GREYC@TEXTMINE2023 : RECONNAISSANCE D'ENTITÉS NOMMÉES DANS LES SIGNATURES D'E-MAILS

Tanguy Gernot et Emmanuel Giguet[1]

[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE

emmanuel.giguet@unicaen.fr, tanguy.gernot@unicaen.fr

GREYC
Laboratoire de recherche en sciences du numérique

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Objectif du défi TextMine'2023 :

- Reconnaître 13 types d'entité dans des signatures d'e-mails : personne, organisation, fonction, projet, lieu, CEDEX, CS, code postal, numéro de téléphone, e-mail
- Structurer l'information ; La stocker en base de données

3 jeux de données :

- 2 jeux d'entraînement : Réaliste (JDR) et Factice (JDF)
- 1 jeu de validation : Authentique (JDA)
- Plusieurs langues, français majoritaire ; format texte brut

La mission :

- Analyser chaque signature du jeu de données
- Associer un type d'entité à une liste de tokens attendus

Table: Évaluation sur le jeu d'entraînement JDR

	precision	recall	f1-score	support
micro avg				
macro avg	0.98	0.98	0.98	6691
weighted avg	0.99	0.99	0.99	6691
F1	0.9912928493115661			

Table: Évaluation sur le jeu d'entraînement JDF

	precision	recall	f1-score	support
micro avg	1.00	1.00	1.00	5225
macro avg	0.46	0.46	0.46	5225
weighted avg	1.00	1.00	1.00	5225
F1	0.9997124612816258			

Table: Évaluation sur le jeu de validation JDA

	precision	recall	f1-score	support
micro avg	0.8243	0.8241	0.8242	8888
macro avg	0.7305	0.7249	0.7220	8888
weighted avg	0.8414	0.8241	0.8312	8888
F1	0.8311914814126119			

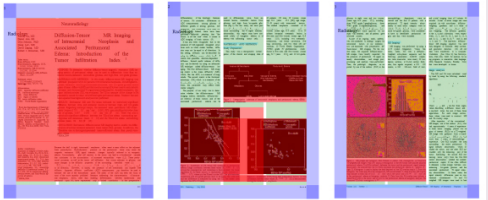
- Comment avons nous abordé le défi ?
- Pourquoi un écart de 17 points entre mise au point et validation ?
- Quelles pistes d'amélioration ?

D’où venons-nous ?

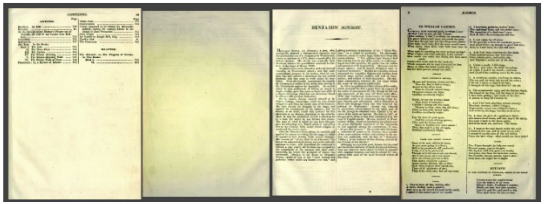
Équipe de recherche en cybersécurité, laboratoire GREYC UMR 6072, Caen, France

Contexte : Structuration automatique des documents

RÉSURGENCE: Analyse de la structure d'articles biomédicaux (2005)

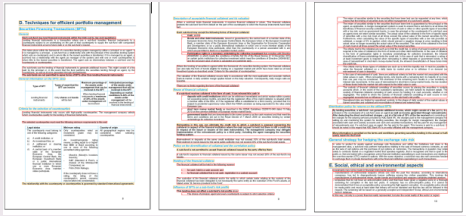


Participations aux défis INEX/ICDAR "Book Structure Extraction" (2009-2011)



Contexte: Structuration automatique des documents

Challenge FinToc: Structure des documents financiers (2019-2022)



X-COTE – Extraction de Contenus Textuels du Web (TALN-RECITAL'2021)

- Identification du contenu principal des pages Web : "Web Scraping"
- Macro-segmentation du contenu des pages Web

Contexte: Lutte contre le phishing

GREYC Safe : Équipe de recherche en cybersécurité

- Axe de recherche en investigation numérique
- Construction d'une plateforme logicielle pour l'investigation numérique
  - Évaluation de technologies existantes
  - Conception de composants à base de machine ou deep-learning
  - Analyse de vidéos, d'images, de textes
- Projet interne de "Lutte contre le phishing et les courriers malveillants"
- Montage en cours d'une thèse sur la détection des mails de phishing
  - Cohérence structurelle et textuelles
  - Macro-segmentation et extraction d'information

- Défi TextMine'2023 : une opportunité à saisir

Notre Approche

Une démarche commune à toutes nos participations

- Concevoir notre propre chaîne d'analyse de bout en bout
  - $\Rightarrow$ Maîtriser l'intégralité de la chaîne de traitement

La démarche pour le défi TextMine

- Traiter les énoncés textuels fournis au format brut
- Réaliser notre propre segmentation des signatures
- Procéder à l'identification des entités dans notre format
- Produire un produit dérivé : un format de sortie TextMine

La segmentation des signatures (1/2)

Penser une segmentation en constituants logiques

- Viser une macro segmentation en constituants logiques :
  - personne, fonction, organisation, adresse, téléphone ...
- Utiliser la catégorie des constituants pour catégoriser les tokens internes
- Catégoriser les constituants par une analyse morpho-syntaxique
  - analyse morphologique
  - analyse contextuelle

Recourir à une segmentation en constituants physiques

- 2 types de constituant physique : phrastique et intra-phrastique
  - Délimiteurs dits ”phrastiques” : saut de ligne
  - Délimiteurs intra-phrastiques : tiret long, barre oblique, ...
  - Présupposé : Segment intra-phrastique = unité minimale

La segmentation des signatures (2/2)

Intérêt : Disposer d'une hiérarchie de constituants

- Se donner la possibilité de travailler à différents niveaux
- Faciliter les déductions contextuelles
- Réduire le nombre d'unités à considérer : "diviser pour régner"

La catégorisation des constituants

Considérer 3 grands types de catégorie

1. Morphologiquement non ambigus : e-mail, numéro de téléphone, site internet
   - Identifiés à l'aide d'expressions régulières
2. Morphologiquement incertains (indices partiels) : fonction, adresse
   - Identifiés à l'aide d'expressions régulières partielles
     + de lexiques spécialisés (liste de fonctions, types de voie)
     + de liste de suffixes caractéristiques (-ogue, -iste)
3. Non catégorisables morphologiquement : personne, organisation, projet

Catégorisation contextuelle et positionnelle

- A base de règles contextuelles :
  - Nom du signataire : En début de signature = position 1 + présence obligatoire
  - Fonction : Position intermédiaire entre Nom et Organisation
  - Organisation : Obligatoire, à chercher dans les constituants non catégorisés

Les déductions par calcul de la référence

Calculer les chaînes de coréférence

- Idée : utiliser le fait qu'une signature à une cohérence textuelle
  - Mis en œuvre pour la détection du nom du signataire et du nom de l'organisation
- Calcul d'une distance (Jaccard) entre chaînes de caractères normalisées :
  - suppression des espaces, ponctuations, caractères spéciaux, accents, bascule en casse minuscule
- Coréférence sur nom de signataire :
  - Recherche des similarités graphiques entre le "local-name" d'une adresse mail (local-name@domain.tld) et un segment de l'adresse, idem pour les identifiants de réseaux sociaux
- Coréférence sur nom d'organisation :
  - Recherche des similarités graphiques entre le "domain-name" d'une adresse mail, d'un site internet et un segment de l'adresse

La production des fichiers attendus

Alignement des segmentations

- Jointure entre notre segmentation en tokens et les tokens attendus

Contrôle qualité

- Vérification de la bonne formation et de la bonne indexation des tokens,
- Vérification et réétiquetage hors-contexte des catégories sures : CEDEX, e-mail, téléphone, nom de domaine),
- Calcul des statistiques de bonne couverture,
- Calcul P/R/F1-score pour les jeux d'entraînement

Résultats

Évaluation

- 99% à 100% de F1-score sur les données de mise au point
- 83% de F1-score sur les données de validation
- Toutes les catégories ne se valent pas :
  - certaines sont plus faciles que d'autres : URL, Téléphone, CEDEX
- La granularité de l'étiquetage sur-pénalise les erreurs :

Explication de la baisse de performance

- Le jeu de validation est très différent des jeux de mise au point
- La segmentation en macro-constituants n'est pas assez robuste sur le corpus de validation qui ne préserve pas la mise en forme originale
- La pseudonymisation affecte la cohérence textuelle et perturbe le calcul des chaines de coréférence et les critères d'unicité

Conclusion

- Une participation très enrichissante,
- Des résultats très encourageants et améliorables significativement,
- Des résultats explicables et interprétables,
- Une gestion de plusieurs types d'unités pour raisonner à différents niveaux,
- Une utilisation originale du calcul des chaînes de coréférence,
- La mise en évidence des effets négatifs d'une pseudonymisation trop directe.

Nos sincères remerciements au comité d'organisation.

The end

[thank you]

Any Questions ?

emmanuel.giguet@unicaen.fr, tanguy.gernot@unicaen.fr