

Soutenance TER

Appariement Questions/Questions

Félix Jamet

30 mai 2018

Outline

- 1 Présentation du sujet
- 2 Approches explorées

Appariement de questions/questions

- Communautés de questions / réponses (AskUbuntu, StackExchange)
- Beaucoup de données (questions dupliquées)
- Intéressant de pouvoir évaluer la similarité de questions

- *Workshop*
- Évaluation de systèmes d'analyse de sémantique computationnelle
- Organisé en tâches

Tâche 3

- Similarité question / réponse
- Sous-tâche 3B : similarité question / question
- Données extraites du forum Qatar Living

Organisation des données

- N questions originales
- $N \times 10$ questions reliées
- Dans un fichier XML

- Fautes de frappe ou langage abrégé
- Parties non pertinentes à la sémantique (ex: remerciements)

Document

- Texte à analyser (en l'occurrence des questions)

Corpus

- Ensemble de documents

Sac de mots

- Approche d'analyse de document en tant que multi-ensemble de tokens

- *Term Frequency - Inverse Document Frequency*
- Donne une idée de l'importance d'un terme dans un document et dans un corpus

$$TF(\text{terme}, \text{document}) = \frac{\text{occurrences}(\text{terme}, \text{document})}{\text{taille}(\text{document})}$$

$$IDF(\text{terme}, \text{corpus}) = \log \left(\frac{\text{taille}(\text{corpus})}{\|\{\text{doc} / \text{doc} \in \text{corpus} \wedge \text{terme} \in \text{doc}\}\|} \right)$$

$$TF-IDF(\text{terme}, \text{document}, \text{corpus}) =$$

$$\begin{cases} TF(\text{terme}, \text{document}) \times IDF(\text{terme}, \text{corpus}) & \text{si } \text{terme} \in \text{corpus} \\ \max(\{IDF(el, \text{corpus}) / el \in \text{corpus}\}) & \text{sinon} \end{cases}$$

Méthode de référence

- Somme des valeurs TF-IDF des tokens du sac de mots
- Corpus : les questions originales et reliées mises bout à bout
- Document : les deux questions concaténées

similaritéRéf érence(Q_1, Q_2) =

$$\sum_{\text{terme} \in Q_1 \cap Q_2} TF\text{-}IDF(\text{terme}, Q_1 \cup Q_2, \text{corpus})$$

Méthode de référence - Scores

Édition	Méthode	Score MAP
2016	UH-PRHLT-contrastive2	77.33
2016	UH-PRHLT-primary	76.70
2016	UH-PRHLT-contrastive1	76.56
2016	<i>IR baseline</i>	74.75
2016	Référence	71.48
2017	KeLP-contrastive1	49.00
2017	SimBow-contrastive2	47.87
2017	SimBow-primary	47.22
2017	<i>IR baseline</i>	41.85
2017	Référence	44.21

Table: Scores SemEval 2016+2017

Méthode de référence avec filtres

- Intuitivement, les mots de faible longueur transportent peu de sens
- Il existe potentiellement des mots trop communs pour être intéressants (mots-outils)
- Une amélioration simple de la méthode de référence consiste à filtrer ces mots

corpus	# tokens tq. $\text{len}(\text{token}) > 4$	# tokens tq. $\text{len}(\text{token}) \leq 4$
2016	13552	31331
2017	19013	41787

Table: Nombre de mots de longueur inférieure et supérieure ou égale à 4

Méthode de référence avec filtres - Scores

Méthode	Score MAP
UH-PRHLT-contrastive2	77.33
UH-PRHLT-primary	76.70
UH-PRHLT-contrastive1	76.56
Mots outils, ≤ 1	75.42
Mots outils, ≤ 2	75.04
<i>IR baseline</i>	74.75
≤ 1	74.58
≤ 3	74.42
Mots outils, ≤ 4	74.21
≤ 4	74.06
Mots outils, ≤ 3	73.97
≤ 2	73.87
Mots outils	73.76
Référence	71.48

Table: Scores SemEval 2016

Méthodes	Score MAP
KeLP-contrastive1	49.00
SimBow-contrastive2	47.87
SimBow-primary	47.22
≤ 1	46.89
Mots outils, ≤ 1	46.35
Mots outils, ≤ 2	46.08
≤ 2	46.07
≤ 3	45.59
Mots outils	45.53
Mots outils, ≤ 3	45.46
Référence	44.21
<i>IR baseline</i>	41.85
Mots outils, ≤ 4	41.80
≤ 4	40.47

Table: Scores SemEval 2017

Comparaison pratique avec et sans filtres (2016)

Question 387

Score AP = 0.1

" Mall of Asia in Qatar soon to open ? . " " Is it true that there is Mall of Asia opening in Doha ; Qatar? .. If yes? .. Then ; is it in justice if I 'll will just receive 1000riyal monthly?excluding the commission . "

Question 387 après filtrage

Score AP = 1

Mall Asia Qatar open Is true Mall Asia opening Doha Qatar? .. If yes? .. Then justice 'll receive 1000riyal monthly?excluding commission

Filtres appliqués : mots-outils et mots de longueur 1

Comparaison pratique avec et sans filtres (2017)

Question 448

Score AP = 0.5

household clearance companies . Anyone know of a company that can clear your house of furniture / appliances / misc when leaving Doha ?

Question 448 après filtrage

Score AP = 1

household clearance companies Anyone know of company that can clear your house of furniture appliances misc when leaving Doha

Filtre appliqué: mots de longueur 1