Combiner analyse syntaxique de surface et apprentissage supervisé pour la fouille d'opinion ciblée : expérimentations sur des données d'opinion concernant les livres

Jeanne Villaneau*, Stefania Pecore*
Farida Saïd*,**, Pierre-François Marteau*

* IRISA, Campus de Tohannic, Université de Bretagne-Sud, 56000 Vannes
jeanne.villaneau, stefania.pecore, farida.said, pierre-francois.marteau@univ-ubs.fr
** LMBA, Campus de Tohannic, Université de Bretagne-Sud, 56000 Vannes

Résumé. La fouille d'opinion ciblée est une tâche complexe, susceptible de bénéficier de l'apport d'approches variées. Nos expérimentations testent des combinaisons de méthodes sur un corpus d'avis d'internautes concernant les livres. Sur ces données et pour ce qui concerne la polarité de l'opinion, des résultats prometteurs ont été obtenus par une approche basée sur une analyse linguistique de surface et un lexique enrichi par les informations discriminantes basées sur des méthodes de classification statistiques supervisées.

1 Introduction

La fouille d'opinion est devenue un champ de recherche important du Traitement Automatique des Langues (TAL) et sa maturité est attestée par les nombreux états de l'art dont elle a fait l'objet [Chapate et all. (2015); Feldman (2013); Liu (2012), etc.]. Son objectif est d'analyser l'ensemble des opinions des internautes sur un objet donné. La tâche se veut plus précise avec la fouille d'opinion ciblée (ABSA ; Aspect Based Sentiment Analysis) [Liu (2012)] : les cibles correspondent aux différentes caractéristiques des objets concernés et les avis des internautes sont analysés en fonction de ces cibles. L'ABSA a fait l'objet de différents challenges, parmi lesquels SemEval-2014, 2015 et 2016 [Pontiki (2016)].

Ciblée ou non, la fouille d'opinion est une tâche extrêmement complexe pour de multiples raisons : forme indirecte de l'expression, humour, subjectivité, nécessité d'un traitement fin de la négation, etc. Les approches utilisées sont nombreuses et variées, généralement divisées entre méthodes statistiques, généralement supervisées [Pang et Lee (2008)] et méthodes basées sur un lexique d'opinion [1], à base de règles [Neviarouskaya et al. (2010)] ou statistiques [Wilson et al. (2005)], une tendance actuelle étant de combiner différentes approches.

La fouille d'opinion ciblée peut être partagée en trois sous-tâches : (i) détection de phrases ou portion de phrases porteuses d'opinion, (ii) détection de la cible relative à l'opinion émise, (iii) détection de la polarité et éventuellement, de l'intensité de l'opinion en question. Nos expérimentations concernent les deux dernières. Par ailleurs, les travaux présentés analysent des

1. Pour un état de l'art des méthodes basées sur l'utilisation d'un lexique, cf. Taboada et al. (2011).

avis laissés par les internautes concernant les livres : un domaine plus complexe que ceux généralement envisagés, où l'opinion est souvent exprimée de manière indirecte ou complexe et les cibles parfois peu différenciées. La langue, intermédiaire entre la langue académique et celle des tweets, rend possible l'utilisation d'une analyse syntaxique de surface (*chunking*). Notre étude s'est plus particulièrement attachée à explorer les améliorations que peut apporter ce type d'analyse, jamais testée en ABSA précédemment à notre connaissance. La section 2 décrit les corpus et la tâche de détection des cibles, où le *chunking* semble inopérant ; la section 3 montre les avantages de la méthode pour la détection de la polarité.

2 Corpus et détection des cibles

2.1 Corpus et annotations

En fouille de données ciblée, l'absence de corpus disponible en langue française dans le domaine des livres qui nous intéressait, nous a conduit à combler ce manque : nous avons annoté 900 avis d'internautes collectés sur le site *Amazon.fr*. Le schéma d'annotation proposé se veut générique et très précis ; il distingue 5 cibles principales, elles-mêmes divisées en attributs pour un total final de 21 classes possibles. L'annotation distingue (a) l'expression de l'opinion ; (b) l'entité à laquelle cette opinion se rattache (si exprimée) ; (c) la cible et (d) une valeur entière comprise entre -2 et 2, qui exprime la polarité de l'opinion en même temps que son intensité. Au total, le corpus comporte 3300 expressions d'opinion ainsi annotées [Pecore et Villaneau (2018)].

Pour les tâches que nous nous proposions de réaliser, nous avons réduit le nombre de cibles à 8 en effectuant des regroupements et nous n'avons pas tenu compte de l'intensité des opinions. Comme corpus de tests, nous avons présélectionné 340 phrases ou portions de phrases porteuses d'opinion dans la partie non annotée du corpus et nous avons choisi pour chacune d'elles l'une des 8 cibles et la polarité de l'opinion correspondantes.

2.2 Détermination des cibles

Pour ce qui concerne la détection des cibles, le domaine étudié se distingue des domaines classiquement étudiés par deux particularités. La première est que la désignation des entités n'y suffit pas. Par exemple, dans la phrase : *« le livre est bien écrit »*, la cible est le *Style* alors que l'entité, *le livre*, ne la désigne pas. La deuxième est que les classes diffèrent très fortement de par leurs importances relatives : avec près de 45% des annotations, la classe qui désigne une opinion globale sur l'ouvrage (*General*), est très fortement prévalente ; à l'inverse, la classe *Illustrations* en regroupe moins de 1%.

Nous avons testé plusieurs modèles statistiques classiques entraînés avec les noms, adjectifs, verbes et adverbes (sauf mots grammaticaux) figurant dans les expressions d'opinion et les entités annotées : kNN, Random Forest, Neural Network, SVM, Fuzzy classification, SOM (Kohonen package in R), etc. La prévalence de la classe *General* pose problème à tous les classifieurs, jusqu'à aboutir à une totale inefficacité de certains qui classent la quasi-totalité des tests en *General* [2]. Nous avons également utilisé Word2Vec pour pouvoir prendre en compte

2. Pour plus d'informations, on pourra se reporter à Villaneau et al. (2018)

les mots du corpus de test absents du corpus d'entraînement (SVMW2V). Par ailleurs, les essais d'introduction de paramètres linguistiques n'ont pas donné de résultats concluants. Les meilleurs résultats ont été obtenus en utilisant une méthode en sac de mots lemmatisés avec les SVM (noyau linéaire) et les Random Forest (ntree=500) pour des macro-moyennes respectives (F1-scores) de 0,673 et 0,642. Si l'utilisation de Word2Vec n'a pas permis d'améliorer les résultats des SVM, un vote majoritaire entre SVM, RF et SVMW2V a permis d'augmenter le score final, avec une macro-moyenne de 0,703.

Dans une tâche similaire (livres scolaires, même nombre de classes), Hamdan et al. (2016) obtient des F1-scores compris entre 0,610 et 0,615. Dans Semeval 2016, la détermination des cibles dans le domaine des restaurants obtient respectivement en Anglais et en Français des F1-scores égaux à 0,73 et 0,61. La qualité de nos résultats confirme l'intérêt de prendre en compte conjointement les lemmes annotés comme entités et comme expression de l'opinion. Elle confirme également que la détermination des cibles est avant tout un problème lexical.

3 Chunking et polarité de l'opinion

3.1 Approche par analyse syntaxique de surface et lexique d'opinion

Nous avons choisi une approche globale proche de la sémantique compositionnelle utilisée par Moilanen et al. (2010) : le système calcule le score des mots, expressions et *chunks* en utilisant le lexique d'opinion et les règles définies sur les différents types de groupes de mots (*chunks*). Le score d'une phrase s'obtient en combinant ces scores et en appliquant des fonctions définies pour prendre en compte le temps et les modes des verbes ainsi que certains éléments structurants de la phrase (*mais, pourtant,* etc.).

Pour la reconnaisance des différents *chunks*, nous avons réadapté un outil précédemment utilisé dans une étude de patrons sémantiques [El Maarouf et al. (2011)]. La difficulté rencontrée à ce stade est la recherche d'un compromis entre les erreurs inévitables du *chunking* - en particulier causées par les agrammaticalités du texte - et la précision nécessaire.

3.1.1 Le lexique d'opinion

Le lexique d'opinion est l'une des pierres angulaires d'un système de détection de l'opinion dans une approche linguistique [Breck et Cardie (2017)]. Nous avons d'abord utilisé un lexique obtenu par compilation de deux lexiques existants : la norme émotionnelle *ValEmo* [Syssau et Font (2005)] et une extension de la norme *F-POL* [Vincze et Bestgen (2011)]. Le lexique obtenu s'est avéré trop général : par exemple le verbe *dormir* qui y est positivement connoté, exprime généralement une opinion négative lorsqu'il est utilisé dans une critique de livres.

De plus, nous sommes arrivés à la conclusion énoncée par Taboada et al. (2011) que *« more words may lead to including more noise" »[plus de mots peut conduire à plus de bruit]* et nous avons réduit le vocabulaire à celui qui exprime clairement une opinion sur un livre. En revanche, nous avons introduit un lexique d'expressions courantes en langue française, une étude plus approfondie étant en dehors du cadre de notre étude.

3.1.2 Traitement de la négation

Le traitement de la négation est une pierre d'achoppement en analyse d'opinion. Wiegand et al. (2010) en présente diverses approches, les deux problèmes principaux étant d'une part, sa détection et d'autre part, la détermination de sa portée.

Bon nombre de difficultés se retrouvent dans d'autres langues, notamment l'anglais : fausses négations (*« non seulement »*), doubles négations, patrons spécifiques à chaque mot négatif, négation exprimée de manière indirecte ou subtile [Pang et Lee (2008); Asmi et Ishaya (2012)], etc. L'un des problèmes spécifique à la langue française est l'usage de l'adverbe *ne* : fréquemment omis dans la langue orale, son omission tend à se répandre dans la langue écrite informelle, jusqu'à conduire à des expressions ambigües dans la langue écrite, telles que *« il y en a plus »*, qui peut signifier aussi bien *il y en a davantage* que *il n'y en a plus*.

Dans notre approche compositionnelle, une fonction est attachée à la négation : elle modifie le score des éléments qui sont dans la portée de cette dernière. Nous avons testé plusieurs fonctions proposées dans la littérature sans constater de modification globale des résultats obtenus : il semble donc que le choix exact de la fonction soit beaucoup moins significatif que la détermination de la portée. Globalement, il convient de prend en compte le fait qu'une négation atténue l'intensité tout en étant généralement l'indice d'une opinion plutôt négative.

3.1.3 Les chunks et leur usage

Le *chunking* permet de prendre en compte partiellement les problèmes liés à l'association des mots si l'on associe un traitement spécifique à chaque forme de *chunk*. Les *chunks verbaux* jouent un rôle majeur dans notre approche. Outre la prise en compte du temps et du mode, ils transmettent ou non une éventuelle négation aux chunks qui les entourent en fonction de la nature du verbe qui est à leur tête. Le rôle essentiel des *chunks nominaux* est la prise en compte des associations entre adjectif et nom, certains adjectifs induisant une polarité stable, positive ou négative (par exemple *mauvais, excellent*) alors que d'autres accentuent, atténuent ou inversent la polarité du nom qu'ils qualifient (*grand*, *faux*). Les *chunks adjectivaux* ou *adverbiaux* permettent de prendre en considération les modifieurs qui peuvent, comme les adjectifs, atténuer, renforcer ou inverser la polarité (*très, trop, un peu*, etc.) [Zhang et al. (2012)].

3.2 Mises en œuvre et résultats

En guise de baseline, nous avons utilisé plusieurs méthodes statistiques classiques pour classifier les 340 phrases de test entre opinion positive et opinion négative. SVM (avec un noyau linéaire), Glmnet (régression logistique), NeuralNet (20-5) (NN) et Random Forest (500 arbres) (RF) obtiennent des F1-scores compris entre 0,768 et 0,803 (cf. table 1), par une approche en sacs de mots utilisant comme variables les lemmes des noms, verbes, adjectifs et adverbes présents dans les annotations. Contrairement à ce qui avait été observé dans la détection des cibles, un vote majoritaire entre méthodes n'améliore en rien les résultats.

Les résultats obtenus par l'approche basée sur le *chunking* (Chv1) avec le lexique utilisé s'avèrent décevants : avec un F1-score de 0,774, ils sont inférieurs à ceux obtenus par trois des quatre baselines. Le nombre important de phrases qui obtiennent un score nul suggère un problème de lexique.

SVM	Glmnet	NN	RF	Chv1	SPLex
0,803	0,800	0,780	0,768	0,774	**0,859**

TAB. 1 – *Polarité : F1-score (macro-moyenne) des quatre baselines, du système linguistique avant (Chv1) et après (SPLex) enrichissement du lexique.*

Glmnet et Random Forest donnent des indications concernant l'importance des variables statistiques . Une sélection des premiers mots (une cinquantaine) rendu par la fonction *importance* du package Random Forest du logiciel R nous a permis de compléter notre lexique d'opinion avec la trentaine d'entre eux qui n'y figuraient pas (noms et adjectifs essentiellement). Ce lexique enrichi améliore considérablement l'efficacité du système (cf. SPLex table 1). Comparé aux résultats obtenus par Hamdan et al. (2016) sur une tâche très similaire (F1-score de 0,794) et les meilleurs F1-scores rapportés sur cette tâche à Semeval 2016 (0,840 et 0,605) dans les domaines des restaurants et des ordinateurs, le F1-score (0,859) obtenu par SPLex est une très bonne performance.

4 Conclusion et perspectives

La cible d'une opinion peut généralement être déterminée en étudiant les mots utilisés pour l'exprimer : les approches lexicales et statistiques s'avèrent très efficaces dans cette tâche et nos expérimentations ne remettent pas en cause cette prépondérance. En revanche, elles suggèrent que, combinée à un lexique d'opinion très ciblé, une analyse de surface est une approche efficace pour déterminer la polarité de l'opinion. Elles suggèrent également que la pertinence d'un lexique d'opinion est très dépendant du domaine et, très probablement, du corpus lui-même, ce qui, en soi, restreint la généricité des approches basées sur de tels lexiques. Ces résultats demandent à être confortés et validés sur d'autres corpus et d'autres domaines.

Références

Asmi, A. et T. Ishaya (2012). The second international conference on advances in information mining and management negation identification and calculation in sentiment analysis. In *IMMM (Advances in Information Mining and Management)*.

Breck, E. et C. Cardie (2017). *Oxford Handbook of Computational Linguistics.* (2nd ed.)., Chapter Opinion mining and sentiment analysis. Oxford University press.

Chapate et all. (2015). Survey on sentiment analysis and its classification technique. In *National Conference on Advances in Computing (NCAC 2015)*, Stroudsburg, PA, USA, pp. 793–801. Association for Computational Linguistics.

El Maarouf, I., J. Villaneau, et S. Rosset (2011). Extraction de patrons sémantiques appliquée à la classification d'Entités Nommées. In *TALN'2011*, Montpellier, France.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM 56*(4), 82–89.

Hamdan, H., P. Bellot, et F. Bechet (2016). Sentiment analysis in scholarly book reviews. In *arXiv preprint arXiv :1610.03106*.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers.

Moilanen, K., S. Pulman, et Y. Zhang (2010). Packed feelings and ordered sentiments : Sentiment parsing with quasi-compositional polarity sequencing and compression. In *Proceedings of workshop WASSA 2010 at ECAI 2010*, pp. 36–43.

Neviarouskaya, A., H. Prendinger, et M. Ishizuka (2010). Affect analysis model ;novel rule-based approach to affect sensing from text. *Natural Language Engineering 17*(1).

Pang, B. et L. Lee (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2*(1-2), 1–135.
english

Pecore, S. et J. Villaneau (2018). Complex and Precise Movie and Book Annotations in French Language for Aspect Based Sentiment Analysis. In *Proceedings of LREC 2018*, Miyazaki, Japan. ELRA.

Pontiki, M. *et al*. (2016). Semeval-2016 task 5 : Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pp. 19–30. Association for Computational Linguistics.

Syssau, A. et N. Font (2005). évaluations des caractéristiques émotionnelles d'un corpus de 604 mots. *Bulletin de psychologie 3*(477), 361–367.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, et M. Stede (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist. 37*(2), 267–307.

Villaneau, J., S. Pecore, et F. Saïd (2018). Aspect detection in book reviews : Experimentations. In *Proceedings of the 2nd Workshop NL4AI 2018*, Volume 2244, Trento, Italy, pp. 16–27.

Vincze, N. et Y. Bestgen (2011). Une procédure automatique pour étendre des normes lexicales par l'analyse des cooccurrences dans des textes. *TAL 52*(3), 191–216.

Wiegand, M., A. Balahur, B. Roth, D. Klakow, et A. Montoyo (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of NeSp-NLP '10*, Stroudsburg, PA, USA, pp. 60–68. Association for Computational Linguistics.

Wilson, T., J. Wiebe, et P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT'05*, Stroudsburg, PA, USA, pp. 347–354. Association for Computational Linguistics.

Zhang, L., S. Ferrari, et P. Enjalbert (2012). Opinion analysis : The effect of negation on polarity and intensity. In J. Jancsary (Ed.), *Proceedings of KONVENS 2012*, pp. 282–290. ÖGAI. PATHOS 2012 workshop.

Summary

Aspect Based Sentiment Analysis (ABSA) is a complex task, for which using combinations of various approaches can be efficient. Our work was led on a corpus of book reviews in French language and tests. On these data and to determining opinion polarity, very good results are obtained for an approach which combines outputs of a chunking with a lexicon enriched by the words as features used by statistical lexical methods.