L'anonymisation, pierre d'achoppement pour le traitement automatique des courriels

Hugues de Mazancourt, Alain Couillault, Gaëlle Recourcé
Eptica, L3I/Université de La Rochelle - AproLab/Aproged, Kwaga
hugues.de-mazancourt@eptica.com, alain.couillault@univ-lr.fr - a.couillault@aproged.org, recource@kwaga.com

Résumé

Constatant que les corpus de courriels disponibles pour la recherche sur les interactions linguistiques sont rares et parcellaires, nous décrivons trois degrés d'anonymisation pour une diffusion maitrisée des courriels de support, respectivement, dans une entreprise, à l'intérieur d'un consortium lié par un accord de confidentialité, à l'ensemble de la communauté des chercheurs.

Index terms— corpus de courriels, obfuscation, anonymisation, ODISAE, e-mail

1. Introduction

Le projet ODISAE, qui réunit 6 entreprises et un laboratoire de recherche, est un projet de recherche opérationnelle dans le cadre du FUI-17 (Fonds Unifié Interministériel). Il se donne pour mission de dessiner les contours de la nouvelle génération d'outils mis à la disposition des plateaux de support client. Ces outils vont exploiter le contenu des interactions avec les clients. Même si le téléphone reste largement majoritaire, la proportion d'interactions par ce media devrait descendre en dessous des 50% à l'horizon 2015. Les autres interactions sont des interactions écrites, en très grande majorité des courriels. L'analyse des interactions par courriel est donc un sujet primordial pour le projet. Or, les corpus de courriels effectivement utilisables à des fins de recherche sont quasi inexistants. Après un rapide inventaire des corpus disponibles, nous exposerons les difficultés pour constituer, utiliser et diffuser des corpus de courriels et exposerons les pistes de solutions mises en oeuvre dans le cadre du projet ODISAE.

2. Des corpus de courriels

Lorsqu'on se confronte au matériau qu'est le courriel, les corpus utilisés dans la littérature montrent certaines limites : les courriels sont confidentiels et ne peuvent être mis à disposition à l'extérieur de l'entreprise (Laval et al., 2009), les travaux sont effectués sur d'autres types de corpus similaires (Carvalho and Cohen, 2004), des corpus artificiels (courriels filtrés par les utilisateurs pour (Estival et al., 2007), ou des courriels échangés dans le cadre de scénarios de simulation d'entreprise pour (Cohen et al., 2004)). Dans le cadre du projet GramLab, l'équipe Eptica/Lingway a rassemblé et mis à disposition sous licence LGPLR un corpus de courriels de taille modeste (54 courriels) qui ont été anonymisés et rendus disponibles en ligne (Couillault et al., 2013), la documentation et les conditions d'utilisation sont décrits dans la Charte Ethique et Big Data associée (Couillault and Fort, 2013).

Ainsi, malgré de nombreux travaux effectués sur des courriels, les corpus effectivement disponibles sont très rares. Or, la disponibilité de corpus est un préalable à toute étude linguistique et à toute démarche de TAL.

Les corpus extraits de forums ou de listes de diffusions ne peuvent être considérés comme représentatifs des corpus de courriels, les situations d'énonciations étant clairement différentes, notamment du fait du caractère privé du courriel, échangé entre un émetteur et un ou plusieurs destinataires identifiés.

2.1. Le corpus Enron

L'assertion qui précède, sur la non disponibilité de corpus de courriels de volume significatif, souffre d'une exception majeure : le corpus Enron. Ce corpus a été rendu public en 2003 suite au scandale judiciaire autour d'Enron et reformatté ensuite pour étude (Klimt and Yang, 2004). A de rares exceptions près, tous les autres corpus d'échange de courriels (on excepte les spams) librement disponibles et de volume suffisant sont en fait des extraits de ce corpus. Ainsi, les deux grandes agences européennes (ELRA) et états-uniennes (LDC) de diffusion de corpus n'en recensent aucun autre. Or, le corpus Enron est inutilisable pour servir de modèle à une analyse d'interactions entre un client et un centre support pour plusieurs raisons :

- il est en anglais, alors que le projet couvre aussi le français,
- la situation d'échange dans laquelle se déroulent les conversations est totalement différente (on est dans de l'échange *corporate*),
- ces échanges datent de plus de 10 ans, avec les outils et les usages de l'époque

Ce dernier point, qui semble n'être qu'un point de forme, est un aspect majeur : les outils ainsi que la manière d'utiliser le medium courriel ont considérablement évolué dans la dernière décennie et bien peu des conclusions qui pourraient être tirées de la forme d'un échange courriel de cette époque seraient applicables aujourd'hui.

2.2. A qui appartient un courriel ?

Beaucoup a été écrit sur le sujet et nombre de droits s'appliquent aux courriels. En premier lieu, un courriel est une correspondance et la Convention européenne de sauvegarde des Droits de l'Homme et des Libertés fondamentales dispose à l'alinéa 1er de son article 8 que *Toute personne a droit au respect de sa vie privée et familiale, de son domicile et de sa correspondance*. Ce point est repris en particulier dans l'article 226-15 du Code Pénal sur la violation du secret des correspondances qui punit d'un an d'emprisonnement et de 45 000 € d'amende *le fait, commis de*

mauvaise foi, d'ouvrir, de supprimer, de retarder ou de détourner des correspondances arrivées ou non à destination et adressées à des tiers, ou d'en prendre frauduleusement connaissance. Cet article s'applique également explicitement à des correspondances effectuées par la voie des télécommunications (voir aussi l'Art. 432-9). Le point de savoir si un courriel est une correspondance privée fait encore débat, essentiellement dans le domaine professionnel, c'est-à-dire lorsque le courriel est émis par un salarié en utilisant le matériel mis à sa disposition par l'entreprise. La jurisprudence permet, dans certains cas, à l'employeur de prendre connaissance de ces échanges [1].

Au niveau européen, la directive 95/46/CE du Parlement européen et du Conseil (European Parliament and Council of the European Union, 1995) stipule en son article 26 *que les principes de la protection ne s'appliquent pas aux données rendues anonymes d'une manière telle que la personne concernée n'est plus identifiable*. Sur la base de cette directive, le G29 (PARTY, 2014), considère que la pseudonymisation n'est pas une technique d'anonymisation, et qu'elle ne fait que réduire le lien potentiel entre une donnée et l'identité de l'individu [2].

Quoi qu'il en soit, prendre connaissance ne signifie pas divulguer et un courriel reste toujours protégé par le droit d'auteur. En d'autres termes, il n'est pas possible de publier un corpus d'échanges de courriels sans demander l'autorisation à chacun des auteurs, ce qui explique l'absence d'une telle ressource publique. Pourtant, les centres de support possèdent des historiques d'échanges de courriels qui sont utilisés quotidiennement par les agents, par exemple pour aider à la résolution de cas similaires, et ce en toute légalité puisque les agents sont tenus au secret sur les données privées et ne diffusent pas les messages échangés.

Au-delà de cet encadrement légal, la diffusion de corpus de courriels s'accompagne de considérations que l'on qualifiera d'éthiques : elle ne doit en aucun cas nuire ni aux émetteurs, ni aux récepteurs, ni aux individus ou entreprises éventuellement mentionnés dans le corps des courriels. Dans le cas des corpus de courriels de support, objets du projet ODISAE, les entreprises qui abritent ces centres de support tiennent à protéger leur image et leur réputation et ne peuvent autoriser la diffusion de contenus qui associeraient leur nom à quelque jugement de valeur.

La diffusion d'un corpus de courriels nécessite donc plusieurs niveaux d'anonymisation :

- anonymisation de premier niveau : obfusquer les méta-données émetteurs et récepteurs (i.e. les champs *from*, *to* et *cc* des courriels)
- anonymisation de deuxième niveau : en plus de l'anonymisation de niveau 1, empêcher toute collecte des informations de contacts (par exemple à des fins publicitaires) contenues dans le corps des courriels
- anonymisation de troisième niveau (ou **anonymisation vraie**) : garantir qu'il n'est pas possible, directement ou indirectement, d'identifier les émetteurs, les récepteurs ni les tiers mentionnés dans le corps des courriels.

Aucun de ces niveaux d'anonymisation ne doit, en outre, biaiser les résultats des travaux scientifiques basés sur ces corpus.

Nous considérons que seule l'anonymisation vraie autorise la diffusion libre des corpus de courriels.

2.3. Quelles contraintes d'anonymisation

Tout en rappelant que le choix d'une procédure d'anonymisation dépend du contexte et de l'application visée, le G29 propose trois critères pour évaluer une telle procédure :

1. L'individualisation : est-il toujours possible d'isoler un individu ?
2. La corrélation : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ?
3. L'inférence : peut-on déduire de l'information sur un individu ?

Cependant, les critères de corrélation et d'individualisation, bien qu'ils puissent être assez simplement mis en oeuvre par des techniques de remplacement aléatoires de chaînes de caractères, interdiraient toute analyse de conversations, ce qui est justement l'objet du projet Odisae.

L'objectif était donc, dans le cadre du projet, tout en s'accordant avec les contraintes légales, de pouvoir disposer de corpus de courriels en provenance d'outils de support client, afin de permettre au consortium de travailler sur ce matériau, de respecter les ces considérations éthiques. Dans le cadre du projet ODISAE, nous avons mis en oeuvre deux niveaux d'anonymisation :

- une anonymisation de deuxième niveau en obfusquant toute information de contact à la fois dans les méta-données et dans les données.
  Cette approche ne permettant pas la diffusion large des corpus, nous avons restreint la diffusion au consortium du projet ODISAE dans le cadre d'un accord de confidentialité stricte.
- une anonymisation vraie, pour une partie des corpus disponibles, effectuée manuellement sur des corpus préalablement transformés automatiquement.

L'anonymisation vraie correspond à la définition donnée par (Med, 2006) :

> Anonymisation is the task of identifying and neutralising sensitive references within a given document or set of documents.

On le voit, disponibilité des corpus et utilisation scientifique paraissent comme des contraintes *a priori* à la fois fortes et contradictoires. (Rehm et al., 2008) contournent cette difficulté par une sorte d'arrangement avec la licence, en proposant des masques obfusqués différement en fonction des usages.

2.4. Contractualisation

Dans le cadre du projet ODISAE, l'accord de confidentialité signé par chacun des partenaires est particulièrement restrictif et s'applique par défaut à l'ensemble des corpus

1. Voir notamment l'arrêt de la Cour de cassation, Chambre sociale, rendu le 16/05/2013, cassation (12-11866)

2. *pseudonymisation is not a method of anonymisation. It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.*

mis à disposition par les partenaires, ce qui a notamment pour effet d'expliciter les règles de (non-)diffusion des données et d'y sensibiliser l'ensemble des acteurs. Une liste de corpus disponibles décrivant leur origine, leur description et la licence éventuellement associée, est annexée à cet accord de confidentialité.

2.5. Modalités d'anonymisation

Une contrainte, non plus législative mais technique cette fois, est qu'un échange soumis à l'anonymisation doit *ressembler* à l'échange initial. Par exemple, remplacer toutes les adresses courriel par une même chaîne de caractères empêche toute exploitation de la donnée (pour reconstituer un fil de discussion). Il en va de même pour les noms de personne (prénom ou nom patronymique). Le processus se base donc sur une liste de référence de prénoms et patronymes communs (utilisé par ailleurs par Eptica pour un produit d'analyse de CV). Cette approche (*pseudonymisation* au sens de (Med, 2006)) permet ainsi de satisfaire à la contrainte d'usage scientifique. Certains biais seront nécessairement introduits dans le corpus, comme des références géographiques aberrantes parce que les noms de ville auront été changés, mais ce biais est considéré comme acceptable pour la mise au point de modèles d'analyse.

2.6. Procédure d'anonymisation

La procédure consiste, pour chaque courriel (ou chaque fil de discussion, si les données ont déjà cette structure) à utiliser un module d'extraction d'entités nommées. Ce module permet de détecter les données sensibles que l'on veut anonymiser : nom, prénom, adresse courriel, mais aussi numéros d'identification (numéros de commande, etc.). Ensuite, on regroupe ces éléments pour masquer des variantes simples (changement de casse, présence ou non d'accents, etc.). Enfin, on calcule, pour chacun, un remplaçant de même nature, choisi de façon déterministe à partir d'un calcul de hashcode sur la donnée initiale. Par exemple, le hashcode sur prénom donnera un rang dans la liste de prénoms et ainsi fournira le prénom de substitution. Dernière étape, toutes les occurrences des chaînes correspondant aux éléments détectés sont remplacés dans le texte en cours d'anonymisation, ceci afin de pallier les éventuels silences des grammaires d'extraction.

2.7. Anonymisation vraie

L'anonymisation vraie des courriels, objectif incontournable à la mise à disposition de corpus réellement exploitables par la communauté scientifique n'est aujourd'hui possible qu'au prix d'un travail manuel. On ne mesure pas aujourd'hui l'impact que pourrait avoir une telle réécriture sur une modéliation de la langue qui en résulterait. En effet, réécrire des portions d'un texte introduit un nouveau rédacteur à ce texte. Pour le moins, le *re-rédacteur* devrait posséder des facultés de mimétisme significatives afin de rendre la ré-écriture plausible et donc utilisable pour un traitement automatique.

3. Conclusion

Nous avons présenté une procédure permettant le partage d'échanges de courriels sans que les données privées ne soient dévoilées, dans le respect de la vie privée et de la législation. Cette procédure implique des processus techniques de TAL, mais aussi la création d'une sorte d'*entreprise étendue* (au consortium) via des accords stricts de confidentialité. L'ensemble du consortium peut ainsi travailler sur l'étude de corpus de courriels partagés, mais la diffusion libre de ces corpus n'est pas possible. Cette diffusion nécessiterait une tâche d'obfuscation, bien plus coûteuse, ainsi qu'une licence adaptée.

4. Remerciements

Ce travail a été effectué dans le cadre du projet ODISAE, financé par le 17e Fonds Unifié Interministériel. Le projet réunit l'Aproged, La Cantoche Productions, le Centre Départemental du Tourisme de l'Aube, Kwaga, Eptica, l'INSEE et TokyWoky.

5. References

Carvalho, V. R. and Cohen, W. W. (2004). Learning to extract signature and reply lines from email. In *CEAS 2004 - First Conference on Email and Anti-Spam*, Mountain View, CA.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. (2004). Learning to classify email into "speech acts". In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 309–316, Barcelona, Spain, July. Association for Computational Linguistics.

Couillault, A. and Fort, K. (2013). Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasburg, France, July. 4 pages.

Couillault, A., Vinckx, A., De Mazancourt, H., Grandry, F., and Recourcé, G. (2013). Rapport technique Projet Gramlab : livrable SP5.1 Use Case Eptica/Lingway : identification d'amorces de reprises. july.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B. (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, Melbourne, Australia.

European Parliament and Council of the European Union. (1995). Directive 95/46/CE. Journal officiel n° L 281 du 23/11/1995, 11.

Klimt, B. and Yang, Y. (2004). Introducing the Enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*.

Laval, P., Meunier, F., Recourcé, G., and Surcin, S. (2009). Kwaga : une chaîne UIMA d'analyse de contenu des mails - proposition de démonstration.

(2006). *An Introduction to NLP-based Textual Anonymisation*, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA).

PARTY, A. . D. P. W. (2014). Opinion 05/2014 on anonymisation techniques. 0829/14/EN WP216, April.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008). The metadata-database of a next generation sustainability web-platform for language resources.