Recherche et utilisation d'entités nommées conceptuelles dans une tâche de catégorisation

Jean-Valère Cossu[1] Juan-Manuel Torres-Moreno [1,2,3,4] Marc El-Bèze[1,2,4]
(1) Laboratoire Informatique d'Avignon - Université d'Avignon et des Pays de Vaucluse
339 chemin des Meinajaries, BP91228 84911 Avignon Cedex 9, France
(2) SFR Agorantic Université d'Avignon et des Pays de Vaucluse, 84000 Avignon Cedex
(3) École Polytechnique de Montréal, 2900 Bd Edouard-Montpetit Montréal, QC H3T1J4
(4) Brain & Language Research Institute, 5 avenue Pasteur, 13604 Aix-en-Provence Cedex 1

{jean-valere.cossu,juan-manuel.torres,marc.el-beze}
@univ-avignon.fr

RÉSUMÉ

Les recherches présentées sont directement liées aux travaux menés pour résoudre les problèmes de catégorisation automatique de texte. Les mots porteurs d'opinions jouent un rôle important pour déterminer l'orientation du message. Mais il est essentiel de pouvoir identifier les cibles auxquelles ils se rapportent pour en contextualiser la portée. L'analyse peut également être menée dans l'autre sens, on cherchant dans le contexte d'une cible détectée les termes polarisés. Une première étape d'apprentissage depuis des données permet d'obtenir automatiquement les marqueurs de polarité les plus importants. A partir de cette base, nous cherchons les cibles qui apparaissent le plus fréquemment à proximité de ces marqueurs d'opinions. Ensuite, nous construisons un ensemble de couples (marqueur de polarité, cible) pour montrer qu'en s'appuyant sur ces couples, on arrive à expliquer plus finement les prises de positions tout en maintenant (voire améliorant) le niveau de performance du classifieur.

ABSTRACT

Search and usage of named conceptual entities in a categorisazion task

The researchs presented are part of a text automatic categorization task. Words bearing opinions play an important role in determining the overall direction of the message. But it is essential to identify the elements (targets) which they are intended to relativize the scope. The analysis can also be conducted in the reverse direction. When a target is detected we need to search polarized terms in the context. A first step in an automatic learning from data will allow us to obtain the most important polarity markers. From this basis, we look for targets that appear most frequently in the vicinity of these opinions markers. Then, we construct a set of pairs (polarity marker, target) to show that relying on these couples we can maintain (or improve) the performance of the classifier.

MOTS-CLÉS : Fouille d'opinion, Marqueurs de polarité, Reconnaissance d'entités nommées.
KEYWORDS : Opinion Mining, Named Entity Recognition.

1 Introduction

Depuis fort longtemps, la prise de décision se fait toujours après consultation des points de vue d'autres personnes. On prend très souvent connaissance des critiques émises par d'autres consommateurs avant de consommer un produit ou service. Cette interaction entre

individus peut être élargie à ce qui se produit durant les campagnes précédant des élections. Depuis le développement d'Internet, de plus en plus de personnes donnent leurs avis et ces derniers étant de plus en plus disponibles, il est facile d'avoir accès à de larges corpus d'opinion. Les applications possibles de la fouille d'opinion sont multiples (Pang et Lee, 2008), systèmes de recommandation, outils de marketing, suivi de tendances etc. . Certains moteurs de recherche proposent d'ailleurs déjà des applications pour résumer les opinions des consommateurs dans des interfaces dédiées (Blair-Goldensohn et al, 2008).

L'analyse d'opinion peut se décomposer en trois sous-tâches :
1. Détection de la présence ou non de l'opinion ;
2. Classification et intensité : (très) positif, (très) négatif ou neutre ;
3. Identification des cibles et sources de l'opinion (sur quoi porte l'opinion et qui l'exprime).

Pour autant, alors qu'il est bon d'avoir un avis « général », extraire ce qui est exprimé sur un point précis est tout aussi voire plus utile. La tâche 3 pouvant être répétée en changeant de granularité, en se situant au niveau du texte entier, du paragraphe, de la phrase ou bien du fragment selon les applications envisagées. L'exemple le plus frappant peut être pris en politique, où l'enjeu n'est pas tant de convaincre ses opposants à l'ensemble à adhérer à ses propositions mais plutôt de pousser ceux qui hésitent encore à basculer de son coté en prenant appui sur un sujet donné. Dans ce cas-là, ce n'est pas sur l'ensemble de l'entité qu'il faut agir mais plutôt sur des points précis. Points qu'il reste à déterminer et que nous appellerons par la suite cibles ou Entités Nommées Conceptuelles (ENC).

2 Entités nommées

La reconnaissance des entités nommées (REN) fait partie de l'extraction d'information. Elle consiste à délimiter et catégoriser certaines expressions linguistiques. Ces dernières correspondent à des ensembles de noms (entités, expressions temporelles, géographiques, etc.). Toutefois les entités nommées peuvent être plus spécifiques à un domaine et on parle alors d'entités nommées d'intérêt spécifique. Ici, il s'agira plus véritablement de sous-entités dans la mesure où elles correspondront à des éléments constitutifs d'entités (personnes, entreprises, produits ou services) et représenteront donc des concepts qui seront déterminés en fonction des données d'apprentissage sans connaissance *a priori* de la langue et du domaine. La délimitation d'EN se fait habituellement sous la forme d'annotations en utilisant des listes de connaissances ou avec l'aide d'experts (Dutrey et al, 2012). Les besoins en connaissances linguistiques (et en connaissances du domaine) deviennent vite très importants. Nous proposons de les détecter de façon semi-automatique, puis de les utiliser à des fins de classification tout en gardant à l'esprit que ces dernières peuvent faire un excellent support permettant de produire automatiquement des résumés « polarisés ».

2.1 Propositions

L'hypothèse de travail est la suivante : lorsqu'une opinion est exprimée, cette dernière l'est forcément sur un élément de l'entité critiquée ou sur l'entité dans sa globalité. Cet élément sera appelé cible ou sous-cible selon son niveau de granularité. A rebours, si dans un message une cible est citée par une personne, nous supposons que c'est parce qu'elle souhaite en dire ce qu'elle en pense ou à la limite dire qu'elle n'en pense pas grand-chose et le fait de l'exprimer ainsi n'est probablement pas à négliger. *A contrario*, nous pouvons

également considérer qu'en l'absence de cible dans la critique (critiques très courtes) le marqueur de polarité doit probablement porter sur l'entité par exemple : « super film ».

Un des objectifs visés est l'extraction de couples (cible, marqueur de polarité) permettant à la fois de catégoriser le message mais également de constituer un résumé de la représentation de l'entité (ou du produit dans le cas d'un système de recommandation). Ces couples ne sont pas limités aux seuls concepts identifiés par des experts du domaine ou par ce qui est communément admis (Pupier, 1998), mais sont censés émerger des avis analysés conformément à la façon dont ils ont été exprimés. Cette façon de procéder tient implicitement compte de la restriction des différents sens d'un mot à ceux qui ont cours dans le domaine abordé par les auteurs des critiques (Riloff et Wiebe, 2003). C'est le cas du terme « navet » qui est un légume plus ou moins apprécié par les gastronomes mais aussi et surtout pour ce qui nous concerne un mauvais film dans le domaine du cinéma. On pourrait se baser sur des listes de marqueurs d'opinion comme le propose (Navigli, 2009) mais s'il nous fallait préétablir leur polarité cela impliquerait une coûteuse désambiguïsation lexicale.

La méthode consiste à extraire dans le corpus les éléments les plus porteurs d'opinions (marqueurs de polarité). Une fois ceux-ci extraits, nous cherchons, à proximité de ces derniers, s'il existe des éléments à pouvoir discriminant modéré, non présents dans un anti-dictionnaire (SL composé principalement de mots-outils). Si la fréquence de ces éléments dépasse un plancher déterminé empiriquement, nous pouvons les considérer comme des « cibles ». Nous pouvons considérer l'ensemble de ces cibles de même que les métadonnées film ou pseudo comme des ENC.

3 Données

Des expériences de classification automatique ont été menées sur un corpus de micro-critiques (µC) de cinéma provenant du portail communautaire Vodkaster[1].

Chaque µC étant un tuple[2] : utilisateur, film, note[3], critique correspondant à la définition d'une opinion donnée par (Liu, 2012).

- L'échelle des notes comporte dix barreaux espacés de 0,5 point entre 5 et 0,5 ;
- La critique est dite µ-critique car d'une longueur maximale de 140 caractères.

Le corpus contient 77 000 µC, les 20 000 plus récentes constituent les corpus de développement et test (10 000 chacun), le reste étant considéré comme apprentissage[4]. L'échelle des notes est dans le cadre de nos expériences ramenée de façon volontaire à deux barreaux. Nous avons tablé sur le fait que les positions les plus tranchées feraient ressortir plus de cibles associées à des qualificatifs. Les seuils des deux barreaux ont été déterminés de façon empirique : Positif (note>4) et Négatif (note<2). Les critiques dites neutres (dont la note vaut entre 2 et 4 non inclus) sont pour l'instant exclues des corpus d'apprentissage, développement et test.
Malgré les tailles restreintes des critiques et la liste de cible, les utilisateurs arrivent à exprimer plusieurs opinions (parfois opposées) sur les différents éléments des films.

[1] http://www.vodkaster.com
[2] Nous envisageons d'utiliser par la suite d'autres métadonnées comme : acteurs, réalisateurs, genre.
[3] Note mise par l'utilisateur lorsqu'il a déposé sa critique sur le portail.
[4] Bien évidemment, les 3 intersections de ces corpus pris 2 à 2 sont vides.

Les critiques nuancées ou équilibrées (µC contenant un des « pivots » prédéterminés) sont retirées des différents corpus. Nous avons à cet effet sélectionné uniquement les deux « pivots » les plus fréquents[5] dans le corpus d'apprentissage : « mais » et « malgré ». Ne seront donc présentes dans le corpus de test que les µC *a priori* fortement polarisées contenant au moins une cible et ne contenant aucun de ces « pivots » de langage ce qui réduit à 5 010 critiques sur l'ensemble des 10 000 présentes à l'origine dans le corpus.

Deux systèmes concurrents ont été mis en place : l'un prenant en compte le couple (cible-marqueur de polarité), l'autre se basant sur l'ensemble des termes présents dans la µC. Toutes les expériences présentées tiennent compte de la polarité du pseudo de l'utilisateur ainsi que celle du titre du film, qui ont été intégrés comme des termes à l'intérieur de la µC et deviennent de fait porteurs d'opinions.

3.1 Classifieurs

Le premier classifieur utilisé est un CosinusGini (M1) (Torres et al, 2011). Il est basé sur l'ensemble des termes présents dans la critique. Le classifieur Cosinus a été préféré à d'autres méthodes plus classiques et parfois plus performantes comme les SVM (Collobert et al, 2002) du fait que ces méthodes ne permettent pas d'avoir facilement accès aux éléments ayant contribué à la classification.

Le second (M2) est une variante du premier, ne prenant cette fois en compte que les couples (cible, marqueur de polarité) comme mentionné en 2.1 et repris en 3.2 ; les marqueurs de polarité seront recherchés avec un rayon de ***R*** (variable entre 1 et 9) termes de part et d'autre de la cible. Nous avons fait varier le rayon afin d'évaluer l'impact du contexte sur la catégorisation de la cible.

Les performances sont mesurées en termes de rappel et de précision. Il arrive parfois pour des petits rayons qu'il n'y ait aucun couple présent dans une µC pour cette raison nous comparons M1 et M2 sur la précision à un même niveau de rappel, celui déterminé par M2.

3.2 Liste de cibles

Les cibles sont déterminées de manière semi-automatique selon le protocole suivant :

A partir d'un apprentissage, le système détermine les termes ayant la plus forte contribution dans chacune des catégories. Puis il cherche à proximité de ces derniers s'il existe des éléments, non présents dans l'anti-dictionnaire et n'étant potentiellement pas de forts marqueurs de polarité. Le travail de relecture se trouve être ici assez limité, il consiste à contrôler les sorties du système pour valider ce qui est conservé comme cible ou non. Ceci est nettement moins coûteux qu'un travail de *brainstorming* avec des experts du domaine. Cette façon présente un autre avantage de taille : celui de coller à la langue dont on perçoit l'évolution rapide notamment dans les réseaux sociaux. Pour illustrer notre propos nous donnons quelques exemples (extraits à partir d'une première liste d'environ 550 cibles) avec leur nombre d'apparitions sur l'ensemble du corpus : « *acteurs* » (3 000), « *mise en scène* » (2 000), « *réalisation* » (931), « *esthétique* » (630).

[5] On aurait pu en rajouter d'autres comme « bien que » et « et pourtant » (130 et 150 occurrences).

Listons aussi quelques termes porteurs de polarité purement positive[6] auxquels on n'aurait pas forcément pensé en premier lieu[7] : « coup de poing » (42;14), « *norme* » (33;9), « *exaltant* » (32;10). Ce dernier apparaît d'ailleurs une fois en négatif dans le test à propos du film Juno. On en comprend la raison au vu de l'ironie de son contexte : « aussi exaltant qu'un fœtus mort ». En tête des termes à polarité négative, on trouve : « regardable » (17;1), « bidon » (16;5), « beurk » (16;5) …

Puis, l'enrichissement de la LC se fait selon les deux procédures suivantes :

- Procédure P1 : trouver des cibles permettant de couvrir des µC où aucun terme n'appartient à la liste de cibles. Ne sont alors retenus que les termes présents dans le plus grand nombre de µC résiduelles mais qui permettraient également d'améliorer la couverture des µC déjà sélectionnées. Les termes ayant un pouvoir discriminant proche de celui des mots outils sont filtrés.

- Procédure P2 : dans le cas de µC correctement étiquetées par M1 mais pas par M2. L'objectif est de chercher dans le voisinage du terme de polarité P, qui a le plus contribué à la bonne décision de M1, un terme T répondant au critère : $T \Subset (LC \cap SL)$. Seront alors proposés les termes se trouvant dans le plus grand nombre de µC résiduelles, avec fréquence élevée et pouvoir discriminant supérieur à celui des mots outils.

Itérer ces deux procédures a permis d'augmenter facilement la couverture de la LC en refrénant l'accroissement de sa taille (550 puis 982 cibles). Parmi les 5 010 critiques restant dans le corpus après retrait de celles contenant un pivot. 4 580 contiennent au moins une des cibles présentes dans la liste. La couverture est d'environ 2,9 cibles par µC traitée.

En s'appuyant sur les marqueurs de polarité se trouvant à proximité des cibles, et donc en filtrant ce que l'on peut considérer comme du bruit, on cherche à éliminer une partie de ce qui pourrait amener à prendre une mauvaise décision.

4 Expérimentations

4.1 Résultats

Les recherches présentées ici mettent en avant l'utilisation des couples (cible-marqueur de polarité). L'extraction d'un couple peut suffire à catégoriser un tweet. Au lieu d'opter pour un protocole lourd d'évaluation de la pertinence des cibles détectées nous avons choisi d'en faire une estimation certes grossière mais peu coûteuse. Leur extraction peut être considérée comme valide dès que la prise en compte des seuls couples présents permet de faire aussi bien qu'un classifieur utilisant l'intégralité des termes de la µC (Table 1 pour R=7).

Une première série d'expériences a été menée (pour une LC comprenant 550 cibles). Avec un rayon égal à 7, on trouvait 4 449 critiques du développement contenant au moins une cible, M1 en classait correctement 3 957 (soit 88,94%) contre 3 975 (89,35%) pour M2. En ramenant le rayon à 1, il ne restait que 3286 µC, M1 retrouve correctement la classe de 2977 µC (90,59%) et 2 827 (86,03%) pour M2.

[6] La polarité de ces termes dans le corpus prend parfois le contrepied des usages courants.
[7] Avec leurs fréquences d'apparitions (apprentissage et développement).

Afin de pouvoir intégrer dans le test des critiques ne contenant pas de cibles (c'est le cas des critiques très courtes ne contenant qu'un seul terme très souvent porteur de polarité) nous avons considéré que l'entité (ici le film) pouvait être une cible. LC a donc été enrichie et on arrive à 982 « cibles » potentielles. La couverture passe à environ 3,4 cibles par µC traitée contre 2,9 avec la première liste.

R	Dev (M1)	Dev (M2)	Corpus	Test (M1)	Test (M2)	Corpus
1	3540 (91.09)	3399 (87,47)	3886	3453 (90.00)	3311 (86.36)	3834
2	4137 (90.00)	4031 (87.77)	4593	4025 (89.54)	3892 (86.59)	4495
3	4288 (89.70)	4242 (88.84)	4780	4179 (89.39)	4083 (87.34)	4675
4	4318 (89.60)	4278 (88.77)	4819	4216 (89.38)	4153 (88.04)	4717
5	4327 (89.51)	4316 (89.28)	4834	4228 (89.31)	4182 (88.34)	4734
6	4337 (89.51)	4349 (89.76)	4845	4234 (89.25)	4204 (88.62)	4744
7	4340 (89.52)	**4366 (90.06)**	4848	4237 (89.22)	4227 (89.01)	4749
8	4344 (89.51)	4365 (89.94)	4853	4239 (89.20)	4231 (89.04)	4752
9	4346 (89.51)	4363 (89.87)	4855	4239 (89.20)	4235 (89.12)	4752

TABLE 1 – Résultats des 2 méthodes en fonction du rayon *R* en termes de précision

En effets les résultats obtenus sur le test confirment la robustesse de la méthode car, compte tenu de l'intervalle de confiance, ils sont du même niveau que ceux obtenus sur le corpus de Développement. L'intervalle de confiance est de 0,8% pour le corpus Dev et 0,9% pour le corpus Test.

4.2 Analyse des résultats et exemples

En réduisant le rayon de la fenêtre dans laquelle, autour d'une cible, sont pris en compte des marqueurs de polarité, les résultats de M2 et notamment le rappel chutent logiquement, comme le montre la dernière colonne de la Table 1. Par contre pour la première méthode (M1) le rappel reste stable car M1 prend en compte l'ensemble du contenu de chaque µC. Toutefois, la méthode M2 permet d'identifier les critiques pour lesquelles M1 est bien plus performant que sur l'ensemble du corpus (89,50 sur le Dev et 88,92 sur le Test). Cette mesure permet donc de faire ainsi un premier filtrage des données à tester. Nous constatons également que le passage de 550 à 982cibles a permis d'améliorer les résultats, il ne serait pas improbable que les résultats s'améliorent encore avec une liste de cibles plus grande.

Une des retombées essentielles de la méthode que nous proposons ici réside dans sa capacité à être utilisée pour savoir ce qui a été dit sur une entité particulière. Il suffit pour cela de

retenir avec leur orientation les couples (cible, marque de polarité) de fréquence et pouvoir discriminant élevée. Par exemple pour le film « *Skyfall* » nous avons extrait les couples suivants : « *film, d'actions raté* », « *beauté, stupéfiante* », « *mise-en-scène, classique* ». Il en va de même si l'on souhaite savoir précisément quelles sont les expressions les plus employées et les plus marquantes utilisées par un membre donné du réseau. Par exemple, parmi les expressions employées de façon marquante par « IMTHEROOKIE » un des plus gros contributeurs du site Vodkaster, on trouve : « *chef d'œuvre, ultime* », «*mise-en-scène, radicale* ». Quelques exemples avec les genres de film : « *drame, sentimental* », « *comédie, jubilatoire* ».

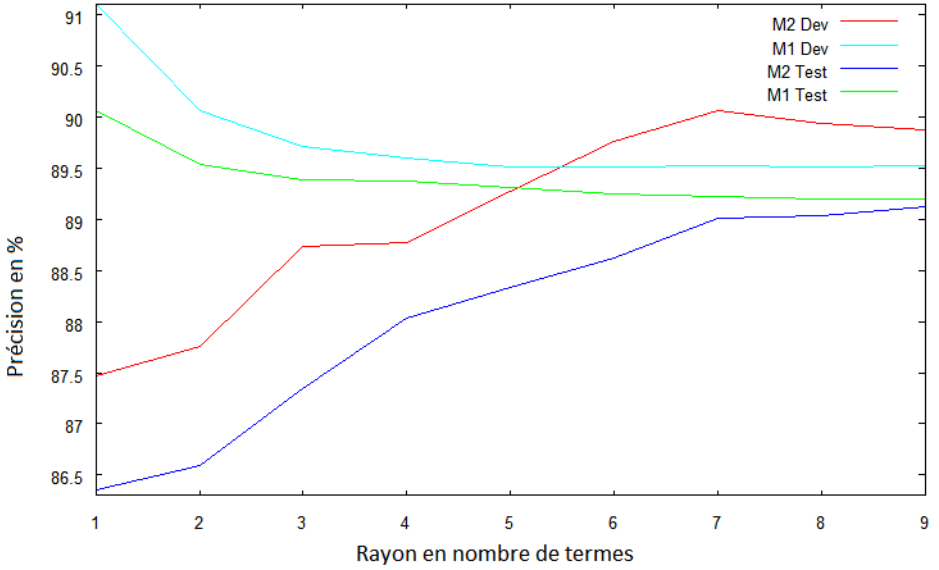

FIGURE 1 – Evolution des résultats en fonction du rayon.

Le corpus pourrait alors devenir une base de données interrogeable en fonction des besoins de chacun. On se donne ainsi la possibilité de répondre aux questions que pourrait se poser un producteur de film ; « *comment est appréciée la mise en scène des James Bond* ?», « *tous les volets de la saga sont-ils critiqués de la même manière ?* », « *quelle opinion ceux qui ont apprécié Casino-Royal ont pu avoir sur Skyfall* ? » ou encore « *alors que la saga Twillight est mal cotée, quels sont les points mis en avant par les gens qui ont aimé ces films ?* ».

4.3 Perspectives

Les couples extraits pourront servir pour des tâches d'analyse plus fine ou de « reporting », par exemple dans l'analyse d'un service à besoin relationnel où il apparaît important de connaître les points à améliorer tout autant que les points appréciés. Dans le but de produire des résumés de ce que pensent les consommateurs d'un produit, la procédure présentée en 4.2 est réutilisable pour dresser un tableau de bord résumant l'ensemble des avis émis sur un produit et, à partir de la liste des cibles, il ne reste plus qu'à extraire tous les marqueurs de polarité qui leurs ont été associés.

La méthode proposée permet d'extraire des cibles en fonction d'une liste constituée de manière semi-automatique. La principale perspective d'évolution vise à automatiser totalement le processus d'élaboration et d'enrichissement de la liste afin de faciliter le portage du système à un autre domaine ou à une autre langue.

Il serait possible en appliquant des méthodes de généralisation de remonter jusqu'au concept des cibles extraites. Dans le cadre du projet ImagiWeb, nous disposons à l'inverse de concepts de cibles avec des exemples et il nous faudrait, par annotations manuelles, en rechercher l'ensemble des marqueurs. La méthode présentée deviendrait encore plus intéressante dans la mesure où elle permettrait de pré annoter certains passages et limiter ainsi le travail des annotateurs.

Remerciements

Ce travail a été subventionné par l'ANR, Projet IMAGIWEB contrat n° 2012-CORD-002-05 et par le Pôle de Compétitivité SCS. Le corpus sur lequel ont porté les expériences a été mis à notre disposition par les fondateurs du Site Vodkaster. Nous tenons à les en remercier.

Références

BLAIR-GOLDENSOHN, S, HANNAN, K, MCDONALD, R, NEYLON, T, REIS, G, REYNAR, J. (2008) Building a sentiment summarizer for local service reviews. *In WWW Workshop on NLP*.

COLLOBERT R, BENGIO S. et MARIETHOZ J. (2002). *Torch: a modular machine learning software library*. In Technical Report IDIAP-RR02-46, IDIA

DUTREY, C, CLAVEL, C, ROSSET, S, VASILESCU, I, ADDA-DECKER, M, (2012). Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ? *In Actes de TALN12.*

LIU, B. (2012). Sentiment Analysis and Opinion Mining *A Comprehensive Introduction and Survey* Morgan & Claypool, May 2012, 167 pages.

NAVIGLI, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys*.

PANG, B. et LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrival*, 2:1–2.

PUPIER, P. (1998). Une première systématique des évaluatifs en français. *Revue québécoise de linguistique*, 26(1).

RILOFF, E, WIEBE, J. (2003). Learning extraction patterns for subjective expressions. *In EMNLP*.

TORRES-MORENO, J-M, EL-BEZE, M, BELLOT, P, BECHET F. (2011) Peut-on voir la détection d'opinions comme un problème de classification thématique ? in *Modèles statistiques pour l'accès à l'information textuelle* sous la direction de GAUSSIER, E, YVON, F, Hermes, 2011