Pseudoanonymization of Data from a Fragile Population (Mind-It Project)

Daya Messaoudi*[1], Cédrick Fairon[2], Louise-Amélie Cougnon[3], and Bernard Hanseeuw[4]

[1]Université catholique de Louvain (UCL) – Belgique

[2]Centre de traitement automatique du langage (UCLouvain) (CENTAL) – Place Blaise Pascal B-1348 Louvain-la-Neuve, Belgique

[3]MiiL - ILC - UCLouvain – Belgique

[4]Cliniques universitaires St Luc [Bruxelles] – Cliniques universitaires St LucAvenue Hippocrate 10 - 1200 Bruxelles - Belgique, Belgique

Résumé

Since the adoption in 2016 of the General Data Protection Regulation (GDPR), research that work with a corpus containing personal data must go through anonymization or pseudoanonymization (Adams et al. 2019: 1). The NLP field is highly concerned by this issue, as there is a double challenge of processing large quantity of corpora which may contain personal data and of providing solutions for the anonymisation/pseudoanonymization of the data (Ahrenberg and Megyesi, 2019: 5). Further, personal information may be needed for some linguistic studies (Raffay and Teutsch, 2007: 4).
In this work, we will present our corpus of electronic messages from a fragile population and the process of automatic pseudoanonymization of the personal data that we decided to follow in order to respect the GDPR. Both methods of pseudoanonymization and anonymization are based on identity information hiding but the former is reversible while the latter is not (Atanassova et al., 2019: 57). Therefore, pseudoanonymous data is within the scope of the GDPR unlike anonymous data (Francopoulo and Schaub, 2020: 9). Our corpus is part of the Mind-it project which aims to build a machine learning algorithm to help the preclinical detection of Alzheimer's disease based on the analysis of linguistic changes that mark the disease in electronic messages. The study is thus doubly concerned by the GDPR as it involves data from social networks and electronic messaging containing a lot of identification data (De Mazancourt et al. 2014: 2), and a fragile population which may be subject to Alzheimer's disease and different forms of senility. We face a dilemma of hiding any personal data in the text on the one hand and keeping intact its semantic to be able to apply relevant treatments on the other hand. Further, we should be able to link our anonymized texts to the medical, social, and cultural profile of their authors. This dilemma raises many questions such as which kind of data should be masked according to the research objectives? To what extent can we modify a corpus without altering the quality of the linguistic study? Which pseudoanonymization techniques should we use? Which tools to apply?

In this presentation, we first detail the numerous challenges faced by a project dedicated to older adults. We will then focus on the testing of different anonymization techniques and tools (python, Unitex/GramLab, dedicated software). We will also propose to adapt some of the existing programs to make them correspond to our project's criteria. Finally, the results

*Intervenant

produced by these solutions will be compared and discussed.

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Fernander Valencia, F. & Wechsler, R. (2019, September). AnonyMate: A toolkit for anonymizing unstructured chat data. In Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland (No. 166, pp. 1-7). Linköping University Electronic Press.

Ahrenberg, L. & Megyesi, B. (2015). Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland.

Atanassova, I., Bertin, M., & Le Béchec, M. (2019). Sécuriser le traitement des traces numériques dans le cadre du Règlement général sur la protection des données (RGPD) : anonymisation et pseudonymisation. I2D Information, donnees documents, (1), 55-58.

De Mazancourt, H., Couillault, A., & Recourcé, G. (2014, November). L'anonymisation, pierre d'achoppement pour le traitement automatique des courriels. In Journée d'Etude ATALA Ethique et TAL.

Francopoulo, G., & Schaub, L. P. (2020, May). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In workshop on Legal and Ethical Issues (Legal2020) (pp. 9-14). ELRA.