Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Désidentification de données texte produites dans un cadre de relation client

Guillaume Dubuisson Duplessis Elliot Bartholme Sofiane Kerroua
Mathilde Poulain Ahès Roulier Anne-Laure Guénet
EDF Commerce, Direction Numérique, Tour PB6, 178 Rond-Point de la Défense, 92800 Puteaux, France
{guillaume.dubuisson-duplessis, mohamed-sofiane.kerroua,
anne-laure.guenet}@edf.fr

RÉSUMÉ

Cette démonstration présente une solution performante de désidentification de données texte selon 13 types d'entités nommées et entraînée sur des données issues de la relation client.

ABSTRACT

De-identification of customer relationship text data

This demonstration presents an efficient text de-identification system based on 13 named-entity types. It has been trained on customer relationship data.

MOTS-CLÉS : désidentification, reconnaissance d'entités nommées, RGPD.

KEYWORDS: de-identification, named-entity recognition, GDPR.

1 Désidentification de données texte à EDF Commerce

1.1 Minimiser l'usage des données à caractère personnel

Chaque mois des millions de données texte sont produites dans le cadre de la relation client au sein d'EDF Commerce de la part des clients (e.g., e-mails, réponses libres à des questionnaires de satisfaction) et des conseillers (e.g., commentaires de contact). Ces données majoritairement en français sont riches : elles offrent un large panel de structures allant d'expressions libres et spontanées à des formes contraintes comme des formulaires ; tout en manifestant une grande variabilité en termes de respect de l'orthographe, de la syntaxe et de niveau de langue. Ces données sont utilisées pour répondre au mieux aux attentes de nos clients. En outre, elles sont exploitées dans de nombreux cas d'utilisation « métier » visant à optimiser la relation client. Le cas le plus répandu est celui des tableaux de bord permettant aux opérationnels de suivre une activité comme, par exemple, le traitement des e-mails client (Dubuisson Duplessis *et al.*, 2019). Afin de nourrir ces cas d'utilisation, des données texte peuvent être consultées par des individus (e.g., lors de tâches d'annotation manuelle, lors de retour aux données dans des tableaux de bord), et intervenir dans de nombreuses tâches de modélisation du TALN (e.g., classification, analyse de ressenti client). Or, ces données texte font généralement apparaître des données à caractère personnel (DCP) telles que les noms, prénoms, adresses postales, e-mails, et de nombreux identifiants numériques. Le règlement général sur la protection des données (RGPD) renforce les droits des individus quant à l'utilisation qui peut être

Volume 4 : Démonstrations et résumés d'articles internationaux, pages 10–13.
HAL : hal-02768496.

faite de leurs DCP. A cette fin, il exige une minimisation de l'usage des DCP au regard de la finalité pour laquelle elles sont traitées.

1.2 Une approche par désidentification pour respecter les contraintes du RGPD

Afin de limiter efficacement l'usage des DCP, notre approche vise à les supprimer par une procédure de désidentification automatique qui fonctionne en deux temps : (i) une phase de reconnaissance d'entités nommées (NER) correspondant aux DCP (Nouvel *et al.*, 2016), suivie de (ii) une phase de délexicalisation substituant le texte des entités nommées par le type des entités nommées. Par exemple, « Je suis Jean Dupont (Paris 12è). Je ne comprends pas ma facture n° 12 345 6. » peut être délexicalisé en « Je suis _PERSON_ (_LOCALISATION_). Je ne comprends pas ma facture n° _NUMBER_. ». L'avantage de la désidentification est de prévenir la divulgation de DCP à des personnes non-habilitées tout en conservant la substance du document. Dans le cadre de la modélisation, la désidentification en amont sur les données utilisées permet de prévenir des biais d'apprentissage en empêchant l'utilisation de DCP dans les décisions des algorithmes. Par exemple, elle évite l'impact du genre ou de l'origine des noms et prénoms. Une limite notable d'une approche par désidentification est atteinte par l'usage de périphrase (« Président de la République française, je souhaite ... »). L'évaluation de nos algorithmes indique que ce type de phénomène est très rare dans nos données de relation client.

1.3 Une solution de désidentification performante en français

La construction de notre solution de désidentification a bénéficié d'un processus d'annotation rigoureux et de qualité. 13 types d'entités nommées parmi lesquels des adresses, des noms/prénoms, des e-mails, des informations bancaires et des numéros client ont été annotés en interne via une plateforme web inspirée du projet Camomile (Poignant *et al.*, 2016). Le corpus d'apprentissage contient environ 1000 e-mails et 700 conversations de chat client/conseiller représentant 11459 tours de parole. Ces données sont en français. Le corpus contient approximativement 6200 instances d'entités. La répartition des catégories d'entités est détaillée dans le Tableau 1. Les e-mails et les conversations

Localisation	Personne	Info. numériques	URL/e-mails	Info. bancaire
19.1% / 41.9%	47.7% / 27%	29.6% / 22.9%	3.3% / 8%	< 0.5%

TABLE 1 – Répartition des catégories d'entités dans le corpus d'apprentissage. Lecture : proportion en nombre d'instances d'entité / proportion en nombre de tokens.

sont pré-traités pour uniformiser l'encodage (en particulier au niveau des « smileys »), supprimer les balises et normaliser les entités HTML. Les données de conversations de chat ont été simplement segmentées au niveau des tours de parole. Les e-mails ont subi une segmentation plus lourde visant à isoler les informations techniques (par exemple, les informations d'en-têtes telles que l'expéditeur, les destinataires, les dates), des champs structurés (par exemple, « Numéro client : ... » et des champs de texte libre. Les informations techniques et les champs structurés sont découpés via des heuristiques. Les parties de texte libre sont segmentées en phrases.

Algorithmiquement, notre solution se fonde sur l'hybridation de règles et d'apprentissage profond afin d'obtenir le meilleur compromis entre performance et temps de calcul. L'approche par règles est

utilisée pour les entités fortement structurées et peu dépendantes du contexte comme les identifiants numériques, les informations bancaires, les e-mails et les URL. L'approche par apprentissage profond se concentre sur les entités telles que les prénoms, noms, adresses postales et lieux. Plusieurs modèles sont maintenus afin de maximiser la performance en NER tout en satisfaisant les deux principales contraintes techniques impliquées que sont le temps de calcul et les limitations liées à la taille des documents. La première partie de nos modèles se fonde sur la famille des RNN et utilise des bi-LSTM+CRF (Akbik *et al.*, 2018) combinant des plongements vectoriels au niveau des caractères (Lample *et al.*, 2016), des sous-tokens (Heinzerling & Strube, 2018), et des tokens (Pennington *et al.*, 2014). La seconde partie se fonde sur l'architecture « transformer » (Vaswani *et al.*, 2017) et utilise fructueusement les modèles en français créés par la communauté tels que CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020).

De par sa fondation sur des données variées de la relation client, notre système peut être considéré comme spécialisé pour ce domaine. Néanmoins, les types d'entités traités ont une portée générique à l'exception de quelques identifiants numériques spécifiques à EDF Commerce (par ex., numéro client, numéro de contrat, index de compteur).

Notre solution obtient de bonnes performances en NER sur des énoncés de chat conseiller/client confirmant ainsi la maturité des approches par apprentissage profond pour des cas opérationnels (Barriere & Fouret, 2019). Sur les onze entités gérées par la partie « règle », nous obtenons une micro F1 à 0.94 sur notre test. Sur les deux entités gérées par la partie « apprentissage profond », nous obtenons une micro F1 à 0.97. En outre, nous avons évalué manuellement la performance en anonymisation sur 1000 conversations. Nous avons lu des conversations désidentifiées en cherchant à recroiser des informations dans le texte pour vérifier leur anonymisation. Sur cette évaluation exigeante, nous avons obtenu un taux d'anonymisation de 96%. Cette évaluation nous a permis d'identifier quelques cas limites pour notre système. Sans prétendre à une présentation exhaustive, notre système semble manquer certaines entités peu communes (comme des noms de village, des noms de personne ou des prénoms peu courants) ou à l'opposé des mots plutôt communs (par exemple, un nom de famille pouvant dénoter un métier comme « M. Boulanger », un prénom comme « Claire » qui peut également être un adjectif). Les erreurs de typographie, variées dans des expressions spontanées de client, fragilisent également la reconnaissance des entités nommées. Sur les conversations client-conseiller, cela s'est particulièrement manifesté autour de la reconnaissance d'adresses e-mails. Une perspective intéressante de nos travaux réside en une analyse quantitative et qualitative approfondie des erreurs systématiques de notre solution.

2 Démonstrateur

Notre solution de désidentification est destinée à être utilisée aussi bien en amont des projets « data » afin de désidentifier les données textes avant qu'elles ne soient manipulées qu'en production dans projets industrialisés. Un cas d'utilisation fréquent est celui des tableaux de bord de pilotage permettant des retours aux données anonymisées (voir, par exemple, le projet Cameli@ (Dubuisson Duplessis *et al.*, 2019)).

Cette démonstration propose une interface permettant de saisir un texte libre en se mettant à la place d'un client EDF et de le désidentifier suivant 13 types d'entités nommées en utilisant les algorithmes précédemment évoqués. Le démonstrateur inclut la possibilité de tester plusieurs approches alternatives en laissant libre choix du système d'apprentissage profond utilisé (RNN, « transformer »).

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Gilles Pouëssel, Mélanie Cazes, Meryl Bothua, Lou Charaudeau, Ibtissem Menacer, Uta Hosokawa, Aurore Hamimi, Anaël Cabrol, Sylvain Boucault, François Bullier, Jean Vidal et Marie Hervé.

Références

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, p. 1638–1649.

BARRIERE V. & FOURET A. (2019). May I Check Again ? A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. arXiv preprint : 1909.03453.

DUBUISSON DUPLESSIS G., KERROUA S., KUZNIK L. & GUÉNET A.-L. (2019). Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 623–625.

HEINZERLING B. & STRUBE M. (2018). BPEmb : Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1030.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. arXiv preprint : 1912.05372.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. arXiv preprint : 1911.03894.

NOUVEL D., EHRMANN M. & ROSSET S. (2016). *Named Entities for Computational Linguistics*. John Wiley & Sons.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

POIGNANT J., BUDNIK M., BREDIN H., BARRAS C., STEFAS M., BRUNEAU P., ADDA G., BESACIER L., EKENEL H., FRANCOPOULO G. *et al.* (2016). The Camomile collaborative annotation platform for multi-modal, multi-lingual and multi-media documents. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.