Evaluation de la détection des émotions, des opinions ou des sentiments : dictature de la majorité ou respect de la diversité d'opinions ?

Jean-Yves Antoine[1], Marc Le Tallec[1], Jeanne Villaneau[2]

(1) Université François Rabelais de Tours, LI, 37000 Blois
(2) Université Européenne de Bretagne, VALORIA, 56100 Lorient
Jean-Yves.Antoine@univ-tours.fr, Marc.Le-Tallec@univ-tours.fr, Jeanne.Villaneau@univ-ubs.fr

Résumé - Détection d'émotion, fouille d'opinion et analyse des sentiments sont généralement évalués par comparaison des réponses du système concerné par rapport à celles contenues dans un corpus de référence. Les questions posées dans cet article concernent à la fois la définition de la référence et la fiabilité des métriques les plus fréquemment utilisées pour cette comparaison. Les expérimentations menées pour évaluer le système de détection d'émotions *EmoLogus* servent de base de réflexion pour ces deux problèmes. L'analyse des résultats d'*EmoLogus* et la comparaison entre les différentes métriques remettent en cause le choix du vote majoritaire comme référence. Par ailleurs elles montrent également la nécessité de recourir à des outils statistiques plus évolués que ceux généralement utilisés pour obtenir des évaluations fiables de systèmes qui travaillent sur des données intrinsèquement subjectives et incertaines.

Abstract - Emotion detection, opinion identification and sentiment analysis are generally assessed by means of the comparison of a reference corpus with the answers of the system. This paper addresses the problem of the definition of the reference and the reliability of the metrics which are commonly used for this comparison. We present some experiments led with *EmoLogus*, a system of emotion detection, to investigate these two problems. A detailed analysis of the quantitative results obtained by *EmoLogus* on various metrics questions the choice of a majority vote among several human judgments to build a reference. Besides, it shows the necessity of using more sophisticated statistical tools to obtain a reliable evaluation of such systems which are working on intrinsically subjective and uncertain data.

Mots-clés : Détection d'émotion, analyse de sentiments, fouille d'opinion ; Evaluation : métrique d'évaluation, constitution de référence, analyse statistique des résultats.
Keywords: Detection of emotion, sentiment analysis, opinion mining, Evaluation: objective measures, test reference, statistical analysis of the results.

1 Evaluation en détection des émotions / opinions / sentiments

La détection d'émotions, la fouille d'opinion ou l'analyse des sentiments sont des tâches très proches qui consistent à trouver et catégoriser dans des flux langagiers oraux ou écrits des passages porteurs d'un état émotionnel ou traduisant un jugement. La granularité de la détection est variable suivant l'application : il peut s'agir d'un document ou d'un discours complet, d'un paragraphe, d'une phrase (qui peut être spécifiquement un titre, par exemple) ou d'un tour de parole dans le cas du dialogue oral homme-machine. Le grain de catégorisation recherché peut également varier d'une tâche à l'autre. On peut ainsi ne considérer que trois classes principales (valence positive, négative ou neutre) ou rechercher une caractérisation plus fine sous la forme de modalités correspondant aux émotions principales définies en psychologie : colère, joie, dégout, peur, surprise, tristesse et émotion neutre (Ekman 1999).

Quelle que soit la tâche considérée, l'évaluation suit toujours le même paradigme : celui de la comparaison des réponses du système à une référence prédéfinie. Les différentes campagnes d'évaluation qui ont été menées à bien se différencient essentiellement par le choix de la métrique de comparaison. Soit C le nombre de classifications correctes (identiques à la référence) effectuées par le système et E le nombre de ses erreurs.

- La **robustesse** (*accuracy* en anglais) correspond à la proportion de réponses correctes du système sur l'ensemble du corpus de test. On a : R = C / (C+E). Cette mesure est surtout utilisée pour l'évaluation individuelle des performances d'un système comme dans (Callejas et Lopez-Cozar 2008).

- La **précision** et le **rappel** sont souvent utilisés en détection d'opinion et d'émotion. Le système peut choisir la valence neutre qui peut être interprétée comme une non-décision de la part de classifieurs entraînés pour identifier une expression émotive donnée. La précision quantifie la justesse des décisions prises par le système (P = C/(C+E)), tandis que le rappel estime sa part d'indécision (R = C/$C_{R,}$, où C_R est le nombre d'énoncés classés dans la référence). Ces deux indices peuvent être calculés globalement (macro-moyenne) ou être la moyenne d'un calcul fait pour chacune des classes (micro-moyenne). La **F-mesure** correspond à la moyenne harmonique de la précision et du rappel. Elle estime le compromis entre une prise de risque se traduisant par un fort rappel mais une faible précision, et inversement. On a F = 2.P.R / (P+R). Ces métriques ont été adoptées lors de l'*Emotion Challenge* d'Interspeech 2009 (Schuller et. al. 2009) et pour l'évaluation de la valence dans tâche *Affective Text* de la campagne SemEval-2 (Straparava & Mihalceva 2007). La campagne DEfi Fouille de Texte innove quelque peu en définissant un indice de confiance qui correspond à une F-mesure pondérée lorsque les réponses du système sont une distribution de probabilités sur chaque classe (Grouin *et. al.* 2007).

- D'usage moins fréquent, le coefficient de **corrélation** de Pearson est une autre manière d'estimer la proximité des réponses avec la référence. On a CP = $\sigma_{sr}/(\sigma_s.\sigma_r)$ où σ_{sr} est la covariance entre les réponses du système et la référence, et σ_s et σ_r correspondent à la variance des réponses du système et des juges qui ont établi la référence sur le corpus de test. Cette mesure est utilisée dans la campagne SemEval-2 pour l'évaluation de la catégorisation en modalité émotionnelle (Straparava & Mihalceva 2007).

- Enfin, certaines évaluations choisissent d'évaluer le système comme un expert humain, on cherchant à estimer l'**accord inter-annotateur** observé entre les réponses du système et la référence. Le plus souvent, cet accord est estimé par un calcul de **Kappa** (Cohen 1960) : soit P_o la proportion d'accord observée effectivement et P_a celle correspondant à une annotation aléatoire, on a K = P_o - P_a / (1 – P_a).

La diversité des situations de test fait qu'il est malaisé de comparer les résultats d'une campagne à une autre. Il n'en reste pas moins que peu de travaux ont étudié l'influence des pratiques de test sur les résultats. Les multiples mesures de Kappa pour estimer l'accord inter-annotateur ont fait l'objet d'expérimentations critiques (Callejas & Lopez-Cozar 2008, Hayes & Krippendorf 2007). Mais en dehors de cette réflexion méthodologique (souvent le fait de psycholinguistes ou statisticiens), aucune campagne d'évaluation n'a par exemple utilisé de concert plusieurs métriques. Ce papier s'appuie précisément sur l'évaluation d'un système de détection des émotions pour répondre aux questions suivantes :

1) Quelles influences ont les métriques d'évaluation sur l'estimation des systèmes ?
2) Comment interpréter clairement les résultats en termes de performances ?
3) Quelle est l'influence de l'utilisation systématique du vote majoritaire sur l'évaluation ? Rend-elle bien compte de la diversité des jugements humains et donc des attentes utilisateurs ?

2 Cas d'école : évaluation du système de détection des émotions *EmoLogus*

Pour répondre à ces questions, nous avons mené plusieurs expérimentations avec un système de détection des émotions, *EmoLogus* développé dans le cadre du projet ANR *EmotiRob*. Ce projet intervient dans un contexte d'application original: la réalisation d'un robot compagnon émotionnel pour des enfants fragilisés. *EmoLogus* est un composant qui sert à caractériser l'émotion du locuteur en se basant sur le contenu linguistique de ces messages oraux, et non, comme c'est souvent le cas, en conduisant une analyse acoustique du signal de parole. Il se base sur le principe de compositionnalité des émotions: les mots lexicaux possèdent une valeur émotionnelle fixe définie par une norme psycholinguistique, tandis que les verbes et les adjectifs agissent comme des prédicats, dont le résultat dépend de la valeur émotionnelle de leurs arguments (LeTallec *et al.* 2010). *EmoLogus* analyse ainsi la structure de l'énoncé pour identifier l'émotion qu'il porte. Il a été comparé avec un système de base (*baseline*) qui ne considère l'énoncé que comme un sac de mots: on se contente ici de sommer les valences émotionnelles des mots qui le composent.

Notre objectif étant de comparer les performances des deux systèmes d'un point de vue purement linguistique (i.e. sans intégrer l'influence des erreurs de reconnaissance de la parole), les expérimentations ont été menées sur un conte enfantin (*Le Grand Nord*), comme souvent sur ce type d'applications (Volkova *et al.* 2010). Le corpus de test est de taille limitée (93 phrases) mais la référence a été obtenue, à la différence de la plupart des campagnes d'évaluation, par un nombre élevé de juges (31 en pratique). Notre objectif est en effet d'étudier l'influence de la dispersion des jugements sur l'évaluation. L'annotation a consisté à attribuer à chaque énoncé une valeur émotionnelle sur une échelle de 5 classes variant de -2 (émotion très négative) à +2 (très positif). Deux évaluations ont donc été conduites qui ont porté sur :

- la **valence + intensité émotionnelle** – Les cinq classes d'annotation sont considérées.
- la **valence** seule – on ne considère ici que trois classes : émotion positive, négative et neutre

Enfin, nous avons évalué l'influence du contexte de discours en réalisant deux annotations manuelles successives. Dans un premier cas (évaluation hors contexte), les énoncés étaient présentés dans un ordre aléatoire. Dans le second, l'ordre de présentation respectait le fil du conte. Les tableaux 1 et 2 présentent les performances des deux systèmes suivant les métriques présentées précédemment, à la fois pour l'annotation en valence seule ou celle en valence et intensité.

Hors-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	82,8 %	0,77	0,77	0,69
Baseline	64,6%	0,69	0,67	0,5
En-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	75,3%	0,62	0,73	0,58
Baseline	53,8 %	0,39	0,42	0,31

Tableau 1 : Evaluation des performances en annotation en valence (3 classes).

Hors-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	69,9 %	0,59	0,82	0,56
Baseline	52,7 %	0,41	0,69	0,30
En-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	59,2%	0,47	0,79	0,48
Baseline	39,8 %	0,25	0,5	0,22

Tableau 2 : Evaluation des performances en annotation en valence et intensité (5 classes).

Fait rassurant, certaines conclusions se retrouvent sur toutes les métriques retenues : on observe dans tous les cas une dégradation des performances lors de la prise en compte du contexte. L'analyse des réponses des systèmes nous montre que ceux-ci peinent à intégrer aussi finement qu'un sujet humain le contexte dans l'appréciation des émotions, l'enchainement des actions créant une ambiance difficile à modéliser. Par ailleurs, trois des métriques sur quatre semblent indiquer que, sans surprise, la catégorisation est plus difficile avec un nombre de classes supérieur : les performances des systèmes sont meilleures sur la tâche de classification en valence seule. Seul le coefficient de corrélation de Pearson se distingue ici. Cette particularité nous amène à considérer avec réserve cette métrique comme indicateur de performance. Il est d'ailleurs malaisé de rapprocher corrélation et identité de réponses comme le fait l'usage de cette métrique à fin d'évaluation. Dans les autres cas, nos observations, corroborées par d'autres études (Grouin et al. 2009) posent la question de la mise en place d'une métrique qui serait moins sensible au nombre de classes afin de faciliter les comparaisons entre évaluations.

On note enfin que le système de base présente des performances inférieures à *EmoLogus* pour toutes les métriques. Ce résultat montre l'intérêt d'une prise en compte de la structure des énoncés. Cette cohérence des résultats pourrait laisser croire à l'absence d'influence de la métrique sur le classement des systèmes d'une campagne d'évaluation. Les résultats présentés ci-après montrent qu'il n'en est rien.

3 Prise en compte de la diversité des opinions dans l'évaluation des systèmes

Les résultats précédents suggèrent que les métriques utilisées couramment sont d'assez bons indicateurs de la performance relative des systèmes. Leurs indications restent toutefois difficiles à interpréter en termes de qualité absolue, avant tout du fait de la dispersion des jugements humains. Les émotions, les opinions correspondent en effet à des états cognitifs complexes influencés par le contexte à court-terme (historique de l'interaction, situation d'énonciation) comme à long terme (vécu personnel et socioculturel) et dont la perception varie de manière sensible d'une personne à l'autre. Il n'est donc pas étonnant que toutes les études montrent que l'accord inter-annotateur est bien plus faible sur ce type de tâche que sur, par exemple, l'annotation en catégories morpho-syntaxique. L'annotation du corpus de test que nous avons réalisée s'est ainsi traduit par des valeurs de Kappa assez faibles entre les 31 juges : 0,48 en contexte et 0,51 hors contexte. Plutôt que de discuter de la pertinence des mesures de Kappa, la figure 1 donne la répartition des votes des 31 juges pour la phrase «*c'est court comme belle saison, je sais, mais les jours seront très longs*».

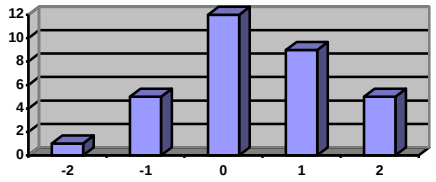

Figure 1 : Exemple de distribution des votes sur un énoncé du corpus

On voit que les votes des experts se distribuent sur l'ensemble de l'échelle d'annotation. Face à une telle dispersion, on peut s'interroger sur la pertinence d'une référence obtenue par vote majoritaire. Dans le cas présent, la catégorie retenue ne représente que 38,7% des votes ! La référence en valence et intensité peut ainsi représenter moins de la majorité des votes dans 25,6% des annotations en contexte. Face à cette observation, deux questions se posent dès lors : 1) comment situer les performances des systèmes face aux capacités des sujets humains et 2) comment intégrer la dispersion des jugements dans l'évaluation.

3.1 Evaluation des systèmes et performances humaines

Compte-tenu de la dispersion des jugements humains observés, il est clair qu'aucun juge humain ne peut égaler la référence de test, qui ne représente qu'une approximation de l'opinion moyenne d'une population donnée. Dès lors, les performances des systèmes ne devraient pas être estimées en termes de métriques absolues, mais comparées à celles des sujets humains. Aussi avons-nous repris les jugements des 31 annotateurs que nous avons soumis aux mêmes métriques que les systèmes testés. En ordonnant les résultats, nous avons déterminé le classement des systèmes parmi les annotateurs humains (tableaux 3 et 4).

Hors-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	9 ème	22 ème	23 ème	10 ème
Baseline	31 ème	31 ème	33 ème	31 ème
En-contexte	**Robustesse**	**F-mesure**	**Pearson**	**Kappa**
EmoLogus	19 ème	32 ème	24 ème	24 ème
Baseline	33 ème	33 ème	33 ème	33 ème

Tableau 3 : Classement des systèmes face aux 31 experts humains : annotation en valence (3 classes)

Plusieurs conclusions peuvent être tirées de ces classements. Tout d'abord, les difficultés des deux systèmes à considérer le contexte se retrouvent, sans doute d'une manière plus visible ici. Par ailleurs, si le système de base a des performances médiocres, il est amusant de constater que, pour certaines métriques, ses performances dépassent celles des juges les plus atypiques. A l'opposé, *EmoLogus* présente un classement très honorable (3° sur 33) sur la tâche d'étiquetage hors-contexte en valence et intensité. Performance que ne traduisait pas vraiment les 75,3% de robustesse données au tableau 2. Il semblerait que, sur une tâche assez complexe (5 classes), le système, basé sur une norme émotionnelle équilibrée, soit plus à même d'adopter un comportement moyen représentatif de la population qu'un individu particulier.

Hors-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	3 ème	13 ème	11 ème	9 ème
Baseline	31 ème	31 ème	30 ème	32 ème
En-contexte	Robustesse	F-mesure	Pearson	Kappa
EmoLogus	27 ème	28 ème	23 ème	18 ème
Baseline	33 ème	33 ème	33 ème	33 ème

Tableau 4 : Classement des systèmes face aux 31 experts : annotation en valence+ intensité (5 classes)

Il nous semble ainsi qu'une évaluation par classement par rapport à des juges humains est plus à même de faire ressortir les qualités et faiblesses des systèmes. On remarque toutefois que ces classements varient significativement suivant la métrique utilisée. Cette observation est de nature à jeter un doute sur les évaluations compétitives ne prenant en considération qu'une seule métrique, comme c'est généralement le cas en détection d'émotions ou d'opinion. Elle nous incite par ailleurs à proposer une métrique qui prenne mieux en compte, à la base, la dispersion des jugements lors de l'établissement de la référence.

3.2 Intégrer la distribution des jugements pour éviter une dictature de la majorité

Les calculs de taux de robustesse, de précision ou de rappel reposent tous sur une évaluation binaire : une réponse est considérée comme bonne ou erronée au vu de la référence. Compte tenu de la fragilité d'une référence obtenue par vote majoritaire, nous avons voulu pondérer ces calculs évaluatifs en tenant compte de la distribution des jugements humains. Considérons une annotation en valence (3 classes). Pour un énoncé donné, supposons que les jugements se sont répartis suivant la distribution suivante :

Négatif : 35% Neutre : 48 % Positif : 17% ⇨ Emotion choisie pour la référence : Neutre

Si le système propose la classe *Neutre*, sa réponse est considérée correcte à 48% dans notre évaluation pondérée. S'il choisit la classe *Négatif*, sa réponse est considérée comme correcte à 35%. Ce qui à nos yeux représente une évaluation plus proche de la réalité qu'une décision binaire. Par ailleurs, notre objectif est d'estimer la performance des systèmes par rapport aux capacités humaines, et non pas vis-à-vis d'une référence quelque peu idéale. Dès lors, le poids d'un énoncé n'est plus de 1 dans le calcul du score global, mais pondéré par le poids de la classe qui a reçu le maximum de votes : 0,48 dans le cas présent. Cette normalisation nous assure par ailleurs qu'un système qui choisirait systématiquement la classe de référence atteindrait par un score maximal de 100%. Ce modèle de calcul fait disparaître en pratique la notion de prise de décision : très peu d'énoncés sont jugés sans émotion par la totalité des juges (dans notre exemple : 1 énoncé hors contexte et 0 en contexte) et la grande majorité des énoncés (82 sur 93 hors contexte et 76 en contexte) sont classés comme neutres par au moins l'un des juges. La notion de précision et rappel ne se distingue donc plus d'un calcul en robustesse. De même, les indices de Pearson et Kappa deviennent, tels quels, inutilisables.

Hors-contexte	Robustesse : valence seule	Robustesse : valence + intensité
EmoLogus	87,4% (18 ème)	85,6% (9 ème)
Baseline	80,7% (32 ème)	71,9% (31 ème)
En-contexte	Robustesse	Robustesse
EmoLogus	79,3% (30 ème)	74,7%(28 ème)
Baseline	64,9 % (33 ème)	59,9%(33 ème)

Tableau 5 : Evaluation pondérée des systèmes : résultats et classements.

Parmi les indices précédents, seule la robustesse reste donc significative. Le tableau 5 donne les résultats de ce calcul pondéré. Ces résultats nous semblent représenter plus fidèlement les performances intrinsèques des

systèmes de deux points de vue. D'une part, les valeurs de performances se sont légèrement accrues, ce qui correspond au ressenti d'une analyse qualitative des réponses. En effet, en cas d'erreur, le système propose souvent la classe arrivée en seconde position au sein de la population de juges, ce que prend en compte cette métrique pondérée. A l'opposé, le classement du système par rapport aux juges humains s'est dégradé. Une fois encore, cette observation montre la dépendance à la métrique d'évaluation d'une campagne de test de type compétitif.

4 Conclusion

L'analyse de ce « cas d'école » montre les difficultés liées à l'évaluation de ces données subjectives que sont les opinions et les émotions. Elle montre également que les indices d'évaluation classiquement utilisés donnent des résultats instables donc peu fiables. Ce constat étant établi, il reste à étudier des pistes pour pallier cette insuffisance. Le coefficient alpha de Krippendorff ouvre des perspectives intéressantes par sa plasticité : dans l'exemple d'évaluation présenté dans ce papier, il permet en effet d'introduire des métriques entre les classes et de mesurer de façon fine l'écart entre la notation proposée par un expert et la référence complète dans sa dispersion, sans recours au vote majoritaire. Une autre piste à étudier est celle des ensembles flous : la référence que constitue l'ensemble des avis des experts correspond à des probabilités d'appartenance floue (Milleman et Scholl 1996). Le meilleur système est celui qui donne l'ensemble net le plus proche de cet ensemble flou.

Remerciements : ce travail a été pour partie réalisé dans le cadre de l'ANR EmotiRob (PSIROB'06).

Références

CALLEJAS Z., LOPEZ-COZAR R. (2008) Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*. 50 (2008), 416-433

COHEN J. (1960) A coefficient of agreement for nominal scales. *Educational & Psycho. Meas.*, 20, 37-46.

EKMAN P. (1999) *Patterns of emotions: New Analysis of Anxiety and Emotion* . Plenum Press.

GROUIN C., HURAULT-PLANTET M., PAROUBEK P., BERTHELIN J.B. (2009) DEFT'07: une campagne d'évaluation en fouille d'opinion. *Revue des Nouvelles Technologies de l'Information.* vol.E-17.1-24.

HAYES A.F, KRIPPENDORFF K. (2007) Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures 1,1: 77-89.

LE TALLEC M., VILLANEAU J., ANTOINE J.-Y., SAVARY A., SYSSAU-VACCARELLA A. (2010) Emologus, a compostional model of emotion detection based on the propositionnal content of spoken utterances Proc. *Text, Speech and Dialogue, TSD'2010*, Brno, Tchéquie, sept. 2010. In. *LNCS/LNAI 6231*, Springer

MILLEMANN S., SCHOLL P. (1996) Estimation de quantités subjectives floues par des techniques connexionnistes. Application à l'évaluation du confort automobile. In. *MODULAD, numéro 17, Juin 1996*.

SCHULLER B., STEIDL S., BATLINER A. (2009) The INTERSPEECH 2009 Emotion Challenge. Proc. *Interspeech'2009*, Brighton, UK.

STRAPPARAVA C., MIHALCEVA R. (2007). SemEval-2007 Task 14: Affective Tex. *Proc. 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Juin 2007, 70–74.

VOLKOVA E., MOHLER B., MEURES D., GERDEMANN D. AND BÜLTHOFF,H. (2010) Emotional perception of fairy tales: achieving agreement in emotion annotation of text, Proc. *NAACL HLT' 2010*