Soutenance TER

Appariement de questions/questions

Félix Jamet
Encadré par Amir Hazem

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
  - Attribut : pertinence vis-à-vis de la question originale
- Dans un fichier XML

Année	Nombre de questions originales
2016	70
2017	88

But

ordonner les questions selon leur pertinence

Average Precision (AP)

- Associe un score à une liste ordonnée de documents

Formule AP

$$AP = \frac{1}{|R|} \times \sum_{i=1}^{n} Precision(i) \times Pertinence(i)$$

- $|R|$: nombre total de documents pertinents
- $Precision(i)$: précision au rang i (Proportion de documents pertinents dans les i premiers rangs)
- $Pertinence(i)$: $\begin{cases} 1 & \text{si pertinent} \\ 0 & \text{sinon} \end{cases}$

Exemples *Average Precision*

- $AP(Vrai, Faux, Faux) = 1$
- $AP(Faux, Vrai) = 0.5$
- $AP(Faux, Faux, Vrai) = \frac{1}{3}$
- $AP(Vrai, Faux, Faux, Faux, Vrai) = \frac{1}{2} \times (1 + 0 + 0 + 0 + \frac{2}{5}) = 0.7$

Mean Average Precision (MAP)

- La moyenne de l'*Average Precision*
- En l'occurence la moyenne des *Average Precision* des N questions originales

Vocabulaire

Définition (Document)

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

Définition (Corpus)

- Ensemble de documents

Définition (Token)

- Unité lexicale extraite d'une phrase

Définition (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 communs au sac de mots
- Corpus : toutes les questions
- 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	**Référence**	**44.21**
2017	*IR baseline*	41.85

Table: Scores SemEval 2016 et 2017 - Référence

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 - Filtres

Méthode	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 - Filtres

Comparaison de la tokenisation avec et sans filtres

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

Lemmes

Définition (Lemme)

- Forme canonique d'un mot
- Permet de regrouper des mots d'une même famille

Exemple

- cherchera $\rightarrow$ chercher
- chercherons $\rightarrow$ chercher

- La sémantique est conservée
- Des termes proches vont prendre une forme commune

Lemmes - Scores

Méthode	Score MAP
UH-PRHLT-contrastive2	77.33
UH-PRHLT-primary	76.70
UH-PRHLT-contrastive1	76.56
Lemmes, Mots outils, ≤ 2	76.48
Lemmes, Mots outils, ≤ 3	75.87
Lemmes, Mots outils, ≤ 1	75.56
Lemmes, Mots outils, ≤ 4	75.38
Lemmes, ≤ 4	75.31
IR baseline	74.75
Lemmes, ≤ 1	73.64
Lemmes, ≤ 2	73.38
Lemmes	73.38
Lemmes, ≤ 3	72.95
Lemmes, Mots outils	72.14
Référence	**71.48**

Table: Scores SemEval 2016 - Lemmes

Méthode	Score MAP
KeLP-contrastive1	49.00
SimBow-contrastive2	47.87
Lemmes, Mots outils, ≤ 1	47.70
SimBow-primary	47.22
Lemmes, Mots outils, ≤ 2	46.61
Lemmes, Mots outils, ≤ 3	46.16
Lemmes, ≤ 1	45.92
Lemmes	45.82
Lemmes, ≤ 3	45.17
Lemmes, Mots outils	44.23
Référence	**44.21**
Lemmes, ≤ 2	42.82
Lemmes, Mots outils, ≤ 4	41.87
IR baseline	41.85
Lemmes, ≤ 4	41.16

Table: Scores SemEval 2017 - Lemmes

Nature grammaticale

- Hypothèse : la nature grammaticale d'un mot a une influence sur son importance sémantique
- Approche : appliquer une pondération sur les noms (0.52)

Scores 2016

Filtres	Score MAP
Mots outils, ≤ 2	76.48
Mots outils, ≤ 3	75.87
Mots outils, ≤ 1	75.56
Mots outils, ≤ 4	75.38
≤ 4	75.31
≤ 1	73.64
≤ 2	73.38
Pas de filtre	73.38
≤ 3	72.95
Mots outils	72.14

Table: Scores SemEval 2016 - Lemmes

Filtres	Score MAP
Mots outils, ≤ 2	76.61
Mots outils, ≤ 1	76.32
Mots outils, ≤ 3	75.97
Mots outils, ≤ 4	75.32
≤ 4	75.30
Pas de filtre	73.73
≤ 1	73.45
≤ 2	73.04
≤ 3	72.97
Mots outils	72.07

Table: Scores SemEval 2016 - Lemmes et pondération

Scores 2017

Filtres	Score MAP
Mots outils, ≤ 1	47.70
Mots outils, ≤ 2	46.61
Mots outils, ≤ 3	46.16
≤ 1	45.92
Pas de filtre	45.82
≤ 3	45.17
Mots outils	44.23
≤ 2	42.82
Mots outils, ≤ 4	41.87
≤ 4	41.16

Table: Scores SemEval 2017 - Lemmes

Filtres	Score MAP
Mots outils, ≤ 1	47.81
Mots outils, ≤ 2	46.63
≤ 1	45.97
Mots outils, ≤ 3	45.66
Pas de filtre	45.57
≤ 3	45.09
Mots outils	44.04
≤ 2	43.59
Mots outils, ≤ 4	42.02
≤ 4	41.27

Table: Scores SemEval 2017 - Lemmes et pondération

Conclusion

Difficultés

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

Perspectives

- Construction d'une liste de mots outils spécifique au corpus
- Utilisation d'un dictionnaire de synonymes
- Utilisation d'une distance d'édition pour contrebalancer les fautes
- Arbres de décision