Un système de vote pour la classification de textes d'opinion

Michel Plantié*, Mathieu Roche**, Gérard Dray*

* LGI2P, Ecole des Mines d'Alès, Site EERIE
(michel.plantie, gerard.dray)@ema.fr
** LIRMM, UMR 5506, Univ. Montpellier 2, CNRS,
mathieu.roche@lirmm.fr

Résumé : Les tâches de classification textuelle ont souvent pour objectif de regrouper thématiquement différents textes. Dans cet article, nous nous sommes intéressés à la classification de documents en fonction des opinions et jugements de valeurs qu'ils contiennent. L'approche proposée est fondée sur un système de vote utilisant plusieurs méthodes de classification.

1 Introduction

La classification de textes a pour objectif le regroupement de documents selon différents critères. Dans les travaux présentés dans cet article, nous nous intéressons à la classification de textes d'opinion qui consiste à classer les textes selon un jugement tel que l'aspect positif ou négatif d'une critique, l'aspect favorable ou défavorable donné par un expert, etc. Nous proposons dans cet article une approche fondée sur plusieurs classifieurs combinés à un système de vote. Dans un premier temps, nous présentons les corpus du défi DEFT'07(Grouin *et al.*, 2007) sur lesquels nous avons mené nos expérimentations ainsi que les représentations des textes utilisées. La section 3 décrit les classifieurs et les systèmes de vote proposés. Enfin, la partie 4 présente les résultats obtenus.

2 Représentation des données textuelles

La troisième édition du défi francophone DEFT'07 consistait à déterminer des catégories de jugements à partir de quatre corpus français très différents en terme de thème, taille, tournures de phrases, richesse du vocabulaire, représentation des catégories de jugement :

- ✔ **Corpus 1 :** Critiques de films, livres, spectacles et bandes dessinées. Trois catégories : bon, moyen, mauvais.
- ✔ **Corpus 2 :** Critiques de jeux vidéo. Trois catégories : bon, moyen, mauvais.
- ✔ **Corpus 3 :** Commentaires de révision d'articles de conférences scientifiques. Trois catégories : acceptation, acceptation sous conditions, rejet.

- ✔ **Corpus 4 :** Interventions des parlementaires et du gouvernement dans les débats sur les projets de lois. Deux catégories : pour, contre.

La première étape de notre approche consiste à appliquer un certain nombre de prétraitements linguistiques. Ceux-ci consistent à extraire du corpus toutes les unités linguistiques (mots lemmatisés ou lemmes) utilisées pour la représentation des textes. En effet, le prétraitement consistant à lemmatiser les données textuelles a globalement tendance à améliorer les tâches de classification (Plantié, 2006). Par ailleurs, ces prétraitements consistent également à éliminer certains mots ayant des types grammaticaux peu discriminants pour classifier les textes d'opinion : articles, ponctuations. Dans notre approche nous avons souhaité conserver les lemmes associés à tous les autres types grammaticaux. En effet, le fait de traiter spécifiquement des textes d'opinion nous encourage à conserver un maximum de types grammaticaux susceptibles d'exprimer des nuances d'opinions ou des contributions à des opinions (comme les adverbes). Les expériences que nous avons menées sur les textes d'opinion propres aux corpus DEFT'07 ont en effet montré que la suppression de types grammaticaux diminuait les performances. Dans la suite de cet article, nous appellerons « index » la liste de lemmes constitués pour chacun des corpus.

Chaque corpus est représenté sous forme matricielle en utilisant l'approche classique dite de Salton (Salton *et al.*, 1975). Dans cette représentation, les lignes sont associées aux différents textes du corpus et les colonnes sont relatives aux lemmes. Chaque cellule de la matrice représente le nombre d'occurrences du lemme dans chaque texte du corpus.

L'ensemble des textes d'un corpus et donc les vecteurs associés constituent dans notre approche l'ensemble d'apprentissage qui permet d'identifier un classifieur associé. L'espace vectoriel défini par l'ensemble des lemmes du corpus d'apprentissage et dans lequel sont définis ces vecteurs comporte un nombre important de dimensions. Nous avons choisi d'effectuer une réduction de l'index. Nous utilisons la méthode présentée par Cover qui mesure l'information mutuelle associée à chaque dimension de l'espace vectoriel (Cover & Thomas, 1991). Cette méthode expliquée en détail dans (Plantié, 2006) permet de mesurer l'interdépendance entre les mots et les catégories de classement des textes par la différence d'entropie entre celle de la catégorie et celle de la dimension en cours d'étude de l'espace vectoriel. Notons que plus la différence est grande, plus la quantité d'information de discrimination est importante et plus le mot est important pour la tâche de catégorisation. Dans notre approche, nous avons appliqué un seuil de zéro pour effectuer cette réduction. Un tel seuil signifie que les mots retenus sont réellement discriminants. Cette opération diminue de manière très importante l'ensemble des corpus avec un pourcentage de réduction de plus de 90% de tous les corpus de DEFT'07. Cette étape améliore de façon sensible les résultats de classification (environ 10% sur la « F-mesure » comme indiqué ci-après).

Dans le cadre de DEFT'07, nous avons appliqué un prétraitement supplémentaire. Ainsi, avant d'effectuer la classification des textes, nous avons cherché à améliorer les traitements « linguistiques » des textes. Dans ce but, les termes (groupes de mots respectant des patrons syntaxiques spécifiques tels que « Nom Adjectif », « Adjectif Nom », etc.) ont été extraits avec EXIT (Roche, *et al.*, 2004). Ainsi, nous avons considéré la liste des termes extraits

comme l'index du corpus à partir duquel tous les textes ont été vectorisés. Puis la procédure classique a été appliquée : réduction d'index et classification. Le nombre de termes extraits peut se révéler assez faible pour certains textes ce qui réduit significativement la taille de l'index. Ceci met alors en défaut les méthodes statistiques qui ont été mises en place dans le cadre du défi. Par ailleurs, les termes sélectionnés après réduction d'index ne sont pas suffisamment significatifs pour représenter la diversité des textes. Ceci peut expliquer que notre approche uniquement fondée sur les termes dégrade nos résultats. Ainsi, nous proposons ci-dessous une méthode plus générale fondée sur l'utilisation de bigrammes de mots.

Dans l'approche développée, outre les mots qui sont pris en compte, les vecteurs sont aussi constitués de bigrammes de mots. Seuls les bigrammes contenant des caractères spéciaux sont rejetés (caractères mathématiques, ponctuations, etc.). Cette représentation plus riche des textes permet d'obtenir des informations plus adaptées aux textes d'opinion. A titre d'exemple, dans le corpus de relectures d'articles les bigrammes tels que *pas convainquant*, *mieux motiver*, *pas assez* sont des groupes de mots beaucoup plus porteurs d'opinion comparativement à chacun des mots constituant ces bigrammes. Deux différences majeures sont à relever par rapport à la méthode fondée sur la terminologie : (1) L'approche prend en compte les mots ainsi que les bigrammes pour constituer l'index, (2) Le nombre de bigrammes retournés est beaucoup plus important que le nombre de termes respectant des patrons syntaxiques définis.

Les résultats en utilisant cette représentation enrichie des textes par la prise en compte des bigrammes améliore les résultats comme nous allons le montrer dans la section 4. Outre la qualité de la représentation des textes qui améliore les tâches de classification, la prise en compte de différents classifieurs (voir section suivante) reste déterminante pour retourner un résultat de bonne qualité.

3 Processus de classification

Afin d'améliorer la méthode générale de classification des textes d'opinion, nous avons mis en œuvre un système de vote fondé sur différents classifieurs. Le même processus de classification a été appliqué en nous appuyant sur les représentations de textes présentées dans la section précédente. Dans la suite de ce papier, nous appellerons *Copivote* (Classification de textes d'OPInion par un système de VOTE) le système de classification appliqué à la représentation par lemmes uniquement et *CopivoteBi* (Classification de textes d'OPInion par un système de VOTE avec Bigrammes) le système de classification appliqué à la représentation par lemmes et bigrammes.

Le système de vote mis en place s'appuie sur différentes approches de classification. Ces dernières sont fondées sur trois méthodes principales :

- *Vote à la majorité simple* : choix des classes à la majorité des résultats des classifieurs.
- *Vote par choix du maximum (respectivement, minimum)* : choix de la classe allouée par le classifieur qui a donné la probabilité la plus élevée (respectivement, faible).

d'appartenance. Dans ce cas, il y a nécessité que les probabilités exprimées par les différents classifieurs soient comparables.

– *Vote par somme pondérée* : pour chaque document et pour chaque classe la moyenne des probabilités de tous les classifieurs est calculée et le choix de la classe attribuée au document est alors fondé sur la plus forte moyenne.

Notons que plusieurs travaux s'appuient également sur un système de vote de classifieurs (Kuncheva, 2004 ; Kittler *et al.*, 1998).

Le choix de la procédure de classification a été appliqué sur chaque ensemble d'apprentissage. Nous avons conservé la méthode de classification la plus performante pour un corpus donné. Les données sont évaluées par validation croisée sur l'ensemble du corpus en s'appuyant sur les mesures de précision, rappel et F-mesure. La précision d'une classe *i* correspond à la proportion de documents correctement attribués à leur classe *i* par rapport aux documents attribués à la classe *i*. Le rappel calcule la proportion de documents correctement attribués à leur classe *i* par rapport aux documents appartenant à la classe *i*. La précision et le rappel moyens peuvent alors être calculés par rapport à l'ensemble des classes. Un compromis entre la précision et le rappel est alors calculé en mesurant la F-mesure (la F-mesure est la moyenne harmonique du rappel et de la précision).

Le système de vote mis en place étant décrit, les différents classifieurs utilisés par *Copivote* sont succinctement présentés ci-dessous. Une description plus précise de ces derniers est donnée dans (Plantié, 2006).

- ✔ **Bayes Multinomial.** La méthode de Bayes Multinomial (Wang *et al.*, 2003) qui est une approche classiquement utilisée pour la catégorisation de textes combine l'utilisation de la loi de Bayes bien connue en probabilités et la loi multinomiale.
- ✔ **Les Machines à Vecteurs Support** (**Support Vector Machine – S.V.M.**). La méthode Machines à Vecteurs Support (Joachims, 1998; Platt, 1998) consiste à délimiter par la frontière la plus large possible les différentes catégories des échantillons (ici les textes) de l'espace vectoriel du corpus d'apprentissage. Les vecteurs supports constituent les éléments délimitant cette frontière : plus la frontière est large, plus les risques d'erreurs de classification sont rares.
- ✔ **Les réseaux RBF (Radial Basis Function).** Les réseaux RBF sont fondés sur l'utilisation d'un réseau de neurones à fonctions radiales de base. Cette méthode utilise un algorithme de « clustering » de type « k-means » (MacQueen, 1967) avec application d'une méthode de régression linéaire. Cette technique est présentée dans (Parks & Sandberg, 1991).

4 Résultats

Le tableau 1 montre que la procédure de vote que nous avons mise en place améliore globalement les résultats. Notons que toutes les procédures de vote permettent une amélioration du même ordre même si des résultats légèrement meilleurs sont retournés avec

le « vote par somme pondérée ». Dans un deuxième temps, nous pouvons noter que l'utilisation des bigrammes (*CopivoteBi*) améliore globalement les résultats par rapport au traitement sans les utiliser (*Copivote*).

Le tableau 1 présente les méthodes de classification prises en compte dans le système de vote. Nous constatons que le classifieur de Bayes Multinomial est très performant avec un temps de calcul très faible. Les meilleurs résultats sont dans la grande majorité des cas obtenus par le classifieur SVM. Le classifieur RBF Network donne quant à lui des résultats décevants.

Corpus	**SVM**	**RBF-Network**	**Naïve Bayes Mult.**	***Copivote***	***CopivoteBi***
Corpus 1	61,02%	47,15%	59,02%	60,79%	61,28%
Corpus 2	76,47%	54,75%	74,16%	77,73%	79,00%
Corpus 3	50,47%	X	50,07%	52,52%	52,38%
Corpus 4	69,07%	61,79%	68,60%	74,15%	75,33%

TAB 1. *Moyenne des F-mesures des différentes méthodes sur le corpus de test.*

Le tableau 2 présente les résultats propres au corpus de test fourni par le Comité d'Organisation de DEFT'07. Ce tableau montre que seul le corpus 3 (relectures d'articles) donne des résultats décevants. Ceci peut s'expliquer par le nombre de textes constituant l'ensemble d'apprentissage trop faible et souvent bruité. A contrario, les excellents résultats obtenus avec le corpus 4 (débats parlementaires) pourraient s'expliquer par sa taille très importante qui favorise significativement les méthodes statistiques utilisées. Avec ce corpus, le système de vote améliore de manière importante les résultats obtenus par chacun des classifieurs (voir tableau 1). Notons par ailleurs que la F-mesure a une valeur de plus de 4% supérieure à la meilleure soumission de DEFT'07 avec ce corpus.

Précisons enfin que, de manière globale, une comparaison de nos résultats avec la meilleure soumission de DEFT'07 est donnée dans le tableau 2. Ceci montre que nous obtenons des résultats du même ordre voire légèrement meilleurs avec *CopivoteBi*.

Corpus	**Type de vote**	***Copivote***	***CopivoteBi***	***Meilleure soumission DEFT07***
		F-mesure	**F-mesure**	**F-mesure**
Corpus 1	Minimum	60,79%	61,28%	60,20%
Corpus 2	Moyenne	77,73%	79,00%	78,24%
Corpus 3	Minimum	52,52%	52,38%	56,40%
Corpus 4	Moyenne	74,15%	75,33%	70,96%
Total		**66,30%**	**67,00%**	**66,45%**

TAB 2. *Corpus de test de DEFT'07.*

5 Conclusion et perspectives

La classification de textes d'opinion est un thème porteur. La mise en place du défi DEFT'07 montre l'intérêt de l'étude de ce type de textes par la « communauté fouille de textes ». Cet article présente une approche consistant en une combinaison de représentations mots-clés et bigrammes tout en utilisant un système de vote de plusieurs classifieurs. Les résultats obtenus se sont révélés particulièrement intéressants avec une valeur de F-mesure légèrement supérieure à la meilleure soumission de DEFT'07.

Dans nos futurs travaux, nous comptons utiliser des représentations combinées mots-clés, bigrammes et trigrammes qui pourraient encore améliorer les résultats. Nous souhaitons également utiliser des procédures de vote avec un nombre plus important de classifieurs. Enfin, une étude plus globale en utilisant d'autres corpus et surtout des données textuelles de langues différentes pourraient être menée.

Références

Cover, T.M. & J.A. Thomas (1991). *Elements of Information Theory*: John Wiley.

Grouin, C., J-B. Berthelin, S. El Ayari , T. Heitz, M. Hurault-Plantet, M. Jardino, Z. Khalis, M. Lastes (2007). Actes de l'Atelier DEFT'07, Plate-forme AFIA 2007. (http://deft07.limsi.fr/)

Joachims, T. (1998). *Text Categorisation with Support Vector Machines : Learning with Many Relevant Features.* Proceedings of ECML.

Kuncheva, L. (2004). *Combining Pattern classifiers: Methods and Algorithms. J.* Wiley and Sons.

Kittler, J., M. Hatef, Robert P.W. Duin, J. Matas (1998). *On combining classifiers*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(3):226-239.

MacQueen., J.B. (1967). *Some Methods for classification and Analysis of Multivariate Observations.* Paper presented at the Symposium on Mathematical Statistics and Probability.

Parks, J., & I.W. Sandberg (1991). « Universal approximation using radial-basis function networks». In *Neural Computation* (Vol. 3, pp. 246-257).

Plantié, M. (2006). *Extraction automatique de connaissances pour la décision multicritère.* Thèse de Doctorat, Ecole Nat. Sup. des Mines de St Etienne et de l'Univ. Jean Monnet de St Etienne, Nîmes.

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

Roche, M., T. Heitz, O. Matte-Tailliez, Y. Kodratoff (2004). *EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés.* Proceedings of JADT'04 (Journées internationales d'Analyse statistique des Données Textuelles), p. 946-956,

Salton, G., C.S. Yang, C.T. Yu (1975), *A theory of term importance in automatic text analysis*, Journal of the American Society for Information Science, 26 , pp. 33-44.

Wang Y., Hodges J., Tang B. (2003). Classification of Web Documents using a Naive Bayes Method. *IEEE.*

Summary:

Text classification has the goal of gathering text documents by topics. In this paper, we will focus on classifying documents according to the opinion and value judgment they contain. Our approach is founded on a voting system using several classification methods.