Stratégie pour l'anonymisation systématique d'un corpus volumineux d'interactions multilingues : spécification d'un outil interactif

Christophe REFFAY, François-Marie BLONDEL, STEF – ENS-Cachan / IFÉ – ENS-Lyon -
Emmanuel GIGUET, GREYC, Université de Caen Basse-Normandie, CNRS

Mots-clés : Anonymisation, De-identification, Analyse d'interaction
Axe thématique : Constitution et exploitation des corpus, en amont ou en aval des situations d'IC
Langue de la communication : italien
Langue du support visuel : anglais

Pour que les recherches sur des expériences non reproductibles puissent être discutées et comparées, des dispositifs comme Mulce[1] ou Calico[2] ont été développés pour permettre le partage de données entre chercheurs. Les données qui sortent des laboratoires doivent être vierges de toute marque permettant d'identifier les acteurs impliqués dans les expériences analysées. Un procédé d'anonymisation est donc nécessaire.

Dans le cas des interactions médiées par ordinateur, le problème le plus délicat est celui de l'anonymisation à l'intérieur des messages. En effet, les acteurs de la formation (enseignants, tuteurs, apprenants) laissent fréquemment des traces d'identification d'eux-mêmes ou de leurs pairs comme les noms, prénoms, adresses, établissement, identifiants d'outils de communication (téléphone, skype, twitter, facebook, etc.).

Comme évoqué dans (Reffay & Teutsch, 2007), des données personnelles peuvent être utiles à certaines analyses (e.g. : prénom, localisation). En l'illustrant sur des exemples tirés de corpus choisis, nous tenterons de mettre en évidence la tension entre anonymisation et analyse en précisant ce qui doit être masqué pour garantir l'anonymat, et ce qui risque d'être perdu pour l'analyse. Le processus de transformation doit être également guidé par les besoins des analyses envisagées après anonymisation.

Si les volumes à traiter sont tels que les procédures complètement manuelles ne sont pas envisageables, de nombreux obstacles s'opposent aussi à la mise au point d'une méthode d'anonymisation automatisée. Au cours de cette communication, nous identifierons les obstacles à l'anonymisation automatique, puis proposerons une méthode semi-automatique, applicable à un corpus multilingue.

Laissant au chercheur le contrôle sur les données et le processus de transformation, notre méthode s'appuie sur des techniques (Aramaki et al., 2006 ; Mestre et al., 2010 ; Grouin et al., 2009) proposées dans le domaine hospitalier. Le processus d'anonymisation suit 3 étapes :
1. Une construction (interactive et itérative) d'un dictionnaire des entités nommées à masquer ainsi que les balises à utiliser pour le remplacement ;
2. Un repérage automatique des contextes où apparaissent ces entités nommées en vue de détecter de nouvelles formes d'entités nommées. Les étapes 1 et 2 sont répétées jusqu'à ce qu'on ne trouve plus aucune nouvelle entité nommée ;
3. Le remplacement multiple en une seule passe de toutes les entités nommées.

À partir d'un cas issu de Galanet, dont les données seront importées sur plateforme Calico (Blondel & Giguet, 2011), nous détaillerons ce processus d'anonymisation sur un corpus donné, ainsi que les modalités qui permettent au chercheur de contrôler ce processus de transformation et de l'adapter au niveau de protection qu'il souhaite.

---

[1] http://mulce.org : Projet ANR 2007-2010 coordonné par T. Chanier.
[2] http://www.stef.ens-cachan.fr/calico/calico.htm : ERTé 2006-2009 coordonnée par E. Bruillard

Références bibliographiques

Aramaki, E., Imai, T., Miyo, K. & Ohe, K. (2006). Automatic Deidentification by using Sentence Features and Label Consistency. In *Workshop on Challenges in Natural Language Processing for Clinical Data*. http://lululu.com/profile/paper/2006-i2b2/i2b2-deid.pdf

Blondel , F.-M., & Giguet, E. (2011). Analyses et partages de corpus de discussions avec Calico - Leçons tirées d'une expérience récente. In Dejean, C., Mangenot, F., Soubrié, T. (coord.). Actes du colloque Epal 2011 (Échanger pour apprendre en ligne), Université Stendhal – Grenoble 3. http://w3.u-grenoble3.fr/epal/dossier/06_act/pdf/epal2011-blondel-giguet.pdf

Grouin, C., Rosier, A., Dameron, O. & Zweigenbaum, P. (2009) Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. 13èmes Journées francophones d'informatique médicale, Nice, 28-30 avril 2009. *Risques, Technologies de l'Information pour les Pratiques Médicales*, Fieschi, M., Staccini, P., Bouhaddou, O. et Lovis, C. (Eds), *Informatique et Santé*, Vol. 17, Springer, 2009.

Meystre, S.M., Friedlin, F. J., South, B.R., Shen, S. & Samore, M.H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010, 10:70, http://www.biomedcentral.com/1471-2288/10/70

Reffay, C. & Teutsch, P. (2007). *Anonymisation de corpus réutilisables : masquer l'identité sans altérer l'analyse des interactions*. Rapport interne, LIFC, 12 pages. http://edutice.archives-ouvertes.fr/edutice-00158877/fr/