Soutenance TER

Appariement Questions/Questions

Félix Jamet

30 mai 2018

Outline

Appariement de questions/questions

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

SemEval

- *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 Quatar Living

Organisation des données

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

Difficultés

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

Vocabulaire

Document

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

Corpus

- Ensemble de documents

Sac de mots

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

TF-IDF

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

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

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

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

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

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 contaténées

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

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	*IR baseline*	74.75
2016	**Référence**	**71.48**
2017	KeLP-contrastive1	49.00
2017	SimBow-contrastive2	47.87
2017	SimBow-primary	47.22
2017	*IR baseline*	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. len(token) $>$ 4	# tokens tq. len(token) $<=$ 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
IR baseline	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**
IR baseline	41.85
Mots outils, ≤ 4	41.80
≤ 4	40.47

Table: Scores SemEval 2017