Indexation et appariements de documents cliniques pour le Deft 2019

Davide Buscaldi[1], Dhaou Ghoul[2], Joseph Le Roux[1], Gaël Lejeune[2]

2 juillet 2019

(1) LIPN UMR 7030, Université Paris XIII firstname.lastname@univ-paris13.fr
(2) STIH EA 4509, Sorbonne Université firstname.lastname@sorbonne-universite.fr

Plan de la présentation

1. Introduction
2. Appariement
3. Indexation
4. Résultats
5. Conclusion

Introduction

Introduction

Les trois tâches du Défi 2019 :

- **Indexation**
- **Appariement**
- Extraction d'information

Appariement

Appariement

Deux méthodes :

- Distributions de chaînes de caractères (run1)
- Réseau siamois avec algorithme hongrois (run2 et run3)

Appariement fondé sur les affinités (I)

Hypothèse : exploiter les phénomènes de recopie

[Lejeune et al., 2011] : l'association entre un résumé et un article est détectable par des "affinités"

Affinité : chaîne de caractères commune à une discussion et un cas clinique mais hapax dans la sous-collection des résumés

On apparie un prétendant P_i (cas clinique) et un célibataire C (discussion) sous deux conditions :

- Card-Aff : P_i partage le plus d'affinités avec C
- Max-Aff : P_i et C partagent la plus longue affinité détectée

Appariement fondé sur les affinités (II)

Différence avec le Deft 2011 :

- la même discussion peut correspondre à plusieurs cas cliniques
- le phénomène de recopie est moins présent (genre moins normé)

Quelques affinités :

- " tous les critères d'Amsterdam, seules les mutations "
- "permis une amélioration clinique"
- ". La tomodensitométrie abdominale"
- " laparotomie médiane sous ombilicale"

`https://github.com/rundimeco/deft2011`

Appariement fondé sur les réseaux siamois (I)

[Bromley et al., 1993] : architectures neuronales spécialisées pour l'appariement de structures similaires (ex : signatures de chèques)

Deux sous-réseaux identiques : des vecteurs d'entrée vers un espace de caractéristiques commun. La dernière couche calcule une **énergie**, censée représenter la proximité entre les deux structures.

Plus concrètement :

1. Filtrage via SpaCY pour ne garder que les noms communs
2. Pour chaque document, un vecteur représentatif de la moyenne des vecteurs associés à ses mots après filtrage
3. Proximité entre d_1 et d_2 $\rightarrow$ en entrée du sous-réseau du réseau siamois
4. Pour la prédiction : distance euclidienne entre v_1 et v_2.
5. $\rightarrow$ couplage parfait de poids minimal dans $G = (V_1 \cup V_2, V_1 \times V_2)$
6. Poids des arcs (v_1, v_2) : distance euclidienne (algorithme hongrois [Munkres, 1957])

Appariement fondé sur les réseaux siamois (II)

Energie contrastive pour discriminer les paires de structures similaires et dissimilaires [Chopra et al., 2005]

Une série de triplets (d_1, d_2, l) avec deux documents d_1, d_2 et un label $l \in \{0, 1\}$ (similaire ou non)

Label 0 : documents identiques, appariés dans le $\textsc{Train}$ ou associés à la même discussion

Indexation

L'indexation

Deux méthodes exploitées :

- Indexation fondée sur des appariements de documents
- Indexation fondée sur l'annotation terminologique

Indexation par Appariement (I)

Hypothèse : des documents proches ont des mots-clés en commun.

Un cas clinique et la discussion qui s'y rapporte doivent avoir une indexation très proche

Indexer un cas clinique C (resp. une discussion D) revient à

- l'apparier avec une discussion D' (resp. un cas clinique C')
- lui assigner les mots-clés correspondants
- fusionner les listes de mots-clés ainsi trouvées

Indexation par Appariement (II)

- Pour chaque document à indexer (C ou D)
  - On considère que l'on a une paire à indexer (C_i, D_i) :
    - on calcule un appariement (D_j, C_j)
    - on en déduit jeux de mots-clés candidats : KW_{Di} et KW_{Cj}.
    - conserve l'intersection : $KW_{Di} \cap KW_{Cj}$
    - si la taille de l'intersection est inférieure au nombre de mots-clés attendus on utilise l'union : $KW_{Di} \cup KW_{Cj}$
- on les ordonne par longueur en caractères décroissante

Indexation par Appariement (II)

- Pour chaque document à indexer (C ou D)
  - On considère que l'on a une paire à indexer (C_i, D_i) :
    - on calcule un appariement (D_j, C_j)
    - on en déduit jeux de mots-clés candidats : KW_{Di} et KW_{Cj}.
    - conserve l'intersection : $KW_{Di} \cap KW_{Cj}$
    - si la taille de l'intersection est inférieure au nombre de mots-clés attendus on utilise l'union : $KW_{Di} \cup KW_{Cj}$
- on les ordonne par longueur en caractères décroissante

Echec critique

Efficacité significativement inférieure à une simple *baseline* vérifiant la présence des mots-clés de la référence

Indexation par Annotation terminologique (I)

[Buscaldi and Zargayouna, 2016] : combiner annotation terminologique (précision) et annotation supervisée (rappel)

- Identifier des concepts du MESH
- Donner aux concepts associés un poids
- Compléter avec l'Information Mutuelle Ponctuelle (PMI) quand les concepts n'ont pas de représentations lexicales dans le texte
- Filtrer en prenat seulement en compteles concepts qui ont servi pour annoter au moins 3 documents dans le corpus de référence

Indexation par Annotation terminologique (II)

Une matrice M avec $|C|$ lignes (étiquettes) et $|T|$ colonnes (mots pleins), où chaque élément $M_{i,j}$ est calculé comme suit :

$$M_{i,j} = \begin{cases} PMI(t_i, c_j) & if \;\; PMI(t_i, c_j) > 0 \\ else 0 \end{cases}$$

LSA avec décomposition M en valeurs singulières (SVD) $M = U\Sigma V^T$, et en approximant M avec $\hat{M} = U_k \Sigma_k V_k^T$, avec les meilleurs k valeurs singulières sélectionnés ($k = 100$).

Annotation finale

Pour chaque document d, on a un vecteur binaire $\mathbf{b}$ de taille $|T|$ (taille du vocabulaire) où chaque élément $\mathbf{b}_i$ est à 1 si le terme correspondant est dans le texte du document d, 0 autrement.

Résultats

Résultats et discussion : Indexation

Run	MAP	R-Precision
Run1 (Appariements)	0.126	0.122
Run2 (Baseline)	0.220	0.240
Run4 (MeSH)	0.044	0.034

Table 1: Résultats officiels sur la tâche 1

- Deux hypothèses inadaptées :
- appariement de documents (Run1)
- annotation terminologique à partir du MeSH (Run4)

Résultats et discussion : Appariements

Run	Précision
Run1 (Similarité en caractères)	0.617
Run2 (Réseau Siamois *average*)	0.107
Run3 (Réseau Siamois *single*)	0.126

Table 2: Résultats officiels sur la tâche 2

- Réseaux siamois, variante *average* (modèle moyenné) et variante *single* (meilleure itération) ont très vite plafonné
- Modèle fondé sur les affinités : peu efficace dans le contexte du concours mais non-supervisé

Conclusion

Résultats et discussion (III)

- Résultats décevants . . . mais attendus
- Tâche 1 : 16 pp. sous la moyenne et 18 sous la médiane.
- Tâche 2 : plus éloigné en pp. de la moyenne (19 points) comme de la médiane (25 points)
- Comment améliorer sans en dénaturer l'esprit ?

Bibliographie


Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säackinger, E., and Shah, R. (1993).
Signature verification using a siamese time delay neural network.
International Journal of Pattern Recognition and Artificial Intelligence, 7(04):669–688.


Buscaldi, D. and Zargayouna, H. (2016).
LIPN@DEFT2016 : Annotation de documents en utilisant l'Information Mutuelle.
In *DÉfi Fouille de Texte 2016 – DEFT2016*, Paris, France.


Chopra, S., Hadsell, R., LeCun, Y., et al. (2005).
Learning a similarity metric discriminatively, with application to face verification.
In *CVPR (1)*, pages 539–546.


Lejeune, G., Brixtel, R., Giguet, E., and Lucas, N. (2011).
Deft 2011: appariements de résumés et d'articles scientfiques fondés sur des distributions de chaînes de caractères.
In *Proceedings of DEfi Fouille de Texte (DEFT'11)*, pages 53–64.


Munkres, J. (1957).
Algorithms for the assignment and transportation problems.
Journal of the society for industrial and applied mathematics, 5(1):32–38.