Classification de phrases courtes: des approches non-supervisées aux approches faiblement supervisées

Kaoutar Ghazi, Sébastien Marchal, Andon Tchechmedjiev
Pierre-Antoine Jean, Nicolas Sutton-Charani, Sébastien Harispe

LGI2P, IMT Mines Alès

Contexte

Contexte industriel

Contexte industriel

Contexte industriel

Administration : **Mairie**

24 motifs de visite

Chaque motif est **décrit** par un **texte court**

Échantillon de **180 demandes** clients

Contexte scientifique

Problème de classification de phrases courtes

Contexte scientifique

Problème de classification de phrases courtes

$$Ensemble\ de\ motifs\ M = \{ m_{1,} m_{2,} \ldots, m_k \}$$

$$Ensemble\ de\ descriptions\ de\ motifs\ DM = \{ desc_{1,} desc_{2,} \ldots, desc_k \}$$

$$Ensemble\ de\ demandes\ D = \{ d_{1,} d_{2,} \ldots, d_n \}$$

Contexte scientifique

Problème de classification de phrases courtes

$$f : D \rightarrow M$$

$$Ensemble\ de\ motifs\ M = \{ m_1, m_2, \ldots, m_k \}$$

$$Ensemble\ de\ descriptions\ de\ motifs\ DM = \{ desc_1, desc_2, \ldots, desc_k \}$$

$$Ensemble\ de\ demandes\ D = \{ d_1, d_2, \ldots, d_n \}$$

Approches étudiées

1. Recherche d'Information
   (approche non-supervisée)

2. Plongement de mots
   (approche non-supervisée)

3. Apprentissage faiblement supervisé

1. Recherche d'Information

TF-IDF

Pour toute demande d de D

Corpus = $\{d\} \cup M \cup DM$

Calculer la matrice **TF-IDF** sur Corpus

Sélectionner le motif de la demande

$$m_d = \nabla(Vect(d), Vect(m)), m \in M$$

1. Recherche d'Information

PMI

Pour toute demande d de D

Corpus = $\{d\} \cup M \cup DM$

Calculer la matrice **Terme-Terme** sur Corpus

Sélectionner le motif de la demande

$$m_d = \nabla(Vect(d), Vect(m)), m \in M$$

2. Plongement de mots

Modèle de langue	Corpus d'apprentissage	Dimension
CBOW	WaC	500
SkipGram	WaC	500
CBOW	Wikipédia	700
SkipGram	Wikipédia	1000
Fasttext	Wikipédia	300

$$m_d = \nabla\left(Vect(d), Vect(m)\right), m \in M$$

3. Approche supervisée

Modèle de classifieur : **Flair**

Modèle de langue : **CamemBERT**

Représentation de demande/motif : **FlairEmbeddings**

Corpus de train

1. Ensemble des descriptions des motifs
2. 180 demandes labellisées
3. Combinaison de (1) et (2)

Évaluations

Accuracy : $\frac{\textit{Nombre de demandes bien classées}}{\textit{Nombre total de demandes}}$

Validation croisée (K-Fold)

Résultats et analyse

Distance	TF-IDF	PPMI	Fasttext	CBOW		SkipGram	
				WaC	Wiki	WaC	Wiki
Euclidienne	44 %	5 %	31 %	32 %	26 %	27 %	27 %
Cosinus	44 %	5 %	27 %	44 %	55 %	39 %	37 %
Correlation	46 %	5 %	31 %	44 %	56 %	40 %	37 %

Résultats et analyse

Distance	TF-IDF	PPMI	Fasttext	CBOW		SkipGram	
				WaC	Wiki	WaC	Wiki
Euclidienne	44 %	5 %	31 %	32 %	26 %	27 %	27 %
Cosinus	44 %	5 %	27 %	44 %	55 %	39 %	37 %
Correlation	46 %	5 %	31 %	44 %	**56 %**	40 %	37 %

Résultats et analyse

Distance	TF-IDF	PPMI	Fasttext	CBOW		SkipGram	
				WaC	Wiki	WaC	Wiki
Euclidienne	44 %	5 %	31 %	32 %	26 %	27 %	27 %
Cosinus	44 %	5 %	27 %	44 %	55 %	39 %	37 %
Correlation	46 %	5 %	31 %	44 %	**56 %**	40 %	37 %

Corpus	Accuracy
1. Ensemble des descriptions des motifs	39 %
2. 180 demandes labellisées	95 %
3. Combinaison de (1) et (2)	95 %

Résultats et analyse

Distance	TF-IDF	PPMI	Fasttext	CBOW		SkipGram	
				WaC	Wiki	WaC	Wiki
Euclidienne	44 %	5 %	31 %	32 %	26 %	27 %	27 %
Cosinus	44 %	5 %	27 %	44 %	55 %	39 %	37 %
Correlation	46 %	5 %	31 %	44 %	**56 %**	40 %	37 %

Corpus	Accuracy
1. Ensemble des descriptions des motifs	39 %
2. 180 demandes labellisées	**95 %**
3. Combinaison de (1) et (2)	**95 %**

Conclusion

Approche non-supervisée basée sur une représentation pertinente des demandes et des motifs

Approche faiblement supervisée en exploitant les descriptions des motifs

Perspectives

Mettre en place une stratégie d'augmentation des données

Proposer une représentation appropriée à notre cas d'étude

Mettre en place un système d'apprentissage «online»

MERCI POUR VOTRE ATTENTION !