora.

MOTS-CLES : ressources lexicales, morphologie dérivationnelle, analyse de sentiments

KEYWORDS : lexical resources, derivational morphology, opinion mining

1 Introduction

Depuis quelques années, l'analyse de sentiments suscite de l'intérêt dans la communauté du traitement automatique des langues (TAL), comme conséquence d'un réel besoin dans le traitement de grandes masses de données : services web pour le tourisme ou la culture, discours politiques, etc. Par analyse de sentiments, on entend la détection de la polarité d'un texte, c'est-à-dire, l'obtention automatique de la tendance ou de l'opinion qui s'en dégage.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 495–502,
Grenoble, 4 au 8 juin 2012.

Deux approches ressortent dans la littérature. Les approches statistiques supervisées, fondées sur les co-occurrences de mots dans des corpus, et les approches plus linguistiques qui s'appuient, elles, sur des ressources lexicales. L'idée des méthodes statistiques est de calculer, à partir d'un ensemble de co-occurrences annotées, d'autres co-occurences polarisées (Hatzivassiloglous et McKeown 1997 ; Turney 2002). La procédure de classification se fait automatiquement à partir d'exemples, le modèle attribue une polarité en fonction d'un processus inductif (Pang et al. 2002). L'autre type de méthode est considéré plus linguistique, dans la mesure où l'on utilise des ressources comme des thésaurus, des réseaux lexicaux, etc. (Kim et Hovy 2004 ; Esuli et Sebastiani 2005), ce qui permet d'améliorer les performances des systèmes d'analyse (Choi et Cardie 2009 ; Lu et al. 2011). Les méthodes qui utilisent des lexiques présupposent deux hypothèses : toute unité lexicale aurait une orientation sémantique intrinsèque, indépendamment de son contexte d'apparition ; cette orientation peut être exprimée avec une valeur numérique (Taboada et al. 2011).

Notre travail se situe dans cette perspective : nous nous sommes proposées de construire un lexique de polarités. A partir d'une liste initiale de 3 882 adjectifs annotés manuellement par trois annotateurs, nous avons voulu observer l'impact de la morphologie dérivationnelle dans le maintien ou non de la polarité. C'est-à-dire, nous avons voulu tester l'hypothèse selon laquelle la polarité intrinsèque d'un adjectif est la même que celle des unités lexicales de sa famille morphologique. L'idée a été de voir (i) si on pouvait construire une ressource qui capitalise sur les liens morphologiques pour propager des informations sémantiques, et (ii) si une telle ressource améliore les résultats d'un système d'analyse d'opinions.

Si l'estimation de la polarité d'un texte passe par des phénomènes contextuels (intensificateurs, négation, etc.) et syntaxiques (Brun 2011), la qualité du lexique à la base du système reste cruciale. La construction d'un tel lexique demeure donc un aspect important. Des lexiques existants pour l'anglais ont été construits à partir de WORDNET, par exemple, WORDNET-AFFECT (Strapparava and Valitutti 2004), SENTIWORDNET (Esuli and Sebastiani 2006). Pour le français, on peut citer les travaux de Vernier et Monceaux (2010) pour obtenir automatiquement une liste de 982 termes subjectifs à partir de l'indexation de documents sur le Web ou l'application LIKEIT de JEUXDEMOTS (Lafourcade 2007) où l'enrichissement de la liste de mots polarisés se fait de façon contributive. En dehors de ces exemples, à notre connaissance, il n'existe pas de lexique de polarités pour le français.

Dans cet article, nous décrirons la méthodologie de construction de notre lexique (section 2) et dans la section 3, nous évaluerons la qualité des données obtenues au regard des familles morphologiques (propagation des polarités). Nous présenterons, enfin, les résultats d'un système d'analyse d'opinions qui intègre la ressource.

2 Construction de la ressource lexicale

Pour constituer le lexique de polarités, nous avons utilisé la deuxième version de POLYMOTS, une ressource lexicale regroupant 19 009 mots, à ce jour, en 2 069 familles morpho-phonologiques (Gala et Rey 2008). La deuxième version de cette ressource, outre une description plus fine de quelques familles de mots en clusters sémantiques

(Gala et al. 2011), contient des étiquettes grammaticales, ce qui nous a permis d'extraire les 3 785 adjectifs et de les annoter manuellement avec trois valeurs (positif, négatif, neutre). Cette liste initiale d'adjectifs a été complétée avec une centaine d'adjectifs supplémentaires provenant d'un lexique de l'analyseur XIP. Nous totalisons 3 882 adjectifs annotés.

2.1 Accord inter-annotateurs

Afin de prendre en compte l'accord inter-annotateurs, nous avons transformé les étiquettes en pourcentages (*100%neg, 33%neutre_66%neg, 66%pos_33%neutre*, etc.). Sans compter l'accord à 100%, nous obtenons 22 étiquettes différentes que nous avons regroupées en accord majoritaire (75%-25% si il y a eu quatre annotations -les trois annotateurs initiaux plus l'annotation provenant du lexique de XIP- ou 66%-33%). Enfin, nous avons considéré comme non significatif les cas où il y a eu un seul annotateur ou bien les cas où il n'y a pas eu de tendance claire (*33%pos_33%neutre_33%neg, 50%pos_25%neutre_25%neg,* etc.). La distribution des étiquettes en termes d'accord inter-annotateurs est la suivante : 1 341 adjectifs avec accord total (34,5%), 969 accord majoritaire (25%) et 1 572 accord non significatif (40,50%)[1].

2.2 Propagation des polarités vers les familles de mots

Pour chacun des 3 882 adjectifs, nous avons étendu automatiquement sa polarité vers les mots de sa famille morphologique. Trois cas de figure se sont présentés : (i) la famille contient un seul adjectif, (ii) la famille en contient plusieurs, (iii) la famille n'en contient aucun.

Dans le cas où plusieurs adjectifs sont présents dans la famille (36,97% des cas, 765 familles au total), la difficulté réside dans le choix du critère d'attribution de la polarité lorsqu'elle est différente. A ce stade, le choix de l'étiquette à propager devait être arbitraire, nous n'avons donc pas utilisé l'ensemble de ces données.

Le 32,19% des familles de POLYMOTS (666 au total) qui ne contiennent pas d'adjectifs, a également échappé au processus d'annotation automatique. Il s'agit de familles pour lesquelles la dérivation est nulle (*agrume, aisselle, falaise*, *oncle*, *taie*, etc.) ou quasi nulle (*cage/cageot, poutre/poutrelle/poutraison, nid/nidation/nidification/nidifier,* etc.). Il s'agit aussi de familles pour lesquelles des adjectifs ont été rajoutés après avoir constitué notre liste d'adjectifs initiale.

A ce jour, le lexique que nous avons créé par propagation des polarités à partir des adjectifs initialement annotés contient, 4 065 mots correspondant aux 638 familles avec un seul adjectif (30,84%). La moyenne de mots par famille est de 6,4 mots. Le lexique est constitué de 662 adjectifs (16,3%), 2 337 noms (57,4%), 878 verbes (21,6%) et 193 adverbes (4,7%), cf. http://polarimots.lif.univ-mrs.fr.

[1] Cette classification donne une pondération supérieure aux adjectifs pour lesquels une polarité se dégage. Ce poids est encodé dans le lexique : poids = 1 si 100% d'accord, poids = 0.5 si 75% ou 66% d'accord, poids = 0 si les valeurs sont distribuées (50%-50% ou 33%-33%-33%).

2.3 Remarques méthodologiques

La liste d'adjectifs initiale a été annotée par trois annotateurs bénévoles différents. Dans le cas de sens multiples, nous avons considéré le sens propre en priorité en nous appuyant sur les définitions du TLFi. Ainsi, par exemple, lancinant a comme définition « qui se fait sentir par élancements aigus ». Lors de l'annotation des mots de la famille de cet adjectif, la polarité négative a été propagée. Cela ne pose pas de problèmes pour *lance* (« arme ») ou *lanciner* (« se faire sentir par des élancements douloureux ») mais en pose pour d'autres termes de la famille comme *lancement, élan,* etc. qui devraient être annotés comme neutres. Dans notre lexique, les mots polysémiques (ex. *lancement* sens « départ ») ou homonymiques (ex. *élan* sens « animal ») n'ont qu'une polarité correspondant à un seul sens. Le traitement des sens figurés et de la polysémie en général reste un problème crucial qui mérite un travail beaucoup plus approfondi (qui sort du cadre de la construction de notre ressource et, de surcroît, de la présentation dans cet article). Par ailleurs, même si nous faisons l'hypothèse qu'un mot possède une polarité intrinsèque, nous sommes conscientes que celle-ci est susceptible de varier en fonction des mots co-occurrents (par exemple, pour *gorgé*, elle peut osciller entre positif et négatif dans, respectivement, « un fruit gorgé de vitamines » par rapport à « un terrain gorgé d'eau »). Dans ces cas, faute de contexte, nous avons attribué une polarité neutre.

Enfin, une fois la ressource annotée, un certain nombre de polarités attribuées automatiquement ont été modifiées en fonction d'affixes porteurs d'altérations sémantiques. C'est le cas d'affixes de négation/opposition (*anti-, contre- dé-/dés-, i-/im-/in-, mal-/mé-*) par exemple dans *antiatomique*, *contrepoison*, *déboisement, inachevé, malaisé*.... Étant donné la variété de polarités et de cas (opposition nette positif/négatif : *poison/contrepoison, salissure/antisalissure*, etc. ; dégradation ou amélioration par rapport à une marque neutre : *obligeant/désobligeant*, *créditer/discréditer*, etc.) nous avons modifié les polarités des mots avec des affixes de négation au cas par cas.

2.4 Évaluation intrinsèque

Une évaluation manuelle a été faite pour 2 954 mots correspondant à 450 familles annotées automatiquement par propagation de la polarité à partir d'un adjectif (environ 70% du lexique). Les résultats de cette évaluation sont les suivants : 355 familles maintiennent la polarité de l'adjectif (78,89%) et 95 ne la maintiennent pas (21,11%).

L'impact de la taille des familles morphologiques est un facteur essentiel dans le maintien d'une polarité. Ainsi, plus la taille de la famille est réduite, plus la polarité reste identique, étant donné une cohésion sémantique plus forte. C'est la cas de familles de moins de huit dérivés, par exemple : *acariâtre/acariâtreté* (100%neg), *vertèbre/vertébral/vertébré* (100%neutre), *allègre/allégrement/alléger/allégresse* (100%pos), *aromal/aromatique/aromatiser/arôme... (66%pos_33%neutre), etc.*

Dans le cas de familles avec un seul adjectif mais de taille plus grande (surtout au delà d'une dizaine de dérivés), la dispersion sémantique est trop importante, ce qui fait varier considérablement les polarités au sein de la famille (par exemple, *sec/dessécher/séchage/séchoir/...*, *bombé/bombe/bombage/bombardement/...*, *nerveux/nerf/nerveusement/nervure/...*, etc.). Comme évoqué plus haut, les problèmes de polysémie ont également un impact très important. Enfin, dans d'autres cas, les suffixes

formateurs d'adjectifs ou participes (*-é, -ant, -al, -eux*) modifient souvent le sens du mot dérivé et font basculer la polarité généralement de neutre à négatif (*âge/âgé, angle/anguleux, bête/bestial, bouillir/ébouillanté*, *larme/larmoyant*, etc.) mais aussi de neutre à positif (*baraque/baraqué, rayon/rayonnant*).

3 Évaluation par le biais d'un système d'extraction d'opinions

Pour évaluer la qualité de la ressource lexicale, nous l'avons intégrée à un système d'extraction d'opinions afin de mesurer l'impact de cette intégration sur la capacité de ce système à classer correctement des revues en ligne selon l'opinion globale de l'utilisateur.

3.1 Constitution des corpus d'évaluation

Nous avons collecté deux corpus de revues en ligne : l'un concernant des livres (*evene.fr/livres)* et l'autre des restaurants (*www.linternaute/restaurant)*. Ces revues sont au format html et semi-structurées. Elles ont été converties au format XML en filtrant les informations pertinentes : titre du livre/nom du restaurant, auteur du commentaire, date, note globale et commentaire libre. Le corpus de revues de livres contient 3 110 revues, le corpus de revues de restaurants contient 99 373 revues.

3.2 Le système d'extraction d'opinions

Notre système d'extraction d'opinions utilise les résultats de l'analyse syntaxique profonde fournie par l'analyseur syntaxique XIP (Ait-Mokthar et al. 2002). Une version de l'extracteur d'opinions a été conçue pour l'anglais (Brun 2011), nous l'adaptons pour le français : l'analyseur est enrichi avec un lexique contenant les polarités associées aux mots, et par un ensemble de règles syntactico-sémantiques qui extraient de relations d'opinions à « grain fin », c'est-à-dire des opinions associées aux éléments sur lesquelles elles portent. Par exemple, pour la phrase suivante, le système extrait :

« *Ce livre est très prenant, l'histoire est vraiment bien racontée.* »

OPINION[POSITIVE](prenant, livre) & OPINION[positive](vraiment bien racontée, histoire).

Où le premier élément de la relation est le prédicat porteur de la polarité et le deuxième élément est la cible de l'opinion. La première relation est déduite de la relation attributive détectée par XIP entre *livre* et *prenant*, adjectif de polarité positive. La deuxième relation découle du fait que le sujet passif de la phrase est en relation avec un prédicat modifié par un adverbe de polarité positive.

Actuellement, outre le développement des règles d'extraction de relations d'opinions, l'emphase a été mise sur le traitement des phénomènes de négation. Du point de vue des ressources lexicales, nous disposions initialement d'un lexique de polarités préliminaire, construit à la fois manuellement et en adaptant à notre système une partie du lexique de Blogoscopie, (Vernier et Monceaux 2010). Il était constitué de 714 mots dont 418 adjectifs (58,54%), 163 noms (22,83%), 111 verbes (15,55%) et 22 adverbes (3,08%). L'objectif des présents travaux est l'extension de ce lexique de polarités en utilisant les familles de mots.

3.3 Expériences et résultats

Pour évaluer la qualité de la ressource, nous utilisons une méthode similaire à celle utilisée dans (Brun 2011) qui a été conçue pour évaluer la performance de l'extracteur d'opinions pour l'anglais. Ici, le système est testé afin d'évaluer l'apport de la ressource lexicale. Nous avons donc exécuté une série des tests concernant l'intégration des lexiques de polarité. Les deux corpus constitués précédemment sont utilisés pour évaluer les performances du système de détection d'opinions sur la classification des revues. Ces corpus peuvent être considérés comme annotés pour la classification, l'auteur assignant une note globale à la revue. Nous utilisons les relations d'opinions extraites par le système pour entraîner et tester un classifieur SVM binaire (*SVMLight*, (Joachims 1999)) afin de classer les revues comme positives (note entre 3 et 5) ou négatives (note entre 0 et 2). La référence de cette évaluation est calculée avec le lexique initial, avant intégration de la ressource lexicale constituée. Le protocole d'évaluation utilise 200 revues du corpus « livre » (100 négatives, 100 positives, choisies aléatoirement) pour entrainer, valider et tester et 4000 revues du corpus « restaurant » (2000 négatives, 2000 positives, choisies aléatoirement) pour entrainer, valider et tester.

Les traits utilisés pour le SVM correspondent au nombre d'occurrences d'une relation d'opinion portant sur un une cible donnée : par exemple, si le système extrait, pour une revue donnée, 1 fois la relation *OPINION[positive](vraiment bien racontée, histoire),* le trait du SVM et sa valeur sont OPINION_POSITIVE_CIBLE_histoire = 1 ; si le système extrait les relations *OPINION[negative](décevant, cuisine)* et *OPINION[negative](mauvais, cuisine)*, le trait et sa valeur sont OPINION_NEGATIVE_CIBLE_cuisine = 2, etc.

	Taux d'accord	Exactitude classif. Corpus « livres »	Exactitude classif. Corpus « restaurant »
Référence	*(1)*	79,80%	74,80%
Adjectifs	100% *(2)*	81,50%	76,20%
	100% + 75%/66% *(3)*	82,50%	77,10%
Lexiques obtenus par propagation	100% *(4)*	**83,00%**	**77,90%**
	100% + 75%/66% *(5)*	82,40%	77,20%

TABLE 1 – Résultats de la classification des revues selon le type d'information lexicale

Pour les deux corpus, nous avons calculé l'exactitude de la classification par le SVM (« accuracy ») selon l'intégration des données lexicales. Nous avons distingué les ressources pour lesquelles l'accord entre annotateurs était de 100% lors de l'annotation initiale des adjectifs et celle pour lesquels l'accord était majoritaire (75% et 66%[2]). Nous

[2] Nous n'avons pas intégré les ressources pour lesquelles l'accord initial était plus faible.

avons donc mené 5 jeux de tests pour les deux corpus : (1) avec les ressources initiales, (2) en intégrant les adjectifs annotés avec un accord de 100%, puis (3) en intégrant les adjectifs dont le taux d'accord d'annotation manuelle était majoritaire, enfin les lexiques obtenus par propagation de la polarité des adjectifs aux familles de mots. Nous avons distingué celles pour lesquelles le taux d'accord inter-annotateur était à 100% (4) et celles pour lesquelles l'accord était entre 75% et 66% (5).

Les résultats sur les deux corpus de test sont synthétisés dans le tableau 1. Nous constatons que dans la plupart des cas, il y a un gain, et que la configuration optimale est obtenue dans le cas du test (4). Seule l'intégration du lexique obtenu automatiquement par propagations aux familles de mots pour les adjectifs dont l'accord inter-annotateurs était de 75-66% fait régresser le taux d'exactitude, ce qui montre l'importance de cet accord. En outre, on observe que les tendances sont les mêmes que le corpus traite de littérature ou de restaurants.

4 Conclusion

Dans cet article, nous nous sommes intéressées à la création automatique d'un lexique pouvant améliorer les résultats d'un système d'extraction d'opinions. Pour le construire, nous avons capitalisé sur une ressource regroupant les mots en familles morphologiques et avons observé si les informations sémantiques concernant la polarité des mots pouvaient se maintenir ou non au sein d'une même famille. Nous avons ainsi propagé automatiquement les polarités d'une liste d'adjectifs annotés manuellement vers les mots de leurs familles.

Les résultats de cette expérience montrent que, dans 78,89% des familles avec un seul adjectif, notre hypothèse se vérifie, à condition que certains affixes porteurs de modifications sémantiques (négation) soient pris en compte. L'utilisation de ce lexique dans un système d'extraction d'opinions montre que les résultats de la classification des opinions s'améliorent. Les améliorations plus significatives concernent les mots pour lesquels l'accord inter-annotateur est de 100%, ce qui implique que l'étape d'annotation initiale conditionne fortement les résultats finaux.

En perspective, différentes pistes sont envisageables. Concernant la construction du lexique, notre intérêt porte sur le traitement de la polysémie ainsi que sur la recherche de critères plus fins pour propager les polarités dans le cas de familles à plusieurs adjectifs. Concernant l'utilisation du lexique par le système d'extraction d'opinions, nous poursuivrons les expériences avec ces ressources améliorées ainsi que sur d'autres corpus.

Références

AIT-MOKTHAR, S., CHANOD, J.P. (2002). Robustness beyond Shallowness: Incremental Dependency Parsing. *Special Issue of NLE Journal.*

BRUN C. (2011). Detecting opinions using Deep Syntactic Analysis. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria.

HU M. ET LIU B. (2004). Mining and summarizing customer reviews. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle,

Washington, USA.

CHOI Y. ET CARDIE C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *In Proceedings of the conference on empirical methods in NLP, EMNLP-09*, pages 741-748, Edinbourg, Ecosse.

ESULI ET SEBASTIANI (2005). Determining semantic orientation of terms through gloss classification. *In Proceedings of CIKM,* pages 617-624,

ESULI A. ET SEBASTIANI F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. *Dans les actes de LREC-06,* Gêne, Italie.

GALA, N., HATHOUT, N., NASR, A., REY, V. ET SEPPÄLÄ, S. (2011) Création de clusters sémantiques à partir du TLFi. Actes de *Traitement Automatique des Langues Naturelles* (TALN 2011). Montpellier, juin 2011.

GALA N. ET REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. Actes de *Traitement Automatique des Langues Naturelles* (TALN 2008), Avignon, juin 2008.

HATZIVASSILOGLOUS V. ET MCKEOWN K. (1997). Predicting the semantic orientation of adjectives. Actes de ACL-97, pages 174-181, Madrid, Espagne.

JOACHIMS T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press.

KIM S. M. ET HOVY E. (2004). Determining the sentiment of opinions. Actes de *COLING-04*, pp. 1367-1373, Barcelone, Espagne.

LAFOURCADE M. (2007) Making people play for Lexical Acquisition. Actes de *7th Symposium on Natural Language Processing*, Pattaya, Thailand.

LU Y., CASTELLANOS M., DAYAL U., ZHAI CH. (2011). Automatic construction of context-aware sentiment lexicon: an optimization approach. Actes de *WWW Conference, session semantic analysis*. Hyderabad, India, pp. 347-356.

PANG B., LEE L. ET VAITHYANATHAN S. (2002) Thumbs up? Sentiment classification using machine learning techniques. *In Proceedings of the conference on empirical methods in NLP, EMNLP-02*, pp. 79-86, Philadelphia, USA.

STRAPPARAVA C. ET VALITUTTI A. (2004). WordNet Affect : an affective extension of WordNet. Actes de *4th International conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

TABOADA M., BROOKE J., TOFILOSKI M., VOLL K., STEDE M. (2011) Lexicon-based methods for sentiment analysis. Dans *Computational Linguistics,* Volume *37 (2)*.

TURNEY P. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Actes de *ACL-02*, pages 417-424, Philadelphia, USA.

VERNIER M. ET MONCEAUX L. (2010) Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Revue TAL volume 51 (1),* pp. 125-149.