

# Soutenance TER

Appariement de questions/questions

Félix Jamet  
Encadré par Amir Hazem

30 mai 2018

## 1 Présentation du sujet

- Appariement de questions/questions
- SemEval
- Mesures

## 2 Approches explorées

- Méthode de référence
- Filtres
- Lemmatisation
- Nature grammaticale

## 3 Conclusion

# Table des matières

## 1 Présentation du sujet

- Appariement de questions/questions
- SemEval
- Mesures

## 2 Approches explorées

- Méthode de référence
- Filtres
- Lemmatisation
- Nature grammaticale

## 3 Conclusion

# 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
  - 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*

$$\begin{array}{lcccccc} \text{Pertinence(i)} & 1 & 0 & 0 & 0 & 0 \\ \text{Précision(i)} & 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ AP & = 1 \times 1 = 1 \end{array}$$

$$\begin{array}{lcccccc} \text{Pertinence(i)} & 0 & 0 & 0 & 0 & 1 \\ \text{Précision(i)} & 0 & 0 & 0 & 0 & \frac{1}{5} \\ AP & = 1 \times \frac{1}{5} = \frac{1}{5} = 0.2 \end{array}$$

$$\begin{array}{lcccccc} \text{Pertinence(i)} & 1 & 0 & 0 & 0 & 1 \\ \text{Précision(i)} & 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{2}{5} \\ AP & = \frac{1}{2} \times (1 + \frac{2}{5}) = 0.7 \end{array}$$

# Mean Average Precision (MAP)

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

# Table des matières

## 1 Présentation du sujet

- Appariement de questions/questions
- SemEval
- Mesures

## 2 Approches explorées

- Méthode de référence
- Filtres
- Lemmatisation
- Nature grammaticale

## 3 Conclusion

## Définition (Document)

- Texte à analyser

## Définition (Corpus)

- Ensemble de documents

## Définition (Tokenisation)

- Procédé de séparation d'un document en unités appelées tokens

## Définition (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

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

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

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

# Méthode de référence

- Somme des valeurs TF-IDF des tokens communs au sac de mots
- Corpus : toutes les questions
- Document : les deux questions concaténées

$similaritéRéférence(Q_1, Q_2) =$

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

# 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	<i>IR baseline</i>	74.75
2016	<b>Référence</b>	<b>71.48</b>
2017	KeLP-contrastive1	49.00
2017	SimBow-contrastive2	47.87
2017	SimBow-primary	47.22
2017	<b>Référence</b>	<b>44.21</b>
2017	<i>IR baseline</i>	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	# mots tq. $\text{len}(\text{mot}) > 4$	# mots tq. $\text{len}(\text{mot}) \leq 4$
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, $\leq 1$	75.42
Mots outils, $\leq 2$	75.04
<i>IR baseline</i>	74.75
$\leq 1$	74.58
$\leq 3$	74.42
Mots outils, $\leq 4$	74.21
$\leq 4$	74.06
Mots outils, $\leq 3$	73.97
$\leq 2$	73.87
Mots outils	73.76
<b>Référence</b>	<b>71.48</b>

Table: Scores SemEval 2016 - Filtres

Méthode	Score MAP
KeLP-contrastive1	49.00
SimBow-contrastive2	47.87
SimBow-primary	47.22
$\leq 1$	46.89
Mots outils, $\leq 1$	46.35
Mots outils, $\leq 2$	46.08
$\leq 2$	46.07
$\leq 3$	45.59
Mots outils	45.53
Mots outils, $\leq 3$	45.46
<b>Référence</b>	<b>44.21</b>
<i>IR baseline</i>	41.85
Mots outils, $\leq 4$	41.80
$\leq 4$	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 true Mall Asia opening Doha Qatar? .. yes? .. justice 'll receive 1000riyal monthly?excluding commission

Filtres appliqués : mots-outils et mots de longueur 1

## Définition (Lemme)

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

## Exemple

- cherchera → chercher
  - chercherons → 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, $\leq 2$	76.48
Lemmes, Mots outils, $\leq 3$	75.87
Lemmes, Mots outils, $\leq 1$	75.56
<b>Mots outils, <math>\leq 1</math></b>	<b>75.42</b>
Lemmes, Mots outils, $\leq 4$	75.38
Lemmes, $\leq 4$	75.31
<i>IR baseline</i>	74.75
Lemmes, $\leq 1$	73.64
Lemmes, $\leq 2$	73.38
Lemmes	73.38
Lemmes, $\leq 3$	72.95
Lemmes, Mots outils	72.14
<b>Référence</b>	<b>71.48</b>

Table: Scores SemEval 2016 - Lemmes

Méthode	Score MAP
KeLP-contrastive1	49.00
SimBow-contrastive2	47.87
Lemmes, Mots outils, $\leq 1$	47.70
SimBow-primary	47.22
<b><math>\leq 1</math></b>	<b>46.89</b>
Lemmes, Mots outils, $\leq 2$	46.61
Lemmes, Mots outils, $\leq 3$	46.16
Lemmes, $\leq 1$	45.92
Lemmes	45.82
Lemmes, $\leq 3$	45.17
Lemmes, Mots outils	44.23
<b>Référence</b>	<b>44.21</b>
Lemmes, $\leq 2$	42.82
Lemmes, Mots outils, $\leq 4$	41.87
<i>IR baseline</i>	41.85
Lemmes, $\leq 4$	41.16

Table: Scores SemEval 2017 - Lemmes

- 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, $\leq 2$	76.48
Mots outils, $\leq 3$	75.87
Mots outils, $\leq 1$	75.56
Mots outils, $\leq 4$	75.38
$\leq 4$	75.31
$\leq 1$	73.64
$\leq 2$	73.38
Pas de filtre	73.38
$\leq 3$	72.95
Mots outils	72.14

Table: Scores SemEval 2016 -  
Lemmes

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

Table: Scores SemEval 2016 -  
Lemmes et pondération

# Scores 2017

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

Table: Scores SemEval 2017 -  
Lemmes

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

Table: Scores SemEval 2017 -  
Lemmes et pondération

# Table des matières

## 1 Présentation du sujet

- Appariement de questions/questions
- SemEval
- Mesures

## 2 Approches explorées

- Méthode de référence
- Filtres
- Lemmatisation
- Nature grammaticale

## 3 Conclusion



# Conclusion

- Possible d'appareiller des questions avec des caractéristiques linguistiques simples
- Approche perfectible

- 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
- Tester sur d'autres corpus
- Arbres de décision

Questions ?