

Construction d'un lexique affectif pour le français à partir de Twitter

Alexander Pak Patrick Paroubek
Université de Paris-Sud,
Laboratoire LIMSI-CNRS,
Bâtiment 508, F-91405 Orsay Cedex, France
alexpak@limsi.fr, pap@limsi.fr

Résumé. Un lexique affectif est un outil utile pour l'étude des émotions ainsi que pour la fouille d'opinion et l'analyse des sentiments. Un tel lexique contient des listes de mots annotés avec leurs évaluations émotionnelles. Il existe un certain nombre de lexiques affectifs pour la langue anglaise, espagnole, allemande, mais très peu pour le français. Un travail de longue haleine est nécessaire pour construire et enrichir un lexique affectif. Nous proposons d'utiliser Twitter, la plateforme la plus populaire de microblogging de nos jours, pour recueillir un corpus de textes émotionnels en français. En utilisant l'ensemble des données recueillies, nous avons estimé les normes affectives de chaque mot. Nous utilisons les données de la Norme Affective des Mots Anglais (ANEW, Affective Norms of English Words) que nous avons traduite en français afin de valider nos résultats. Les valeurs du coefficient τ de Kendall et du coefficient de corrélation de rang de Spearman montrent que nos scores estimés sont en accord avec les scores ANEW.

Abstract. Affective lexicons are a useful tool for emotion studies as well as for opinion mining and sentiment analysis. Such lexicons contain lists of words annotated with their emotional assessments. There exist a number of affective lexicons for English, Spanish, German and other languages. However, only a few of such resources are available for French. A lot of human efforts are needed to build and extend an affective lexicon. We propose to use Twitter, the most popular microblogging platform nowadays, to collect a dataset of emotional texts in French. Using the collected dataset, we estimated the affective norms of words present in our corpus. We used the dataset of Affective Norms of English Words (ANEW) that we translated into French to validate our results. Values of Kendall's τ coefficient and Spearman's rank correlation coefficient show that our estimated scores correlate well with the ANEW scores.

Mots-clés : Analyse de sentiments, ANEW, Twitter.

Keywords: Sentiment analysis, ANEW, Twitter.

1 Introduction

Les listes de mots annotés sont l'un des outils fréquemment utilisés dans l'analyse des sentiments pour détecter une humeur ou classer des textes en fonction des émotions qui y sont exprimées. Ces listes contiennent pour chaque mot son évaluation émotionnelle qui peut être représentée par un ensemble de scores numériques ou une balise (par exemple : « colère », « peur », etc.). ANEW (Affective Norms of English Words) est une liste de mots anglais construite par (Bradley & Lang, 1999). Elle a été souvent utilisée dans les travaux de recherche en fouille d'opinion et dans l'analyse de sentiment (Jung *et al.*, 2007) (Kim *et al.*, 2009).

Un ensemble de sujets humains a affecté pour chaque mot de la liste ANEW, des notes sur une échelle de 1 à 9 décrivant selon trois dimensions : « le plaisir », « l'excitation » et « la dominance ». Les scores finaux associés à chaque mots ont été obtenus en calculant la valeur moyenne et son écart-type pour chaque dimension émotionnelle¹. Dans nos travaux de recherche, nous cherchons seulement à estimer la valeur moyenne de la valence (ie. dimension du « plaisir »), car c'est cette valeur qui est à la base de la classification des textes en fonction des sentiments qu'ils expriment.

Mot en anglais	Mot traduit en français	Valence
happy	heureux	8.21
good	bien	7.47
bored	qui s'ennuie	2.95
angry	en colère	2.85
sad	triste	1.61

TABLE 1 – Des exemples de la liste ANEW traduite en français

Nous utilisons un ensemble de données constitué de messages recueillis depuis Twitter². Cette plateforme de microblogging contient un très grand nombre de messages très courts créés par les utilisateurs. Le contenu des messages peut varier d'un avis d'intérêt personnel à des communiqués publics.

Jansen *et al.* (Jansen *et al.*, 2009) ont identifiés le microblogging comme une source de « bouche à oreille » en ligne pour l'image de marque (*branding*). Dans leur article, les auteurs déclarent que le microblogging a le pouvoir de jouer un rôle important pour l'image de marque des grandes entreprises puisque les messages issus du microblogging peuvent faire ou défaire la réputation d'une organisation ou d'un produit.

Pour obtenir notre matériel d'entraînement, nous avons utilisé les émoticônes comme des indicateurs de l'humeur générale d'un message (Read, 2005).

Nous divisons les messages Twitter en deux ensembles :

1. “un ensemble positif” – formé de messages contenant des émoticônes positives, ie. ‘ :)’
2. “un ensemble négatif” – formé de messages contenant des émoticônes négatives, i.e. ‘ :(’

Ensuite, nous comptons le nombre d'apparition de chaque mot issu des données ANEW et nous estimons le score de valence en fonction du nombre respectif d'occurrences du mot dans chaque ensemble.

1. ANEW contient en réalité trois ensembles de données : la moyenne des scores sur l'ensemble des sujets de sexe masculin, la moyenne des scores sur l'ensemble des sujets de sexe féminin, et l'ensemble de données combinées. Cependant, dans nos travaux de recherche, nous nous intéressons seulement au dernier.

2. <http://twitter.com>

Notre hypothèse de base est qu'un mot a un score de valence élevé s'il apparaît fréquemment dans l'ensemble positif et en même temps, rarement dans l'ensemble négatif. Au contraire, un mot qui apparaît plus souvent dans l'ensemble négatif que dans l'ensemble positif, devrait avoir un indice de valence faible.

Nous avons traduit la liste ANEW en français et l'avons utilisée comme donnée de référence pour valider nos résultats, car cette liste a été produite par des humains. Nous prenons en considération le fait qu'un mot en anglais peut avoir une force émotionnelle différente de celle de sa traduction en français. Toutefois, nous faisons l'hypothèse qu'une corrélation forte entre les scores de la liste ANEW traduite et nos scores estimés, valide notre estimation de valence pour le français.

2 Travaux antérieurs

Il ya un certain nombre de lexiques annotés disponibles pour l'analyse des sentiments en anglais, comme ANEW (Bradley & Lang, 1999), WordNet-Affect (Strapparava & A, 2004), Balanced Affective Liste Word (Siegle, 1994). Des versions régionalisées d'ANEW existent dans d'autres langues comme l'espagnol (Redondo *et al.*, 2007) et l'allemand (Võ *et al.*, 2009), mais il n'y a pas de tels ensembles de données pour le français. Une régionalisation de ce type de données nécessiterait beaucoup d'efforts humains.

Dans (Mathieu, 2006), l'auteur construit un lexique sémantique des verbes français décrivant les sentiments, les émotions et les états psychologiques. Le lexique contient 600 verbes qui sont divisés en 33 classes sémantiques. Toutes les catégories sont regroupées en 3 catégories : positif, négatif et neutre. Le lexique contient également des liens entre les classes. Ces liens sont le « sens », « l'intensité » et « l'antonymie ». Le document présente aussi le logiciel FEELING qui utilise le lexique construit pour l'interprétation des émotions dans les expressions données. Bien que le lexique présenté semble être un outil précieux pour l'analyse des sentiments, il faudrait beaucoup de ressources humaines pour l'étendre, par exemple à l'annotation des adjectifs ou à d'autres verbes.

Dans (Harb *et al.*, 2008), les auteurs proposent une méthode automatique pour construire une liste d'adjectifs positifs et négatifs pour n'importe quel domaine. L'algorithme proposé utilise le moteur de recherche de Google Blog pour collecter un corpus de textes émotionnels. Les auteurs ont commencé par deux types d'adjectifs germes : positifs (comme la « bonne », « agréable », « excellent », etc) et négatif (« mauvais », « méchant », etc.) Ils ont utilisé les adjectifs germes pour effectuer des requêtes sur le moteur de recherche de Google Blog et pour récupérer deux séries de messages blog : une positive et une négative. A partir des ensembles de données récupérés, les auteurs ont extrait les adjectifs et déterminé leur polarité en regardant l'occurrence d'un mot au voisinage des mots germes. Dans leur approche, les auteurs ne se sont concentrés que sur les adjectifs, alors que dans nos travaux de recherches, nous avons construit une liste qui contient aussi des verbes et des noms.

3 Notre méthode

Nous avons utilisé l'API de Twitter pour télécharger et filtrer les messages. L'API permet de spécifier la langue des messages et une requête de recherche. Comme critère de recherche nous utilisons une série d'émoticônes positives (telles que : ' :)', ' :-)', ' :-D', etc) pour récupérer des messages positifs, et une série d'émoticônes négatives (' :(, ' :-(, etc) pour récupérer des messages négatifs. Nous avons collecté

un total de 300000 messages répartis de manière égale entre les ensembles positifs et négatifs.

Parce qu’aucun message ne peut pas dépasser 140 caractères d’après les règles de la plate-forme de microblogging, il est généralement composé d’une seule phrase. Par conséquent, nous supposons que l’émotivité dans un message représente une émotion pour tout le message et que tous les mots du message sont liés à cette émotion. Ainsi, si un mot est plus fréquent dans l’un des ensembles, il est lié plus fortement à l’émotion correspondante.

Pour chaque mot w de la liste ANEW, étant donné le nombre de ses occurrences dans l’ensemble positif $N(w, M^+)$ et dans l’ensemble négatif $N(w, M^-)$, nous estimons la valeur de la polarité d’un mot comme suit :

$$valence^*(w) = \frac{9 \cdot N(w, M^+) + N(w, M^-)}{N(w, M^+) + N(w, M^-)} \quad (1)$$

qui est égale à :

$$valence^*(w) = 8 \cdot \frac{N(w, M^+)}{N(w, M^+) + N(w, M^-)} + 1 = 8 \cdot P(M^+|w) + 1 \quad (2)$$

où $P(M^+|w)$ est la probabilité d’un mot w d’être positif, 8 est un facteur d’échelle, et 1 est une translation initiale. Cependant, nous ne considérons pas les occurrences de mots dans des constructions de négation, tels que : “pas bon”, “n’aime pas”.

Nous avons mené des expériences avec trois listes différentes afin de vérifier laquelle influence plus la qualité de l’estimation :

1. « Tous les mots », c’est-à-dire tous les mots de la liste ANEW.
2. « Adjectifs », seuls les adjectifs de la liste ANEW, car les adjectifs sont considérés comme des mots représentatifs de l’opinion (Voll & Taboada, 2007). Nous avons utilisé TreeTagger (Schmid, 1994) pour étiqueter les mots avec leur partie du discours.
3. « Mots filtrés », les mots de la liste ANEW qui apparaissent au moins 100 fois dans le corpus que nous avons recueilli.

Pour valider notre estimation, nous avons calculé l’erreur quadratique moyenne :

$$MSE = \frac{1}{|W|} \sum_{w \in W} (valence(w) - valence^*(w))^2 \quad (3)$$

où W est la liste de tous les mots et $valence(w)$ est le score de valence à partir de zéro de la liste ANEW.

Une autre façon de valider la valeur estimée est de vérifier la concordance des classements des deux mots : le classement en fonction des valeurs de la polarité dans ANEW et le classement selon les valeurs estimées de la polarité. Deux scores ont été calculés pour mesurer la corrélation : le coefficient τ de Kendall (Kendall, 1938) et coefficient de corrélation de rang de Spearman (Maritz, 1981).

Le τ de Kendall et le coefficient de Spearman prennent leurs valeurs entre -1 et 1 : $\tau = -1$ indique le désaccord parfait entre deux classements ; $\tau = 1$ signifie l’accord parfait.

4 Les expérimentations et les résultats

Nous avons présenté la corrélation entre les valeurs de la polarité estimées et les valeurs ANEW sur les figures 1-3, où chaque point représente un mot, les abscisses x correspondent à un classement en

fonction de la valence définie par ANEW, les ordonnées y correspondent à un classement en fonction de la valence estimée. Un alignement des points le long de la ligne $y = x$ montre une corrélation entre les deux classements.

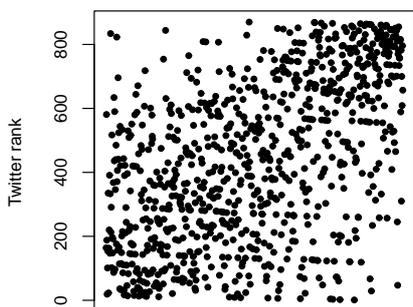


FIGURE 1 – Tous les mots

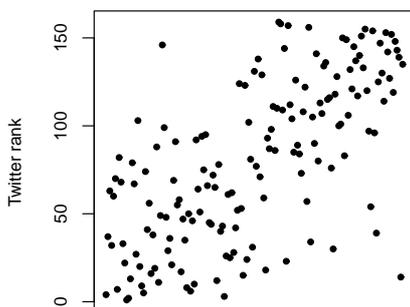


FIGURE 2 – Adjectifs

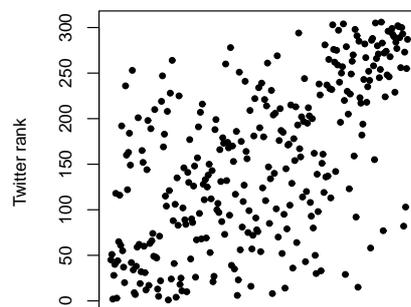


FIGURE 3 – Mots filtrés

Les coefficients de corrélation et l’erreur quadratique sont présentés dans le Tableau 2. Le coefficient τ de Kendall et le coefficient de Spearman montrent un bon accord entre les deux classements. Nous pouvons observer une valeur assez importante de l’erreur quadratique moyenne, mais nous pensons qu’elle peut facilement être réduite en utilisant des techniques d’apprentissage automatique ou d’analyse de régression ; en conséquence nous considérons que les deux coefficients de corrélation que nous avons mesurés sont des mesures d’évaluation appropriées à nos travaux.

Score	τ de Kendall	Coef. de Spearman	Erreur
Tous les mots	0.287	0.376	6.184
Adjectifs	0.384	0.489	4.836
Mots filtrés	0.448	0.543	3.174

TABLE 2 – Les coefficients de corrélation et l’erreur quadratique

Comme nous pouvons l’observer sur les graphes et les résultats numériques, la corrélation s’améliore si l’on ne sélectionne que les adjectifs. Ce fait confirme l’hypothèse que les adjectifs sont des mots représentatifs des opinions. Nous pouvons aussi observer qu’il suffit d’avoir assez de données pour estimer la valeur de polarité. La corrélation est plus élevée si l’on considère uniquement des mots qui apparaissent au moins 100 fois dans le corpus.

5 Conclusion

Dans nos travaux de recherche, nous avons proposé une méthode de construction automatique de lexique affectif des mots français. Nous avons proposé d’utiliser Twitter pour recueillir un ensemble de données de textes émotionnels en s’appuyant sur des émoticônes comme indicateurs de polarité, que nous utilisons pour estimer les scores de valence des mots les plus fréquents. Le lexique construit peut être utilisé dans les études de l’émotion, l’analyse des sentiments et les systèmes de fouille de données.

Nous avons comparé nos résultats estimés avec l’ensemble de données ANEW traduites en français. Les résultats expérimentaux ont montré une bonne corrélation entre le classement de mots en fonction de nos

valeurs estimées et celles d'ANEW. Notre méthode peut être appliquée à n'importe quelle langue pourvu que la quantité de messages soient suffisants.

De cette façon, nous avons montré que des ensembles de données similaires à ANEW peuvent être développés pour d'autres langues que l'anglais. En outre, notre méthode est automatique, évitant ainsi l'effort humain associé à création d'un lexique affectif pour une nouvelle langue. Elle pourrait être utilisée pour étendre la liste ANEW avec un coût quasi-nul.

Dans le cadre de nos travaux futurs, nous prévoyons d'expérimenter notre méthode sur d'autres langues et d'utiliser des techniques d'apprentissage automatique pour une estimation des scores de valence plus précise. Nous prévoyons également de mettre davantage l'accent sur le traitement des négations.

Références

- BRADLEY M. M. & LANG P. J. (1999). Affective norms for English words (ANEW). Gainesville, FL. The NIMH Center for the Study of Emotion and Attention. *University of Florida*.
- HARB A., DRAY G., PLANTIÉ M., PONCELET P., ROCHE M. & TROUSSET F. (2008). Détection d'Opinion : Apprenons les bons Adjectifs ! volume 8, p. 59–66.
- JANSEN B. J., ZHANG M., SOBEL K. & CHOWDURY A. (2009). Micro-blogging as online word of mouth branding. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, p. 3859.
- JUNG Y., CHOI Y. & MYAENG S.-H. (2007). Determining Mood for a Blog by Combining Multiple Sources of Evidence. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, p. 271–274.
- KENDALL M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1/2), 81–93.
- KIM E., GILBERT S., EDWARDS M. J. & GRAEFF E. (2009). Detecting Sadness in 140 Characters : Sentiment Analysis and Mourning Michael Jackson on Twitter. *Web Ecology*, **03**(August).
- MARITZ J. (1981). Distribution-free statistical methods. p. 217.
- MATHIEU Y. Y. (2006). A Computational Semantic Lexicon of French Verbs of Emotion.
- READ J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *In ACL. The Association for Computer Linguistics*.
- REDONDO J., FRAGA I., PADRÓN I. & COMESAÑA M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior research methods*, **39**(3), 600–5.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing. September 1994*.
- SIEGLE G. (1994). <http://www.sci.sdsu.edu/cal/wordlist/origwordlist.html>.
- STRAPPARAVA C. V. & A (2004). WordNet-Affect : an affective extension of WordNet. *In Proceedings of the 4th International Conference on Language Resources and Evaluation , LREC*.
- VOLL K. & TABOADA M. (2007). Not all words are created equal : Extracting semantic orientation as a function of adjective relevance. *pages 337-346. Volume 4830/2007 AI 2007 : Advances in Artificial In.*
- VÕ M. L.-H., CONRAD M., KUCHINKE L., K. URTON, HOFMANN M. J. & JACOBS A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior research methods*, **41**(2), 534–538.