Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploitation et l'archivage des méls

Touria Aït El Mekki, Bénédicte Grailles, Tsanta Randriatsitohaina

Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploration et l'archivage des méls

Touria Ait El Mekki[1], Bénédicte Grailles[1], Tsanta Randriatsitohaina[1]

[1] University of Angers – [prenom.nom@univ-angers.fr]

Abstract

The Pêle-Mél project, supported by the French Ministry of Culture and currently underway, aims to provide archivists with assistance in reading, interpreting, and fine-tuning the context of mailboxes. Our work is based on the observation of archiving practices and on the current state of research in automatic language processing, particularly in computational terminology and information processing. We propose a method of exploration and contextualization of professional mailboxes to facilitate navigation in these systems. It consists, on the one hand, in building a list of terms and named entities to be validated, which we improve by injecting knowledge from specialized thesaurus or message metadata and, on the other hand, in making a classification of messages available via semantic classes grouped around themes chosen by the archivist.

Keywords : archiving, email, machine learning, classification, terminology, named entity

Résumé

Le projet Pêle-Mél, soutenu par le ministère de la Culture français et en cours de réalisation, vise à proposer à l'archiviste une assistance à la lecture, à l'interprétation, à la contextualisation fine des messageries électroniques. Notre travail s'appuie sur l'observation des pratiques d'archivage et sur l'état actuel des recherches en traitement automatique des langues, notamment en terminologie computationnelle et en traitement de l'information. Nous proposons une méthode d'exploration et de contextualisation des messageries professionnelles devant faciliter la navigation dans ces messageries. Il s'agit d'une part de construire une liste des termes et des entités nommées à valider que nous améliorons en injectant des connaissances issues des thesaurus spécialisés ou des métadonnées des messages et, d'autre part, de mettre à disposition une classification des messages via des classes sémantiques regroupées autour de thèmes choisis par l'archiviste.

Mots clés : archivage, courriel, apprentissage, classification, terminologie, entité nommée

1. Introduction

Le projet Pêle-Mél, en cours de réalisation, vise à proposer à l'archiviste une assistance à l'interprétation et la contextualisation fine des messageries électroniques et la rédaction de synthèses à la demande. Notre démarche intègre des méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels, et de l'intervention humaine dans un processus coopératif, et s'appuie sur les résultats de l'élaboration de l'ontologie du domaine à partir du contenu des messageries et de corpus complémentaires. Outre l'identification des termes et la reconnaissance d'entités nommées, nous proposons une projection entre différents plans permettant de faire le lien entre l'ontologie, les entités nommées et les différentes thématiques (Alghamdi and Alfalqi 2015). La méthodologie proposée prend en compte la totalité des informations des messageries (adresses, sujet,

contenu, signature, pièces jointes). Les traitements automatiques permettent de construire un ensemble de connaissances à partir des textes et des ressources disponibles, de guider l'intervention humaine en assurant sa cohérence. Nous explorons comment différents types de ressources peuvent être exploités afin de valider des résultats ou pour spécialiser davantage des données préexistantes. Dans ce cadre, nous proposons une approche qui consiste à construire une liste des termes et des entités nommées que nous validons et que nous améliorons en injectant des connaissances issues des thesaurus spécialisés ou des métadonnées des messages. Après une présentation du contexte, de l'état de l'art et des objectifs (section 2), nous reviendrons sur les corpus et la méthodologie (section 3) puis sur la question de l'analyse de textes appliquée aux messages (section 4). Nous souhaitons montrer que l'état de l'art en terminologie computationnelle permet aujourd'hui d'envisager des outils d'analyse fine de textes beaucoup plus puissants.

2. Contexte et objectifs

Les services d'archives sont compétents pour collecter, sélectionner et conserver les documents et données physiques et électroniques. C'est une compétence réglementaire et une obligation légale. Pour répondre à ce besoin, des missions archives ont été placées au sein des ministères. Elles se chargent d'organiser et recueillir les données pour archivage, notamment pour documenter les processus de décision, et de restituer cette information à la demande. Depuis les années 1990, le passage à grande échelle à une production numérique a amené à développer des solutions d'archivage électronique. À partir des années 2000, la question des messageries électroniques a été incluse dans la réflexion sur la pérennisation (Programme Vitam 2013).

2.1. *Enjeux*

Dans le travail administratif quotidien, ces messageries sont centrales. Support d'informations stratégiques, elles sont les traces uniques de processus décisionnels. Si des outils d'extraction existent, leur volumétrie et leur pérennisation en silos séparés ne suffit pas à satisfaire les besoins d'un bon archivage, à savoir une appréhension par niveaux de granularité et une contextualisation fine. De fait, la recherche dans les corpus se trouve compromise, chaque messagerie étant maintenue dans un entrepôt isolé. C'est l'expérience de recherches difficiles sur des messageries déjà archivées qui a convaincu la mission archives du ministère de la Santé et des Solidarités, forte d'une solide pratique en matière de collecte d'archives électroniques (200 entrées dont 45 % de messageries, avec des boîtes pouvant aller jusqu'à 70 Go) à participer au projet Pêle-mél, initié par les laboratoires Temos et Leria de l'université d'Angers, et soutenu par le ministère de la Culture. Ce partenariat a permis d'accéder à des corpus de messagerie émanant du cabinet, c'est-à-dire de collaborateur.rices personnel.les choisi.es par le.la ministre, ayant pour mission de le.la conseiller et de l'assister dans la réalisation de l'ensemble de ses missions et jouant un rôle administratif et politique à ses côtés.

2.3. *État de l'art*

Si la question de la pérennisation des messages a été largement traitée, notamment dans le cadre du programme Vitam [http://www.programmevitam.fr], celle de l'accès est beaucoup moins travaillée et uniquement en langue anglaise, via la reconnaissance d'entités nommées : projet ePADD de l'université de Standford [https://library.stanford.edu/projects/epadd] ou RATOM de l'université de Caroline-du-Nord [https://ratom.web.unc.edu/].

Le traitement automatique des courriels est devenu indispensable depuis l'explosion de la quantité de méls échangés entre les différents services et entités. Plusieurs méthodes ont été mises en œuvre pour les explorer. La catégorisation des courriels a été mobilisée dans le but de les organiser, par exemple pour la détection de spam (Tang et al. 2013). Les méthodes utilisées peuvent reposer sur des règles (Xia, 2020), sur l'apprentissage (Nadjate et al., 2020) ou la combinaison des deux. Il est nécessaire d'évaluer l'efficacité des extracteurs automatiques des termes (Nazarenko et al. 2009). L'analyse du contenu des courriels a aussi été utilisée pour catégoriser les contacts en retrouvant les contacts ayant le même centre d'intérêt (Johansen, 2007) ou en identifiant les contacts appartenant à une même communauté (Tyler et al., 2003). Les interactions par méls pouvant être considérés comme un réseau social, des méthodes ont été mises en œuvre pour établir un réseau de contacts en fonction de leurs interactions. Cette analyse se base sur les statistiques des contributions de chaque contact au sein du réseau (Karagiannis, 2009) ou par l'apprentissage de l'objectif des méls envoyés, pour informer, enquêter et planifier (Lockerd, 2003). La détection d'évènements à partir des courriels a aussi été prospectée permettant l'identification de ceux (recrutement, discours, événement social) mentionnés dans les textes avec les détails comme la date, l'heure, le lieu. Des méthodes de reconnaissance d'entités nommées (Suárez et al. 2020) sont utilisées pour extraire ces informations qui sont ensuite soumis à validation (Nair et al., 2020).

2.3. *Objectifs*

Au-delà de l'accès à l'information, notre objectif est de proposer à l'archiviste une assistance à la lecture, à l'interprétation et à la contextualisation des messages pris individuellement, comme des boîtes méls dans leur ensemble et une aide à la rédaction de synthèses à la demande. Par exemple, dans un courriel citant le nom d'un.e autre membre du cabinet, nous devons l'identifier sans ambiguïté et faire le lien avec la fonction occupée à la date du courriel. Ainsi, dans un document mentionnant "la ministre des Solidarités et de la Santé" dans un courriel écrit en 2010, nous devons être capable de fournir l'information qu'il s'agit de Roselyne Bachelot-Narquin et permettre la navigation dans les autres courriel envoyés ou reçus et le réseau de correspondant.es. On cherche aussi à identifier des expressions temporelles et des événements. Ce résultat peut être obtenu via l'injection de connaissances complémentaires et externes au corpus de messages, le recours aux méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels, et l'élaboration de l'ontologie du domaine à partir du contenu des messageries (corps du message, fichiers joints, lien) (Maynard et al. 2008) et des listes de vocabulaire contrôlé externes.

L'objectif final est de proposer à l'archiviste une lecture guidée, un instrument d'orientation permettant de progresser dans la restitution du contenu informationnel par sérendipité et des synthèses pour faciliter la gestion des paquets archivés.

3. Corpus et méthodologie

Pour réaliser ce projet, nous disposions initialement de deux messageries de conseillères et de 8 636 messages reçus ou envoyés entre 2007 et 2011. Le contenu des messages en lui-même est court, avec une rédaction correcte, en moyenne 9 phrases par message, mais le contenu reste répétitif. Nous avons donc complété chaque message par les pièces jointes afin d'enrichir le corpus initial et identifié des sources complémentaires : copies d'annuaires papier et organigrammes natifs électroniques ou numérisés, deux thésaurus soit environ 7 000 descripteurs reliés uniquement par des relations hiérarchiques, et 810 discours prononcés par la ministre entre 2010 et 2012. Les annuaires et organigrammes permettent de repérer les

fonctions occupées par une partie des correspondants au sein du ministère et leur évolution au fil des années, ce qui justifie le temps consacré à leur saisie manuelle.

Nous proposons une méthode composée des principaux modules suivants :

- Un module de prétraitement : découpage de courriels, élaboration de réseaux de contacts, extraction des informations comme la fonction, le rattachement des personnes physiques ou morales à partir des adresses méls complétées par les annuaires et les signatures. Une base de données qui contient l'ensemble de ces informations a été construite, ce qui permet de mettre en évidence les liens entre les personnes, les fonctions et les services et facilite aussi différentes visualisations graphiques.

- Un module d'analyse fine de textes : extraction de termes, d'entités nommées, association termes et entités nommées pour appréhender les informations contenues dans chaque courriel.

- Un module de classification thématique afin de parcourir les messages selon les différents thèmes. Il permettra d'avoir une vue plus globale de l'ensemble des messageries.

4. Analyse de textes

À la différence des corpus usuellement utilisés dans les traitements automatiques des textes, les courriels sont plus courts et contiennent plus d'acronymes et d'entités nommées.

4.1. *Reconnaissance des sigles*

De nombreuses abréviations n'apparaissent pas dans le dictionnaire. Nous avons extrait les abréviations du corpus, grâce à l'étiquetage morpho-syntaxique proposé par TreeTagger, via l'étiquette *abbr* pour abréviation. Après un rapide filtrage, une liste de 254 abréviations dont 195 existaient dans les thesaurus fournis, a été établie. Les spécialistes du domaine ont validé les mots restants. Tous les acronymes n'ont pas été identifiés de cette manière. La liste a été complétée en exploitant les mots étiquetés *org* via un rapide filtrage. En tout, près de 400 acronymes ont ainsi été détectés.

4.2. *Extraction de termes et des entités nommées*

Tout mot ou groupe de mots représentant un concept spécifique d'un domaine est un terme. Leur extraction automatique peut être effectuée par des méthodes linguistiques reposant sur les catégories grammaticales, ou par des méthodes statistiques reposant sur les fréquence et distribution des mots dans le texte, ou encore par la combinaison des deux (TermSuite, Cram and Daille, 2016). Les sorties proposent des candidats-termes à soumettre à validation.

Les entités nommées sont généralement assimilées aux noms propres (noms de personne physique ou morale), noms de lieu ou encore à des valeurs (date, heure...). La reconnaissance d'entités nommées peut être effectuée à partir des bases de connaissances ou par méthode d'apprentissage.

Parmi les outils qui traitent le français et qui répondent au besoin de confidentialité des messageries, nous avons testé plusieurs bibliothèques et choisi la bibliothèque SpaCy (Honnibal and Montani, 2017).

La validation des sorties est exposée à la section 4.4.

4.3. *Association des termes et des entités nommées*

Les avancées récentes de l'apprentissage automatique ont permis de mettre en œuvre des méthodes permettant la représentation des mots en fonction de leur contexte. On parle de plongement lexical. Celui-ci capture les similarités sémantiques et syntaxiques des mots en fonction du contexte dans lequel ils apparaissent dans le texte. Pour notre projet, nous avons utilisé Fasttext (Bojanowski et al., 2017) afin d'identifier les termes ayant une similarité élevée et retrouver les termes proches des personnes ou des organismes mentionnés dans les textes. Le terme Solidarités a été ainsi associé avec cohésion sociale*, ministre*, Bachelot Narquin, Varenne, direction générale, autant de désignations cohérentes avec le ministère de Roselyne Bachelot-Narquin. Nous avons pu aussi récupérer des termes associés au nom Bachelot comme ministère*, ministre*, département*, sociale, DGCS*, liés à son poste, jeunesse, politique de la famille*, solidarités, prévention*, sécurité en lien avec des politiques publiques. Les termes avec * sont présents dans les thesaurus. Après filtrage, les résultats sont prometteurs mais pas suffisants. Sur les 7 000 descripteurs du thésaurus, seuls 60 ont des relations.

L'étape suivante est l'exploration de méthodes et approches terminologique, directes et indirectes, pour structurer la liste des termes et l'outillage à l'aide d'une méthode qui formalise le regroupement conceptuel, le regroupement à base de patrons et les méthodes non terminologiques.

4.4. *Validation des termes et des entités nommées*

Les entités nommées et termes extraits automatiquement bénéficient d'une phase de validation. Elle s'organise comme suit : par comparaison avec les descripteurs des thesaurus, puis avec la liste des abréviations validées, enfin en confrontant les noms de personne avec les métadonnées des messages et les annuaires. La méthode présentée à la section 4.3 permet une deuxième étape de validation en gardant les termes et les entités nommées associées, suivant la stratégie de Omrane et al., 2011. Le traitement des métadonnées nous permet ainsi de désambiguïser les mentions d'entités nommées et de retrouver que [Luc] Chatel a été secrétaire d'État dans le même gouvernement.

4.5. *Classification*

Nous avons dans un premier temps développé le module de classification en utilisant les méthodes existantes pour regrouper les courriels selon leurs similarités et différences en appliquant un modèle de clustering KMeans, en donnant en entrée au modèle les courriels représentés par tous les mots qu'ils contiennent, puis uniquement des noms, verbes, adjectifs, adverbes, ou par les termes identifiés afin qu'il les distingue par groupes. Les premiers résultats de cette classification n'ont pas été probants. Les centroïdes des clusters sont très proches. La similarité entre le contenu des courne permet pas au modèle d'identifier des points suffisamment discriminants pour orienter la classification. De plus, nous travaillons sur des textes très spécialisés pour un domaine sans modèle pré-entraîné adapté.

Nous avons amélioré le processus en réalisant une classification guidée par une liste de 50 thèmes structurés en trois niveaux par les spécialistes, donc en regroupant les classes autour de ces thèmes. Après plongements de mots, nous retrouvons le terme sida dans la même classe que ist, vih, maladie, vhc etc.; le terme médicament est associé à usage, produit, pharmaceutique, generique, prescrit, tamiflu etc.

Nous avons ensuite représenté chaque thème par sa classe sémantique et calculé la similarité avec chaque courriel pour l'associer aux thèmes les plus proches. La distribution des messages par rapport aux thèmes est cohérente d'après les experts. La confrontation entre

celle-ci et la répartition en cours issue d'une annotation manuelle à partir d'un échantillon aléatoire de messages permettra d'évaluer la qualité des résultats.

5. Conclusion

Quelle méthodologie mettre en place pour aider à explorer les courriels et les contextualiser via une lecture guidée ? Un pré-traitement s'impose pour collecter les métadonnées qui servent ensuite à valider des résultats de l'extraction des termes et des entités nommées, en combinaison avec des modèles de plongements lexicaux permettant de les identifier. Pour avoir une vue plus globale des messageries, une classification des messages est indispensable et nécessite des temps de validation par les spécialistes. Nous avons proposé à l'archiviste un nouveau mode d'exploration des messageries conservés à partir des termes, de leurs variants, des classes. Les premiers résultats sont encourageants. La méthode proposée peut être déployée dans d'autres environnements d'archivage.

References

Alghamdi R. and Alfalqi K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications,* vol. 6., pp. 147-153.

Bojanowski P., Grave E., Joulin A., and Mikolov T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, vol. 5, pp.135-146.

Cram D. and Daille B. (2016). Terminology Extraction with Term Variant Detection. *In Proceedings of ACL-2016 System Demonstrations*, pp. 13-18.

Honnibal M., and Montani I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Johansen L., Rowell M., Butler K. R., and McDaniel P. D. (2007). Email Communities of Interest. In *CEAS The Fourth Conference on Email and Anti-Spam, 2-3 August 2007, USA*.

Karagiannis T., and Vojnovic M. (2009). Behavioral profiles for advanced email features. *In Proceedings of the 18th international conference on World Wide Web*, pp. 711-720.

Lockerd A., and Selker T. (2003). DriftCatcher: The Implicit Social Context of Email. In *INTERACT'03*, pp. 813-816.

Maynard D., Li Y. and Wim P. (2008). NLP Techniques for Term Extraction and Ontology Population. In *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press, NLD, pp. 107-127.

Nadjate S., Adi K. and Allili M. (2020). Semantic Representation Based on Deep Learning for Spam Detection. *Foundations and Practice of Security*, pp. 72-81.

Nair A. M., Justus A. A., Ramesh A., and Rajan B. (2020). Event Extraction from Emails. *International Journal of Computer Applications,* vol. 176, n°41, pp. 1-8.

Nazarenko A., Zargayouna H., Hamon O. and Puymbrouck J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Revue TAL,* ATALA, vol. 50, pp. 257-281.

Omrane N., Nazarenko A. and Szulman S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. *In Proceedings of Conférence internationale de Terminologie et Intelligence Artificielle*, Paris, pp. 80-86.

Programme VITAM (2013). *L'archivage des messageries électroniques. Preuve de concept*, 103 p.

Suárez P., Dupont Y., Muller B., Romary L. and Sagot B. (2020). Establishing a New State-of-the-Art for French Named Entity Recognition. *Language Resources and Evaluation Conference LREC*.

Tang G., Pei J., and Luk W. S. (2014). Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems*, vol. 41, pp. 1-31.

Tyler J.R., Wilkinson D.M., Huberman B.A. (2003). Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Huysman M., Wenger E., Wulf V. (eds) *Communities and Technologies*. Dordrecht, pp. 81-96.

Xia T. (2020). A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems. *IEEE Access*, vol. 8, pp. 82653- 82661.