Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments

Egle Eensoo Mathieu Valette
INALCO, ERTIM, 2 rue de Lille, 75343 Paris Cedex 07
{prenom.nom}@inalco.fr

RÉSUMÉ

Depuis une dizaine d'années, le TAL s'intéresse à la subjectivité, notamment dans la perspective d'applications telles que la fouille d'opinion et l'analyse des sentiments. Or, la linguistique de corpus outillée par des méthodes textométriques a souvent abordé la question de la subjectivité dans les textes. Notre objectif est de montrer d'une part, ce que pourrait apporter à l'analyse des sentiments l'analyse textométrique et d'autre part, comment mutualiser les avantages d'une association entre celle-ci et une méthode de classification automatique basée sur l'apprentissage supervisé. En nous appuyant sur un corpus de témoignages issus de forums de discussion, nous montrerons que la prise en compte de critères sélectionnés suivant une analyse textométrique permet d'obtenir des résultats de classification satisfaisants par rapport à une vision purement lexicale.

ABSTRACT

About the application of textometric methods for developing classification criteria in Sentiment analysis

Over the last ten years, NLP has contributed to applied research on subjectivity, especially in applications such as Opinion mining and Sentiment analysis. However, corpus linguistics and textometry have often addressed the issue of subjectivity in text. Our purpose is to show, first, what textometric analysis could bring to sentiment analysis, and second, the benefits of pooling linguistic/textometric analysis and automatic classification methods based on supervised learning. By processing a corpus of posts from fora, we will show that the building of criteria from a textometric analysis could improve classification results, compared to a purely lexical approach.

MOTS-CLÉS : linguistique de corpus, textométrie, analyse de sentiments, classification automatique supervisée.

KEYWORDS : corpus linguistics, textometry, sentiment analysis, supervised learning.

1 Introduction

L'extraction d'information subjective (Pang et Lee, 2008) est depuis une dizaine d'années un vaste domaine d'applications en croissance régulière. Malgré quelques travaux (par exemple Vernier, 2009 ; Béchet *et al.*, 2008) le savoir-faire linguistique y est peu sollicité. La subjectivité a pourtant fait l'objet de nombreux travaux linguistiques, dans différents courants théoriques – linguistique de l'énonciation, analyse de discours, sémantique des textes. La textométrie, aux confins de la linguistique générale et du TAL, propose par ailleurs une archive intéressante de travaux sur corpus susceptibles d'intéresser les

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 367–374,
Grenoble, 4 au 8 juin 2012.

applications d'analyse des sentiments (AS). On songe, dans le discours politique, aux travaux de (Salem, 1993), dans les sondages d'opinion, à (Lebart et Salem, 1988) ou dans la littérature, à (Brunet, 2009).

On souhaiterait ici susciter une rencontre entre d'une part le TAL ingénierique et ses applications et, d'autre part, la textométrie, à partir des constats suivants : (1) l'évaluation des méthodes en TAL repose sur un ensemble restreint de mesures (telles que précision, rappel, f-mesure) qui ont pour but de vérifier la qualité des méthodes plus que de valider des hypothèses et des méthodologies linguistiques. Leurs résultats ne nécessitent pas d'interprétation pour être valides ; (2) la textométrie relève, au contraire, d'une tradition descriptive. Elle se focalise sur l'interprétation des résultats de traitements statistiques, davantage que sur l'amélioration desdits traitements. À la différence du TAL, l'évaluation n'est pas un enjeu en textométrie[1].

Notre projet repose sur l'hypothèse que la textométrie, discipline descriptive, est à même d'apporter des solutions méthodologiques pour les applications généralement dévolues au TAL. Nous tenterons d'évaluer l'apport potentiel de la conjonction d'une analyse textométrique et de méthodes d'apprentissage pour une application d'AS.

2 État de l'art

La catégorisation des textes, qu'elle soit bipolaire (positif/négatif) ou multiclasse (mauvais/bon/excellent), est l'application principale en extraction d'information subjective. Elle peut être réalisée au moyen d'algorithmes *ad hoc* (Turney, 2002 ; Snyder et Barzilay, 2007) ou des méthodes d'apprentissage comme Naive Bayes, Support Vector Machines, etc. (Pang *et al.*, 2002 ; Mihalcea et Liu, 2006), en utilisant des attributs différents pour caractériser les documents. Même si perdurent d'autres méthodes – ayant principalement recours à l'utilisation de ressources lexicales, construites (Kim et Hovy, 2004) ou automatiquement acquises (Turney, 2002 ; Riloff *et al.*, 2003), avec la banalisation des corpus annotés, les méthodes de catégorisation basées sur l'apprentissage supervisé sont de plus en plus utilisées. Elles utilisent diverses caractéristiques textuelles : (i) tous les mots du texte (sac de mots, unigrammes ou n-grammes) (Pang *et al.*, 2002 ; Dave *et al.* 2003) ; (ii) la présence ou l'absence d'un ensemble de mots déterminés ; (iii) l'emplacement de certains mots (Kim et Hovy, 2006) ; (iv) certaines parties du discours seules : adjectifs (Kamps et Marx, 2002), collocations adverbe-adjectif (Turney, 2002) ; substantifs ; enfin (v) les dépendances syntaxiques (Nakagawa *et al.*, 2010 ; Wi *et al.*, 2009 ; Wiegand et Klakow, 2010). Nous nous inscrivons donc pleinement dans cette démarche en proposant des critères de classification issus d'analyses textométriques pour servir de base aux divers algorithmes d'apprentissage supervisé.

[1] Autrement dit, les études textométriques ne sont validées que par l'assentiment d'une communauté qui, dans le meilleur des cas, est distante (par exemple, critique littéraire, sociologie), mais, dans le pire des cas, n'est peut-être qu'un avatar du *jugement d'acceptabilité* pourtant honni de ladite communauté.

3 Présentation du corpus

3.1 Contexte applicatif de l'étude

Le corpus est constitué de 300 textes courts réunis par SAMESTORY (http://www.same-story.com), un service d'agrégation d'ego-documents. Il s'agit, en l'occurrence, de témoignages et récits d'histoires vécues postés par les internautes sur différents forums de discussion (aufeminin.com, doctissimo.fr, etc.). Les catégorisations sont multicritères : thématiques, tonalité, conseil *vs* demande, sexe de l'émetteur, situation familiale, etc. Nous traitons, dans des textes portant sur la santé, la tonalité « gai/triste ». De prime abord, elle s'apparente à une analyse thymique, mais il s'agit de catégories complexes où les phénomènes discursifs (ex : structure du récit) interviennent dans la classification autant que l'expression linguistique des sentiments. Ainsi, notre tâche est de modéliser l'art de témoigner d'une histoire vécue.

3.2 Annotation tonale du corpus

L'annotation tonale du corpus a été effectuée par SAMESTORY. Nous en avons analysé un échantillon pour en déduire la stratégie d'annotation de façon à caractériser plus finement l'opposition binaire gai/triste. Un témoignage « triste » est (i) une histoire qui finit mal, (ii) un témoignage exprimant des doutes, des interrogations, ou sollicitant de l'aide. Un témoignage « gai » est (i) une histoire triste qui finit bien, (ii) un témoignage modulant la gravité de la situation et soulignant les points positifs (iii) un conseil.

4 Description de l'expérience

4.1 Étape 1 : Choix des caractéristiques textuelles au moyen des méthodes textométriques

Nous tentons de mettre en évidence les phénomènes textuels qui différencient les témoignages de nos deux catégories. Nous avons une double ambition : trouver des critères de classification linguistiquement explicables et suffisamment robustes pour servir de critères aux méthodes d'apprentissage supervisé. Nous faisons l'hypothèse que les critères de classification *interprétables* sont plus performants que les critères trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage (ex : présence de fautes d'orthographe non pertinentes par rapport aux catégories de classification). Ainsi, lors de l'étape de sélection de critères, l'analyste écarte les critères liés à l'échantillon du corpus et choisit les critères textuels cohérents avec les composantes textuelles (thématique, dialogique, etc. *cf.* § 5) actualisées dans le corpus. Pour l'expérience, nous avons utilisé trois types de critères : (i) critères unitaires : un choix de formes, lemmes ou catégories morphosyntaxiques ; (ii) critères composites adjacents (n-grammes) ; (iii) cooccurrences multiniveaux (combinant les éléments de différents niveaux de description linguistique : formes, lemmes ou catégories morphosyntaxiques). Tous les critères sont sélectionnés selon 4 principes : leur caractère spécifique à un sous-corpus, leur répartition uniforme dans le sous-corpus, leur fréquence et leur pertinence linguistique.

L'analyse du corpus et l'extraction des critères ont été effectuées avec deux logiciels

textométriques – Lexico3 (Salem *et al.* 2003) et TXM (Heiden *et al.* 2010) – qui implémentent les algorithmes de spécificités (Lafon, 1980) et de cooccurrences (Lafon, 1981). Nous avons choisi les deux premiers types de critères selon le procédé suivant :

1. calcul des spécificités des items isolés (formes, lemmes et catégories morphosyntaxiques) et de leur n-grammes (fonction « Segments Répétés » de Lexico 3) pour chaque sous-corpus (gai/triste) ;
2. analyse des contextes d'apparition des items spécifiques (au moyen de concordances textuelles) afin de s'assurer de leur pertinence textuelle et de l'unicité de leur fonction (les critères ayant une seule fonction et signification ont été privilégiés) ;
3. vérification de la répartition uniforme des items dans le sous-corpus (fonctionnalité « Carte de Sections » du Lexico 3) ;

La sélection des cooccurrences s'est fait comme suit :

1. calcul des cooccurrences (fonction « Cooccurrences » de TXM) des items spécifiques fréquents et uniformément repartis sur la totalité du corpus ;
2. analyse des contextes d'apparition de ces cooccurrences ;
3. sélection des cooccurrences spécifiques à un sous-corpus ;

Dans les deux cas, les critères de classification pour chaque texte sont des fréquences ou des valeurs booléennes (présence/absence) des items sélectionnés.

4.2 Étape 2 : Classification

La deuxième étape consiste à utiliser des algorithmes d'apprentissage supervisé pour classer les textes. En utilisant Weka[2], nous en avons expérimenté trois, chacun d'une famille différente : les arbres de décision (J48 ; Quinlan, 1993), Naive Bayes (John et Langley, 1995) et les Machines à Vecteurs de Support (SMO ; Platt, 1998). L'objectif est d'observer les différences et similitudes au niveau des performances en changeant la nature et la quantité des critères.

Le corpus contient 300 textes équitablement répartis entres les deux catégories (147 « gaies » et 153 « tristes »). L'évaluation a été effectuée avec la méthode de validation croisée sur cinq parties.

- *Expérimentation 1.1* : première expérimentation avec des mots simples sans aucune modification (avec pour valeur leur fréquence dans un texte) ; on considère ces résultats comme la base de comparaison (*baseline*) pour d'autres expérimentations. La base de comparaison est donc l'expérimentation qui nécessite l'effort computationnel minimal sur les textes en considérant ces derniers comme un matériau brut, directement accessible (au moyen d'une segmentation en mots). Toutes les autres expérimentations effectuent des traitements supplémentaires sur les textes visant à améliorer les résultats. L'évaluation a été effectuée avec la validation croisée sur 5 parties du corpus.

[2] http://www.cs.waikato.ac.nz/ml/weka/

- *Expérimentation 1.2* : A la place des mots, nous avons utilisés leurs lemmes (casse normalisée).
- *Expérimentation 1.3* : Utilisation des n-grammes de mots (longueur maximale 3).

Dans la série des expérimentations 2, nous avons utilisé les critères élaborés selon la méthodologie décrite dans la partie précédente.

- *Expérimentation 2.1* : Utilisation de critères unitaires et de critères composites adjacents pour un total de 30 critères.
- *Expérimentation 2.2* : Ajout de critères cooccurrentiels et augmentation du nombre (total : 46 critères).
- *Expérimentation 2.3* : Augmentation du nombre de critères (total : 61 critères).

5 Résultat et discussion

Type d'attributs	**Algorithme de classification**	**% des textes bien catégorisés**
1.1. Mots simples (1200 critères)	J48	53
	Naive Bayes	63
	SMO	70
1.2. Lemmes (370 critères)	J48	55
	Naive Bayes	63
	SMO	64
1.3 N-grammes de mots (1357 critères)	J48	56
	Naive Bayes	64
	SMO	74
2.1. Critères textométriques (30 critères)	J48	67
	Naive Bayes	64
	SMO	65
2.2. Critères textométriques (43 critères)	J48	62
	Naive Bayes	72
	SMO	72
2.3. Critères textométriques (61 critères)	**J48**	**70**
	Naive Bayes	**74**
	SMO	**77**

TABLE 1 – Résultat des expérimentations

Comme dans des expériences similaires (Pang *et al.*, 2002), on constate que la

classification sur les mots simples et les n-grammes permet d'obtenir des résultats convenables compte tenu de la difficulté de la tâche. Néanmoins, cela constitue un plafond que l'on ne peut dépasser. La généralisation des critères apportée par la lemmatisation ne permet pas d'améliorer les résultats. Ce phénomène a fait l'objet de nombreux débats dans la communauté textométrique (par exemple Mellet, 2003).

A la différence de la première série d'expérimentations, nos critères textométriques sont peu nombreux mais ils constituent une base facilement extensible. L'ajout des critères augmente systématiquement les performances de Naive Bayes et SMO. Ainsi, nous observons une progression sensible sur l'ensemble des algorithmes. Notre meilleur résultat (avec SMO) dépasse de 7 points celui obtenu avec des mots simples et de 3 points celui des n-grammes. Par ailleurs, l'amélioration des résultats pour J48 et Naive Bayes est systématique.

L'interprétation des résultats chiffrés et des critères obtenus participe selon nous de la validation de l'expérimentation et en constitue une valeur ajoutée. Ainsi, nous avons organisé nos critères selon une typologie inspirée de travaux sémiotiques. Les critères thymiques (Courtès, 1991), qui relèvent d'une lecture axiologique des textes, sont essentiellement dysphoriques et concernent donc les textes tristes : *« avoir peur »*, *« je souffre »*, *« douleur »*, *« stress »*. Le seul critère thymique retenu pour la classification des textes gai est *« heureux »* (euphorique). Au-delà des critères thymiques courants, nous nous sommes intéressés à ceux relatifs à des *composantes textuelles* (Rastier, 2001) parce que, ne relevant pas de typologies axiologiques classiques (positif/négatif) (Charaudeau, 1992), ils sont rarement pris en compte en AS. La composante *dialectique* concerne l'organisation linéaire et temporelle du récit. Ces critères, dans les textes « gais », sont différents marqueurs de structuration argumentative (*« par contre »*, *« car »*) et temporelle (*« après »*, *« puis »*) absents des textes « tristes ». Dans ceux-ci, la structuration est cumulative (*« en plus »*, *« de nouveau »*) ou indice d'incertitude (*« ne pas arriver à »*, *« avoir l'impression de »*). La composante *dialogique* est relative au positionnement des acteurs. Elle met en œuvre un fort contraste entre les textes « gais », où le destinateur-énonciateur s'adresse explicitement à un *« tu »* destinataire actualisé par des pronoms de 2ème personne (pronoms personnel, possessifs, etc.), relate une expérience édifiante (*« mon expérience »*, *« pour ma part »*) et prodigue des conseils (présence d'hyperliens *« www »*) et des encouragements (*« bon courage »*) sans pour autant mettre en avant un *je*. Les témoignages « tristes » mettent en texte un *« je »* massif. Enfin, la composante *thématique* n'a pas été négligée mais nous nous sommes efforcés de ne sélectionner que des critères d'un grand niveau de généralité relatifs au domaine de la santé. Ainsi, aux noms de symptômes, maladies, traitements ou médicaments, nous avons préféré, pour les textes « tristes » : *« urgences »*, *« hôpital »*, *« rendez-vous »*, *« analyses »*, *« médecins »*, ou la locution *« être atteint de »*. Pour les textes « gais » : *« rémission »*, *« produit naturel »*, *« homéopathie »* permettent d'obtenir des résultats convaincants.

6 Conclusion

Il est admis que les méthodes efficaces en classification thématique (par exemple, l'apprentissage supervisé sur mots simples) sont peu performantes pour les tâches d'analyse de la subjectivité. La difficulté réside dans le fait que la subjectivité ne relève pas seulement du lexique, mais d'autres niveaux de description : organisation temporelle

du récit, structure argumentative, etc. Nous avons proposé ici quelques éléments d'analyse pour la prise en compte de ces niveaux de description et leur implémentation pour la classification. Le coût en temps de notre méthode d'élaboration de critères n'a pas été quantifié mais nous estimons qu'il est comparable à d'autres méthodes semi-automatiques. Le domaine manquant de méthodes éprouvées, notre expérience nous a permis de mieux comprendre la tâche et sa complexité et d'esquisser une proposition méthodologique tenant compte d'une caractérisation textuelle de la subjectivité.

7 Références

BRUNET, É. (2009). *Écrits choisis,* Volume 1, *Comptes d'auteurs. Études statistiques. De Rabelais à Gracq*. Textes édités par D. Mayaffre, Champion, Paris

BÉCHET, F., EL-BÈZE, M. et TORRES-MORENO, J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes *Actes de l'atelier de clôture de la 4ème édition du Défi Fouille de Texte*.

CHARAUDEAU P. (1992). *Grammaire du sens et de l'expression*. Hachette Education.

COURTÈS, J. (1991). *Analyse sémiotique du discours. De l'énoncé à l'énonciation*, Paris, Hachette.

DAVE, K., LAWRENCE, S., et PENNOCK, D.M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international WWW conference*, May 20-24, 2003, Budapest, Hungary, pages 519-528.

HEIDEN, S., MAGUÉ, J-P. et PINCEMIN, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.),*JADT 2010*, Vol. 2, pages 1021-1032. [logiciel disponible sur http://textometrie.ens-lyon.fr/]

JOHN, G. H. et LANGLEY, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, pages 338-345.

KAMPS, J. et MARX, M. (2002). Words with Attitude. *1st International WordNet Conference*, pages 332-341.

KIM, S-M. et HOVY, E. (2004). Determining the sentiment of opinions.*Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg*, PA, USA.

KIM, S.-M. et HOVY, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics*, pages 1-8.

LEBART L. et SALEM A., (1988). *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*, Paris, Dunod.

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots,* 1, pages 127-165.

LAFON, P. (1981). Analyse lexicométrique et recherche des cooccurrences, *Mots, 3*, pages 95-148.

MELLET, S. (2003). Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica : Autour de la lemmatisation*, Dominique Labbé, éd.

MIHALCEA, R. et LIU, H. A (2006). Corpus-Based Approach to Finding Happiness AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW).

NAKAGAWA, T., INUI, K. et KUROHASHI, S. (2010). Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. *Proceedings of Human Language Technologies.*

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

PANG, B. et LEE, L. (2008). Opinion Mining and Sentiment Analysis, Now Publishers Inc.

PLATT, J. (1998). Machines using Sequential Minimal Optimization. B. Schoelkopf, C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning.*

RASTIER, F. (2001). *Arts et sciences du texte*, Paris, PUF.

RILOFF, E., WIEBE, J. et WILSON (2003). T. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 25-32.

QUINLAN, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

SALEM, R. (1993). Méthodes de la statistique textuelle, Thèse pour le doctorat d'État ès lettres et sciences humaines, Université de la Sorbonne Nouvelle – Paris 3, 998 pages.

SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., *et al.* (2003). Lexico3 – Outils de statistique textuelle. Manuel d'utilisation. http://www.tal.univ-paris3.fr/lexico/]

SNYDER, B. et BARZILAY, R. (2007). Multiple aspect ranking using the Good Grief algorithm. *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pages 300-307.

TURNEY, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417-424.

VERNIER, M., MONCEAUX, l. et DAILLE, B. (2009). DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes.*

WI, Y., ZHANG, Q., Huang, X., et Wu, L. (2009). Phrase Dependency Parsing for Opinion Mining. *Proceedings of EMNLP-2009*, Singapore.

WIEBE, J.M., WILSON, T., BRUCE, R., BELL, M. et MARTIN, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3), pages 277-308.

WIEGAND, M. et KLAKOW, D. (2010). Convolution Kernels for Opinion Holder Extraction. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, L.A., CA*.