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

Amir Hazem, Basma El Amal Boussaha et Nicolas Hernandez

LS2N, Université de Nantes, France
Défi Fouille de Textes (DEFT 2017)

26 Juin 2017, Orléans (France)

Outline

Outline

Outline

Outline

Introduction : Défi de fouille de texte (DEFT) 2017

- Trois tâches de classification de tweets
  - type de langage (figuratif ou non figuratif)
  - polarité du langage (objective, positive, négative ou mixte)
  - Tâche 1 : détecter la polarité des tweets non figuratifs
  - Tâche 2 : détecter le type de langage utilisé
  - Tâche 3 : détecter la polarité d'un tweet quel que soit son type de langage

Introduction : Défi de fouille de texte (DEFT) 2017

- Trois tâches de classification de tweets
  - type de langage (figuratif ou non figuratif)
  - polarité du langage (objective, positive, négative ou mixte)
  - Tâche 1 : détecter la polarité des tweets non figuratifs
  - Tâche 2 : détecter le type de langage utilisé
  - Tâche 3 : détecter la polarité d'un tweet quel que soit son type de langage
- Première expérience en fouille de texte
- Approche simple
- Les trois tâches sont abordées de manière similaire

Approche

- Appariement de tweets similaires
- Représentation des unités textuelles (tweets) dans un espace de plongements de mots au même titre que les mots
- Similarité en cosinus entre vecteurs de plongements moyens des tweets

Approche

- Construire les vecteurs de plongements des mots à partir des tweets du corpus d'entraînement
  - modèle : Skip-Gram
  - bibliothèque : Gensim
- Représenter chaque tweet par son vecteur de plongement moyen qui est la somme des vecteurs de plongements des mots qui le composent
- Calculer la similarité entre un tweet test et tous les tweets du corpus d'entraînement
- Affecter au tweet test, l'étiquette des n tweets d'entraînement les plus proches

Expériences et résultats (Tâche 1)

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%
F-score	77,1%	56,5%	59,3%	20,5%	53,4%

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

Expériences et résultats (Tâche 1)

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: Résultats sur le corpus de test de la tâche1 (DEFT 2017)

Expériences et résultats (Tâche 2)

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%
F-score	61,6%	82,3%	71,9%

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

Expériences et résultats (Tâche 2)

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

Table: Résultats sur le corpus de test de la tâche2 (DEFT 2017)

Expériences et résultats (Tâche 3)

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%
F-score	71,7%	55,2%	70,0%	16,5%	53,3%

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

Expériences et résultats (Tâche 3)

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: Résultats sur le corpus de test de la tâche3 (DEFT 2017)

Conclusion

- Approche simple qui aborde la classification des tweets selon leur type de langage et selon leur polarité
- 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

Perspectives

- 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.
- Une attention particulière aux caractéristiques du langage figuratif est sans doute nécessaire pour améliorer notre travail.

Questions ?

Questions ?