GREYC@TextMine2023 : Reconnaissance d'entités nommées dans les signatures d'e-mails

Tanguy Gernot, Emmanuel Giguet

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen, France
{prenom.nom}@unicaen.fr

Résumé. Cet article présente notre contribution au défi TextMine'23 portant sur la "Reconnaissance d'entités d'intérêts dans les signatures d'e-mails". La performance de notre système atteint 99% et 100% en f-mesure sur les jeux de données de mise au point et 83% de f-mesure sur le jeu de validation.

1 Introduction

Le défi TextMine'23 (Cousot et al., 2023) "Reconnaissance d'entités d'intérêts dans les signatures d'e-mails" est une initiative du groupe de travail TextMine dont l'objectif est de confronter l'état de l'art scientifique aux problèmes d'analyse de données textuelles rencontrés par les industriels. TextMine'23 est centré sur la reconnaissance d'entités d'intérêts dans les signatures d'e-mails dans le but de structurer l'information et de la stocker en base de données.

Suivant la tradition du traitement automatique des langues et des documents de notre équipe (Giguet et Lucas, 2022 ; Giguet et Lejeune, 2021 ; Giguet et Lucas, 2010), nous avons choisi de réaliser une chaîne de traitement complète, depuis la segmentation des signatures jusqu'à la production du fichier d'annotations au format attendu. Notre participation nous a permis d'explorer la dimension multilingue de l'analyse de signature, la sélection des unités d'analyse pertinentes, ainsi que le calcul original de chaîne de coréférence. La performance de notre système atteint 99% et 100% en f-mesure sur les jeux de données de mise au point et 83% de f-mesure sur le jeu de validation.

2 Travaux précédents

L'analyse des e-mails est sujet de recherches depuis le milieu des années 90. La question de la classification automatique a notamment fait l'objet de nombreux travaux, que ce soit pour la détection de courriers indésirables, le routage automatique des courriers entrants vers le bon interlocuteur, ou le classement thématique dans une arborescence. De nombreuses synthèses sont disponibles et la question de la segmentation des e-mails est une question régulièrement soulevée.

Concernant l'identification de la signature des e-mails et de leur analyse en constituants, (Bendahman et al., 2022) propose un état de l'art mettant en avant les travaux de (Chen et al., 1999 ; Carvalho et Cohen, 2004 ; Estival, 2008 ; Li et al., 2015). L'identification de la signature y est réalisée soit dans une perspective d'extraction d'informations ciblées, soit dans une perspective d'élimination de son contenu pour ne pas perturber l'analyse du corps de l'e-mail.

Après avoir constitué de manière participative un corpus de signatures d'e-mails et effectué leur pseudonymisation, (Bendahman et al., 2022) effectuent un comparatif d'algorithme d'apprentissage automatique (SVM, CRF, Bi-LSTM, Bert) pour la classification des tokens des signatures. Le modèle des CRF obtient le meilleur F-score : 79%.

3 Jeux de données

Deux jeux de données annotés ont été fournis aux participants pendant la phase de mise au point :

— le jeu de données réaliste (JDR), composé de 473 signatures, produit par la société Isahit en lien avec le comité d'organisation du défi ;

— le jeu de données factice (JDF), composé de 606 signatures créées automatiquement, par `fakenamegenerator.com`, une API de création de fausses identités. Certaines étiquettes ne sont cependant pas représentées.

Un jeu de données authentique (JDA), sans annotation, a été fourni pour valider notre système en fin de défi. Les signatures de ce jeu ont été recueillies par un formulaire Web permettant le "don" de signatures qui ont ensuite été pseudonymisées.

Les trois corpus ont été annotés avec le jeu de 13 étiquettes présenté en table 1. Les caractéristiques des trois jeux de données sont présentées en table 2.

Pour chaque signature, on trouve un identifiant `identifier`, un champ textuel `text` qui fournit le texte de la signature, sans mise en forme riche, ainsi que la liste des annotations à produire `annotations`. Pour chaque annotation attendue, on trouve la graphie `form`, son étiquette `label`, l'index `begin` de son premier caractère, l'index `end` du caractère suivant le token.

Dans les jeux de mise au point JDR et JDF le champ `label` est fourni. Dans le jeu de validation JDA, ce champ est absent puisqu'il doit être produit par le système automatique.

4 Méthode

Suivant la tradition du traitement automatique des langues et des documents de notre équipe, nous avons choisi de réaliser une chaîne de traitement complète, depuis la segmentation des signatures jusqu'à la production du fichier d'annotations au format attendu.

TAB. 1 – *Définition du jeu d'étiquettes*

Étiquette	Définition
Human	Identité des personnes qui figurent dans la signature
Organization	Organisation à laquelle l'auteur de la signature est rattachée
Function	Ensemble des fonctions assignées à la personne identifiée dans la signature
Project	Projet dans lequel la personne est impliquée
Location	Bâtiments, bureaux, villes, numéros et noms des rues
Reference_CEDEX	Courrier d'entreprise à distribution exceptionnelle
Reference_CS	Course spéciale
Reference_Code_Postal	Code postal
Phone_Number	Numéro de téléphone ou de fax
Email	Adresses e-mails
Url	URL de site web
Social_Network	Nom des réseaux sociaux
Reference_User	Identifiant d'une personne ou d'une organisation sur un réseau social

Contrairement aux approches automatiques reposant sur la catégorisation d'une suite de tokens, nous avons choisi de réaliser une segmentation des énoncés en constituants, puis de catégoriser chaque constituant. L'étiquetage des mots, attendu dans le défi, a alors consisté à projeter l'étiquette de chacun des constituants sur les mots qui le constituent.

Cette approche qui fait intervenir des unités intermédiaires nous semble particulièrement intéressantes puisqu'elle réduit considérablement le nombre d'unités à traiter et limite la taille de la fenêtre d'observation pour les déductions contextuelles basées sur les unités situées avant ou après l'unité traitée. Cette stratégie peut être vue comme une approche de type *diviser pour régner*.

Bien entendu, une segmentation en constituants logiques ne peut être obtenue d'emblée et une segmentation approximative, en constituants graphiques, est utilisée. Ainsi, nous utilisons deux classes de délimiteurs : les délimiteurs phrastiques de type "saut de ligne", et les délimiteurs intraphrastiques, de type "tiret", "tiret-long", "barre oblique", "puce".

Le typage des constituants graphiques ainsi obtenus est réalisé par analyse morphologique, et analyse contextuelle. Les étiquettes sont divisées en trois classes : les étiquettes attribuées à des contenus peu ambigus et ne nécessitant que peu d'analyse contextuelle (adresse e-mail, numéro de téléphone, nom de domaine, réseaux sociaux), les étiquettes qui pour être posées nécessitent des ressources linguistiques (adresse, fonction), et les étiquettes nécessitant une analyse contextuelle (identité, nom de la société, nom de projet).

Pour l'analyse non contextuelle, nous faisons usage d'expressions régulières, en particulier pour les numéros de téléphone, adresses e-mail, noms de domaine, réseaux

TAB. 2 – *Répartition des étiquettes dans les 3 jeux de données*

Étiquette	JDR		JDF		JDA	
Human	971	14,51%	1023	19,58%	1196	13,46%
Organization	1023	15,29%	943	18,05%	1537	17,29%
Location	2533	37,86%	2150	41,15%	2680	30,15%
Phone_Number	473	7,07%	371	7,10%	688	7,74%
Function	567	8,47%	0	0,00%	1449	16,30%
Email	297	4,44%	371	7,10%	344	3,87%
Url	227	3,39%	0	0,00%	303	3,41%
Social_Network	18	0,27%	0	0,00%	28	0,32%
Reference_User	9	0,13%	0	0,00%	11	0,12%
Reference_Code_Postal	269	4,02%	367	7,02%	349	3,93%
Project	98	1,46%	0	0,00%	124	1,40%
Reference_CEDEX	150	2,24%	0	0,00%	146	1,64%
Reference_CS	56	0,84%	0	0,00%	33	0,37%
	6691		5225		8888	

sociaux. Il en est de même pour l'identification d'adresses basée sur les codes postaux, CEDEX et CS. Pour l'identification des autres constituants adresse, nous avons recours à des ressources de types de voie communs, et à des attendus tels qu'un numéro de voie. Pour la fonction, nous utilisons un lexique de noms de métiers, généralisé, complété par des suffixes communs de noms de métiers (e.g., -ogue, -iste). Enfin, nous exploitons également des introducteurs caractéristiques de type de constituants tels que "Tél :", "Adresse :".

Pour l'analyse positionnelle, nous exploitons le fait que l'identité arrive en début de signature. Une analyse contextuelle permet d'attribuer le nom de société, considéré obligatoire, et les éventuels noms de projet, parmi les constituants non catégorisés, quitte à remettre en cause le pré-étiquetage non fiable d'un constituant. C'est également dans l'analyse contextuelle que sont exploités les noms de métiers et leurs suffixes fréquents.

Afin de fiabiliser les déductions sur les noms de personnes et les noms de société, nous effectuons un calcul des chaînes de coréférence, mettant en relation le nom de la personne et son identifiant d'e-mail, et le nom de la société avec les noms de site internet et noms de domaines de l'adresse e-mail. Ce processus à la fois original et efficace, est basé sur le calcul de distance d'édition de chaînes de caractère. Il n'est cependant pas utilisable sur le corpus de validation pseudonymisé.

Enfin, pour plonger dans le format attendu et de garantir la fiabilité de certaines déductions, un post-traitement est effectué et force le réétiquetage de certains tokens (CEDEX, e-mail, téléphone, nom de domaine). Notre segmentation en token est alignée avec la segmentation attendue.

5 Résultats

La performance de notre système est respectivement évaluée à 99% et 100% de f-mesure sur les jeux de mise au point réaliste (JDR) et factice (JDF), voir les tables 3 et 4. Sur le jeu de données authentique (JDA) utilisé pour la validation de notre système, la performance se dégrade de manière significative à 83% de f-mesure, avec une précision et un rappel sensiblement équivalents, voir table 5.

L'explication d'un tel écart de performance entre la phase de mise au point et la phase de validation pourrait être attribuée à une trop grande spécificité de notre système aux données de mise au point. S'arrêter à cette conclusion serait cependant hâtif. Les causes sont certainement multiples, mais il nous semble que la différence de nature des jeux de données de mise au point et de validation n'est pas neutre et mérite une attention toute particulière.

Les jeux de données JDR et JDF sont en effet artificiels et respectent des contraintes syntaxiques qui ont fait l'objet de spécifications (Cousot et al., 2023). Le jeu de validation JDA est quant à lui composé de données authentiques, sans contraintes de bonne formation. On observe par conséquent une variabilité formelle et lexicale bien plus importante dans ce dernier.

L'absence de marques de structuration dans le JDA n'est pas sans impact. Les données de ce jeu ont été récupérées par l'intermédiaire d'un formulaire de "don" de signature en ligne. Ce dispositif de captation, adapté pour les signatures purement textuelles, ne permet pas la récupération des informations structurelles du format HTML : les éléments de structuration (`div`, `br`, ...) sont ainsi perdus, engendrant des signatures composées parfois d'une unique ligne de texte, sans aucun séparateur exploitable. L'impact est significatif puisque ces marques, observables et exploitables dans la phase de mise au point pour délimiter les constituants et mettre en œuvre le critère positionnel, ne sont plus disponibles. Autre conséquence liée que nous attribuons au mode de captation : nous notons le doublement, voire le triplement de certains constituants de l'adresse dans le jeu de validation, ce qui a pour conséquence de rendre inexploitable le critère d'unicité de certains constituants.

La pseudonymisation appliquée sur le jeu de validation JDA est également à considérer dans la perte de performance. En remplaçant systématiquement et de manière aléatoire les noms de personnes, les noms d'organisation, les noms de domaine des adresses e-mail et des URLs, la procédure casse la cohérence textuelle des signatures, en particulier le schéma relationnel qui est exploitable dans les jeux de mise au point.

Enfin, la dimension multilingue représentée dans les jeux de mise au point JDR et JDF est absente du jeu de validation JDA qui n'est par exemple composé que d'adresse française.

Pour contrecarrer cette différence importante de forme entre les jeux de mise au point et de validation, nos efforts pendant la période de validation ont porté sur la relativisation des séparateurs de constituants, et sur leur délimitation basée sur des critères formels. Le temps limité de la phase de validation ne nous a cependant pas permis de mettre en œuvre l'ensemble des indices potentiellement exploitables.

6 Conclusion

Dans cet article, nous avons présenté notre participation au défi TextMine'23 (Cousot et al., 2023) "Reconnaissance d'entités d'intérêts dans les signatures d'e-mails".

Nous avons choisi de réaliser une chaîne de traitement complète, depuis la segmentation des signatures jusqu'à la production du fichier d'annotations au format attendu. La performance de notre système atteint 99% et 100% en f-mesure sur les jeux de données de mise au point et 83% de f-mesure sur le jeu de validation.

Parmi les options stratégiques que nous avons retenues, nous pensons que le fait de chercher à étiqueter non pas des tokens mais des unités de plus haut niveau est particulièrement différenciant. Nous pensons à ce titre qu'une évaluation basée sur la délimitation et l'étiquetage des constituants logiques serait pertinente et complémentaire à celle menée sur les tokens.

Afin de fiabiliser les déductions sur les noms de personnes et les noms de société, nous avons mis au point un calcul des chaînes de coréférence, mettant en relation le nom de la personne et son identifiant d'e-mail, et le nom de la société avec les noms de site internet et noms de domaines de l'adresse e-mail.

La technique d'analyse utilisée a l'intérêt de fournir des résultats explicables et interprétables.

La chute de performances entre la phase de mise au point et la phase de validation a été longuement analysée. Nous avons mis en évidence les limites du jeu de données authentique. Nous pensons cependant que le jeu de données authentique (JDA) est très certainement celui qui présente le plus d'intérêt puisqu'il propose des données réelles. Le processus d'acquisition devrait cependant être revu pour permettre la préservation des informations structurelles. Le processus de pseudonymisation devrait quant à lui également être reconsidéré pour conserver la cohérence textuelle.

Références

Bendahman, N., K. Cousot, et C. Lopez (2022). Reconnaissance d'entités d'intérêt dans les signatures d'e-mails à partir d'un jeu de données authentique. *TextMine'22*.

Carvalho, V. R. et W. W. Cohen (2004). Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam*, Volume 2004.

Chen, H., J. Hu, et R. W. Sproat (1999). Integrating geometrical and linguistic analysis for email signature block parsing. *ACM Transactions on Information Systems (TOIS) 17*(4), 343–366.

Cousot, K., C. Lopez, P. Cuxac, et V. Lemaire (2023). Défi textmine'23 - reconnaissance d'entités d'intérêts dans les signatures d'e-mails. In *Actes de l'atelier TextMine'23, Conférence Extraction et Gestion des Connaissances 2023 (EGC'23)*, Lyon, pp. à paraître.

Estival, D. (2008). Author attribution with email messages. *Journal of Science, Vietnam National University 1*, 1–9.

Giguet, E. et G. Lejeune (2021). Daniel at the FinSBD-2 task : Extracting list and sentence boundaries from PDF documents, a model-driven approach to PDF document analysis. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, Kyoto, Japan, pp. 67–74. -.

Giguet, E. et N. Lucas (2010). The book structure extraction competition with the resurgence software at caen university. In S. Geva, J. Kamps, et A. Trotman (Eds.), *Focused Retrieval and Evaluation*, Berlin, Heidelberg, pp. 170–178. Springer Berlin Heidelberg.

Giguet, E. et N. Lucas (2022). GREYC@FinTOC-2022 : Handling Document Layout and Structure in Native PDF Bundle of Documents. In M. El-Haj, P. Rayson, et N. Zmandar (Eds.), *4th Financial Narrative Processing Workshop (FNP 2022)*, Marseille, France, pp. 100–104.

Li, J., S. Sen, et N. Zaman (2015). Entity extraction from business emails. *International Journal of Information Technology and Computer Science 7*(9), 15–22.

Summary

This paper presents our contribution to the TextMine'23 Challenge related to "Named Entity Recognition in Email Signatures". The performance of our system reaches 99% and 100% F1-score on the training sets and 83% F1-score on the test set.

TAB. 3 – *Évaluation sur le jeu de mise au point JDR*

	précision	rappel	f1-score	support
email	1.00	1.00	1.00	297
function	0.99	0.99	0.99	567
human	1.00	1.00	1.00	971
location	1.00	1.00	1.00	2538
organization	0.98	0.98	0.98	1018
phone_number	1.00	0.99	0.99	473
project	0.87	0.84	0.85	98
reference_cedex	0.96	1.00	0.98	150
reference_code_postal	1.00	1.00	1.00	269
reference_cs	1.00	1.00	1.00	56
reference_user	1.00	1.00	1.00	9
social_network	1.00	1.00	1.00	18
url	1.00	1.00	1.00	227
micro avg				
macro avg	0.98	0.98	0.98	6691
weighted avg	0.99	0.99	0.99	6691
F1	0.9912928493115661			

TAB. 4 – *Évaluation sur le jeu de mise au point JDF*

	précision	rappel	f1-score	support
email	1.00	1.00	1.00	371
function	0.00	0.00	0.00	0
human	1.00	1.00	1.00	1023
location	1.00	1.00	1.00	2150
organization	1.00	1.00	1.00	943
phone_number	1.00	1.00	1.00	371
project	0.00	0.00	0.00	0
reference_cedex	0.00	0.00	0.00	0
reference_code_postal	1.00	1.00	1.00	367
reference_cs	0.00	0.00	0.00	0
reference_user	0.00	0.00	0.00	0
social_network	0.00	0.00	0.00	0
url	0.00	0.00	0.00	0
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			

TAB. 5 – *Évaluation sur le jeu de validation JDA*

	précision	rappel	f1-score	support
email	1.0000	1.0000	1.0000	344
function	0.8881	0.7723	0.8261	1449
human	0.8582	0.9055	0.8812	1196
location	0.9101	0.8540	0.8811	2678
organization	0.6023	0.6415	0.6213	1537
phone_number	0.9663	1.0000	0.9829	688
project	0.0762	0.1935	0.1093	124
reference_cedex	0.9799	1.0000	0.9898	146
reference_code_postal	0.8876	0.9107	0.8990	347
reference_cs	0.0000	0.0000	0.0000	37
reference_user	0.4000	0.1818	0.2500	11
social_network	0.9310	0.9643	0.9474	28
url	0.9967	1.0000	0.9984	303
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			