Exploitation des plongements de mots pour l'analyse d'opinion et du langage figuratif des tweets (DEFT 2017)

Amir Hazem[1] Basma El Amal Boussaha[1] Nicolas Hernandez[1]
(1) LS2N, Université de Nantes, 44000 Nantes, France
prénom.nom@univ-nantes.fr

RÉSUMÉ

Cet article présente la participation de l'équipe TALN [1] du laboratoire LS2N [2] au défi fouille de textes (DEFT) 2017. Nous avons développé un système fondé sur les plongements de mots pour traiter les trois tâches du défi. Nous avons obtenu un F-score de 53,42% pour la tâche 1, 71,97% pour la tâche2 et 53,38% pour la taĉhe3.

ABSTRACT

This paper presents the participation of the TALN team of LS2N laboratory to the Défi fouille de textes (DEFT) 2017. We developped a system based on word embeddings for this shared task. We obtained an F-score of 53.42% for the first task, 71.97% fot the second task and 53.38% for the third task.

MOTS-CLÉS : fouille d'opinion, plongements de mots.

KEYWORDS: opinion mining, word embeddings.

1 Introduction

Ce travail se place dans le domaine de l'analyse de sentiments, et plus particulièrement, de celui de la détection du langage figuratif et de sa polarité. Si l'analyse d'opinion du langage littéral a montré des résultats plutôt satisfaisants, ce n'est pas encore le cas s'agissant du langage figuratif. Puisque la performance des systèmes d'analyse d'opinion dépend grandement de ce type de discours, une attention toute particulière leur est portée. La détection automatique du sens figuré ou imagé comme l'ironie, le sarcasme ou l'humour reste à ce jour un challenge des plus intéressants et des plus difficiles tant le langage humain est subtil et qui plus est, dans une langue très riche comme le français.

Le défi de fouille de texte (DEFT) 2017, offre un environnement d'évaluation en proposant trois tâches de classification de tweets selon le type de langage (figuratif ou non figuratif) et selon la polarité du langage (objective, positive, négative ou mixte). La première tâche consiste à détecter la polarité des tweets non figuratifs. Le but étant de savoir si l'opinion de l'auteur d'un tweet est positive, négative, mixte ou objective. La deuxième tâche quant à elle, consiste à détecter le type de langage utilisé. C'est-à-dire, d'identifier si un tweet fait référence à un style figuratif ou non figuratif. Enfin, la troisième et dernière tâche consiste à détecter la polarité d'un tweet quel que soit son type de

1. Traitement Automatique du Langage Naturel
2. Laboratoire des Sciences du Numérique de Nantes

langage (figuratif ou non).

Dans le but de proposer une approche simple et générique, nous abordons les trois tâches de manière similaire. Nous considérons les différentes étiquettes (ou classes) de façon abstraite, sans réserver de traitement particulier à chacune d'elle. L'approche proposée se base sur un appariement de tweets similaires en attribuant à un tweet de test, l'étiquette du tweet d'entraînement le plus similaire. L'hypothèse sous-jacente est qu'au même titre que les mots, les unités textuelles plus longues (ici les tweets) peuvent être représentées dans un espace de plongements de mots. Ce qui va permettre de mesurer une similarité en cosinus entre tweets et donc de rapprocher des tweets de même classe.

Le reste de cet article est organisé comme suit. La section 2 décrit l'ensemble de données de la campagne d'évaluation. La section 3 présente le système que nous avons développé. Nous présentons ensuite en section 4 les différentes expériences que nous avons menées et les résultats obtenus. Dans la section 5 nous analysons les différents résultats et enfin la section 6 conclut ce travail et donne quelques perspectives.

2 Description des données

L'édition 2017 du défi de fouille de texte, propose trois tâches de classification de tweets qui traitent de sujets d'actualité (politique, sport, cinéma, émissions TV, artistes, etc.) en français. Les tweets sont à catégoriser selon leur langage, à savoir : un langage figuratif ou non figuratif, et selon leur polarité, à savoir : une polarité objective, positive, négative ou mixte. Trois types de langage figuratif sont représentés : l'ironie, le sarcasme et l'humour. La première tâche traite exclusivement les tweets non figuratifs selon leur polarité. Quatre classes sont donc à prédire. Le corpus d'entraînement contient 3906 tweets et celui du test 976 tweets. La deuxième tâche quant à elle, consiste en l'identification du langage figuratif. Ainsi, deux classes sont à prédire. Le nombre de tweets d'entraînement est de 5853 tweets et celui du test de 1464 tweets. Enfin, la troisième et dernière tâche, englobe les deux premières tâches et consiste en la classification des tweets figuratifs et non figuratifs selon leur polarité. Il y a donc quatre classes à identifier. Le corpus de tweets comprend 5228 tweets pour l'entraînement et 1281 tweets pour le test.

Dans ce qui suit, nous présentons pour chaque tâche quelques statistiques sur le nombre de tweets selon leur classe :

Tâche 1	Polarité				
	Objective	Positive	Négative	Mixte	Total
Train	1643	494	1268	501	3906
	42,06%	12,65%	32,46%	12,83%	100%
Test	411	123	318	124	976
	42,11%	12,60%	32,58%	12,70%	100%

TABLE 1 – Effectif des classes à prédire pour la tâche1 sur les données d'entraînement et de test (DEFT 2017)

Le tableau 1 montre l'effectif et le pourcentage des quatre polarités des tweets non figuratifs à prédire de la tâche 1. Nous observons qu'il y a une forte tendance aux tweets objectifs (42,06%) et négatifs

(32,46%). Pour ce corpus, nous pouvons déduire que dans un langage non figuratif (littéral), il y a une forte tendance à exprimer une opinion objective ou négative. Une opinion positive ou mixte est beaucoup plus rare.

Tâche 2	Style		
	Figuratif	Non figuratif	Total
Train	1947	3906	5853
	33,26%	66,73%	100%
Test	488	976	1464
	33,33%	66,66%	100%

TABLE 2 – Effectif des classes à prédire pour la tâche2 sur les données d'entraînement et de test (DEFT 2017)

Le tableau 2 montre l'effectif et le pourcentage des deux types de langage des tweets (figuratif et non figuratif) de la tâche 2. Dans ce corpus, nous observons une majorité de tweet non figuratifs que ce soit dans le corpus d'entraînement ou de test. Ici, nous pouvons en déduire qu'un tiers des personnes (33,33%) utilise l'ironie, le sarcasme ou l'humour pour exprimer son opinion.

Tâche 3	Styles figuratifs et non figuratifs				
	Objective	Positive	Négative	Mixte	Total
Train	1718	504	2263	633	5118
	33,56%	9,84%	44,21%	12,36%	100%
Test	430	125	568	158	1281
	33,56%	9,75%	44,34%	12,33%	100%

TABLE 3 – Effectif des classes à prédire pour la tâche3 sur les données d'entraînement et de test (DEFT 2017)

Le tableau 3 montre l'effectif et le pourcentage des quatre types de polarité des tweets (figuratif et non figuratif) de la tâche 3. Comme pour le premier tableau (TABLE 1) nous observons que dans ce corpus, la majorité des tweets sont objectifs et négatifs et ceci, indépendemment du type de langage utilisé.

De manière globale se dégage une tendance à utiliser des tweets objectifs et négatifs que ce soit dans le langage figuratif ou non figuratif. En revanche, le nombre d'avis positifs et mixtes est beaucoup plus faible.

3 Description du système

Nous proposons un système simple et indépendant de la tâche traitée. Nous partons du principe que l'étiquette d'un tweet peut être trouvée dès lors que l'on a à disposition un corpus d'entraînement avec un minimum de tweets étiquetés.

Notre approche est fondée sur la similarité entre tweets pour prédire une étiquette. Ainsi, pour un tweet donnée du jeu de test, nous calculons sa similarité avec tous les tweets du corpus d'entraînement. Puis, nous affectons au tweet test, l'étiquette du tweet d'entraînement le plus proche (ayant la similarité la plus élevée). La sélection de l'étiquette peut se faire soit sur le tweet le plus proche, soit sur un ensemble de tweets. C'est-à-dire, que sur les *n* premiers tweets qui sont les plus similaires au tweet test, nous choisissons l'étiquette qui apparaît le plus dans cet ensemble. La valeur de *n* est fixée de manière empirique.

Pour calculer la similarité entre deux tweets, nous utilisons la mesure du cosinus sur leurs vecteurs de plongements. Chaque tweet est préalablement représenté par son vecteur de plongement moyen qui est la somme des vecteurs de plongements des mots qui le composent (Mikolov *et al.*, 2013; Arora *et al.*, 2017). Les vecteurs de plongements des mots sont appris sur les tweets du corpus d'entraînement [3]. L'apprentissage des vecteurs de plongements a été fait via le modèle Skip-Gram en utilisant la bibliothèque Gensim (Řehůřek & Sojka, 2010). Ce choix est motivé par (Mikolov *et al.*, 2013) qui ont observé que le modèle Skip-Gram était plus approprié pour des corpus de petite taille. Nos expériences ont confirmé ces observations. Nous présentons dans la section suivante une étude comparative des principaux paramètres à définir : le nombre de tweets d'entraînement, le seuil de l'effectif des mots, etc.

4 Expériences et résultats

Dans cette section, nous présentons les résultats obtenus sur le jeu de test de la campagne DEFT. Nous avons fait varier deux paramètres, à savoir : l'effectif minimal d'un mot dans le corpus pour considérer son vecteur de plongement dans le calcul du vecteur moyen d'un tweet (noté *effectif*) et le nombre de tweets candidats retenus pour sélectionner la classe qui apparaît le plus dans cet ensemble (noté *n*). La taille de la fênetre pour le calcul des plongement a été fixée à 5 et le nombre de dimensions des plongements de mots à 800. Aucun pré-traitement n'a été effectué sur le corpus mis à part la tokenisation.

Tâche 1	Mesures		
	P	R	F-Score
Run1 (effectif=5, n=50)	**60,84**	51,72	49,28
Run2 (effectif=5, n=10)	55,39	**53.79**	**53,42**
Run3 (effectif=2, n=50)	51,08	51,67	48,65

TABLE 4 – Résultats sur le corpus de test de la tâche1 (DEFT 2017)

Selon le tableau 4, nous constatons que le meilleur F-score est obtenu en utilisant un effectif de 5 et un ensemble de tweets d'entraînement de 10. Le nombre de dimensions a été fixé à 800 [4]. Une valeur plus élevée de *n* (ici *n=50*) améliore la précision (60,84%) mais fait baisser le rappel et par conséquent le F-score.

3. Une autre manière de faire serait d'apprendre ces plongements sur un corpus plus large comme wikipedia ou autres, mais en effectuant plusieurs expériences nous nous sommes rendu compte que cela desservait notre approche.

4. 800 est la valeur qui a montré les meilleures performances sur des expériences menées sur un corpus de développement.

Tâche 2	Mesures		
	P	R	F-Score
Run1 (effectif=2, n=10)	71,47	**72,28**	71,81
Run2 (effectif=2, n=10)	**72,81**	71,41	**71,97**
Run3 (effectif=5, n=50)	72,19	69,51	70,39

TABLE 5 – Résultats sur le corpus de test de la tâche2 (DEFT 2017)

Selon le tableau 5, nous constatons que le meilleur F-score est obtenu en utilisant un effectif de 2 et un ensemble de tweets d'entraînement de 10. Le nombre de dimensions a été fixé à 800 de la même manière que pour la tâche1.

Tâche 3	Mesures		
	P	R	F-Score
Run1 (effectif=2, n=10)	**57,16**	**53,34**	**53,38**
Run2 (effectif=2, n=50)	46,50	48,80	47,09
Run3 (effectif=5, n=50)	46,67	49,51	47,67

TABLE 6 – Résultats sur le corpus de test de la tâche3 (DEFT 2017)

Selon le tableau 6, nous constatons que les meilleurs résultats sont obtenus en utilisant un effectif de 2 et un ensemble de tweets d'entraînement de 10. Une valeur plus élevée de n (ici $n=50$) fait chuter les résultats. De manière globale, une taille de dimensions de vecteurs de plongements égale à 800, un effectif de mots égal à 2 et un ensemble de tweets d'entraînement égal à 10 donnent les meilleurs résultats pour les trois tâches du défi.

5 Discussion

Notre système aborde les trois tâches de la même manière, ce qui rend notre approche simple et générique et constitue donc un avantage en soi. Ceci étant dit, une représentation plus dédiée des vecteurs de plongements selon la tâche et donc, selon les caractéristiques du problème abordé aurait peut-être pu améliorer les performances du système. Par exemple, identifier des marqueurs du langage figuratif et se baser sur ces derniers pour le calcul du vecteur de plongement d'un tweet. Aussi, les mots d'un tweet ont été considérés sans pondération particulière. Sachant que les mots d'une phrase n'ont pas tous la même importance et impact sur la compréhension, il aurait été pertinent d'explorer cette piste.

Les meilleurs résultats ont été globalement obtenus en utilisant un effectif de mot égal à 2 et un ensemble de tweets d'entraînement égal à 10. Ceci s'explique premièrement par la taille limitée des tweets. Filtrer les mots de fréquence inférieure à 5 peut pénaliser le calcul du vecteur moyen d'un tweet. Deuxièmement, un ensemble de 10 tweets d'entraînement les plus similaires à un tweet de test semble être un bon compromis. Sélectionner beaucoup de tweets risque de considérer des scores

faibles de similarité. Une alternative, serait de fixer un seuil de score de similarité au lieu d'utiliser le rang (comme ici avec un n=10 par exemple). Des expériences qui consistaient à filtrer les mots outils ont été menées, néanmoins les résultats obtenus étaient en deçà des performances de notre système avec les paramètres utilisés dans ce travail. Là aussi, la taille des tweets et du corpus d'entraînement peuvent expliquer l'inefficacité de cette démarche.

Au regard des résultats obtenus sur les trois tâches, notre approche semble être plus performante pour la tâche 2 avec un F-score de 71,97%. Les résultats des tâches 1 et 3 (53,42% et 53,38% respectivement), sont plus faibles et semblent indiquer qu'une approche aussi générique n'est pas appropriée. Ceci étant dit, les résultats sont à considérer en fonction de la difficulté de la tâche. La tâche 2 étant plus facile que les deux autres. De plus, la répartition des différentes classes selon la tâche n'étant pas équilibrée, ceci rend difficile l'interprétation des résultats. Quoi qu'il en soit, une étude plus approfondie des caractéristiques du langage figuratif est certainement nécessaire pour améliorer notre système.

6 Conclusion

Nous avons proposé une approche simple qui aborde la classification des tweets selon leur type de langage et selon leur polarité. Notre approche représente chaque tweet par son vecteur moyen de plongements de mots. L'attribution d'une classe pour un tweet du corpus de test se base sur les n tweets les plus similaires du corpus d'entraînement. Plusieurs améliorations peuvent être envisagées, notamment dans la manière de sélectionner les tweets similaires en utilisant par exemple, un système de vote plus pertinent qu'un simple comptage du nombre de classes. Mis à part la tokenisation, aucun pré-traitement n'a été appliqué, cette direction est aussi à explorer sachant qu'il peut y avoir beaucoup de bruit dans la rédaction des tweets. Enfin, une attention particulière aux caractéristiques du langage figuratif est sans doute nécessaire pour améliorer notre travail.

Remerciements

Ce travail a bénéficié d'une aide conjointe de "the Unique Interministerial Fund" (FUI) No. 17 faisant partie du projet ODISAE [5] et de l'ANR 2016 PASTEL [6].

Références

ARORA S., YINGYU L. & TENGYU M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, p. 1–11.

F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

G. DIAS, Ed. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

5. http://www.odisae.com
6. http://www.agence-nationale-recherche.fr/?Projet=ANR-16-CE33-0007

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benarmara *et al.*, 2007), p. 101–110.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

ŘEHŮŘEK R. & SOJKA P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta : ELRA. `http://is.muni.cz/publication/884893/en`.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benarmara *et al.*, 2007), p. 401–410.