Faire du TAL sur des données personnelles : un oxymore ?

Hugues de Mazancourt[1], Alain Couillault[2, 3], Gilles Adda[4, 5], Gaëlle Recourcé[6]
(1) Eptica, 95b rue de Bellevue, 92100 Boulogne-Billancourt
(2) L3i, Laboratoire Informatique, Image et Interaction, Université de La Rochelle, Avenue Michel Crépeau, 17042 La Rochelle, France
(3) APROLAB, Aproged, 43, rue Beaubourg, 75003 Paris
(4) Spoken Language Processing group, LIMSI-CNRS, Orsay (5) IMMI-CNRS, Orsay
(6) Kwaga Lab, Kwaga, 15, rue Jean Baptiste Berlier, 75013 Paris
hugues.de-mazancourt@eptica.com, alain.couillault@aproged.org, gadda@limsi.fr, recource@kwaga.com

Résumé. Le présent travail s'inscrit dans le cadre de la version 2 de la Charte Ethique et Big Data (Couillault & Fort, 2013). Il présente les difficultés inhérentes à l'application de techniques de TAL sur des données à caractère personnel, en même temps que la nécessité d'un tel travail.

Abstract.
Is NLP of personal data an oxymoron ?
We present the work in progress for the version 2 of the Big Data Charter and present some aspects of the use of pesonal data for an industrial system embedding NLP technology.

Mots-clés : Données privées, Big Data, Ethique.

Keywords: Privacy, Big Data, Ethics.

1 Introduction

Notre réflexion, dans le cadre de la révision de la charte Ethique et Big Data, se centre sur l'oxymore qu'est l'impossible nécessité de réaliser des travaux de TAL sur des données à caractère personnel. Elle vise à généraliser et aller au delà des solutions mises en œuvre. Après une présentation de la Charte et du cadre légal de ces données, nous présentons quelques solutions mises en œuvre pour effectuer de tels travaux, en soulignant leurs limites. Nous montrerons que la Charte Ethique et Big Data, si elle ne résout pas tous les problèmes, fournit une méthodologique pour aborder ces traitements.

2 La charte Ethique et Big Data

La charte Ethique et Big Data (Couillault & Fort, 2013) a pour objectif de fournir un outil de documentation des jeux de données - et en particulier les ressources langagières -, afin d'en assurer la traçabilité et la transparence. Son principe de base est celui d'un questionnaire **déclaratif**, l'objet étant de fournir à l'utilisateur de la donnée, quelle qu'elle soit, des informations sur la façon dont elle a été produite et la licence attachée. Elle n'assure donc pas l'éthique mais fournit suffisamment d'informations à l'utilisateur pour qu'il puisse décider du degré d'éthique des données et de l'utilisation qu'il en fait.

2.1 Structure de la Charte

La Charte se présente comme un formulaire à remplir pour décrire une donnée, quelle qu'elle soit. Elle comprend une une première section de description des données fournies, puis s'organise autour des trois axes suivants via des séries de questions précises :

— traçabilité des travaux effectués sur les données (transformations) et des acteurs impliqués,
— explicitation de la licence d'usage,
— description des éventuelles législations et contraintes spécifiques

La nécessité d'une Charte pour les données a été clairement exposée par ses initiateurs (Couillault & Fort, 2013). La première version de la Charte a d'ores et déjà été adoptée par de nombreuses associations dont Cap Digital. Ainsi, de nombreux projets proposés au financement public (FUI et autres Appels à Projets gérés par BPIfrance) mentionnent la Charte dans leur description technique.

2.2 Motivations d'une nouvelle version

L'objet de la version 2 de la Charte est double. Il consiste en premier lieu à en faire un objet *actionnable*, c'est-à-dire un formulaire informatisé qui puisse être rempli, consulté en ligne et directement accessible avec les données qu'il décrit. Ainsi, l'objectif de traçabilité est-il complet, puisqu'on dispose du lien direct entre la Charte et la Donnée. Cette nouvelle version est également l'occasion de renforcer la Charte sur un certain nombre d'aspects. Il s'agit notamment de la confronter aux multiples initiatives assimilées en France ou en Europe. Parmi ces travaux, citons les *ethics guidelines* du programme H2020 (European Commission, 2014) ou les travaux dans le domaine de la santé (plus précisément dans les groupe d'intérêt TIC et Santé des Pôles de Compétitivité). Cette confrontation est l'objet d'un groupe de travail qui se réunit mensuellement, de février à juin 2015. Dernier point, on a parfois assimilé cette Charte à une utilisation uniquement liée aux données linguistiques, en raison de sa genèse. En effet, la Charte trouve son origine dans une réflexion sur les corpus constitués en TAL avec des outils de crowdsourcing peu respectueux du droit du travail (voir (Fort *et al.*, 2014)). La validation de la Charte sur ces différents travaux va définitivement invalider cette critique.

Un dernier objectif de cette nouvelle version est d'aborder explicitement le sujet des données personnelles, sujet qui était traité dans la première version par simple renvoi aux dispositions de la CNIL.

3 Le TAL et les données personnelles

La question des données personnelles agite de façon grandissante le monde du Traitement Automatique des Langues : le TAL a en effet besoin de corpus et un nombre croissant de ces corpus contient de telles données (de façon directe ou indirecte), entre autres avec l'abondance du contenu généré par les utilisateurs sur les forums ou réseaux sociaux. Ce contenu, riche en potentiel d'applications et en phénomènes linguistiques (et qui intéresse de plus en plus l'industrie) recèle des informations personnelles, privées, voire intimes, en particulier dans le domaine de la santé. Ces informations sont publiées par leurs auteurs sans grand souci de leur utilisation au delà d'une publication dans un forum à un instant donné. Faire du traitement automatique sur ces données est tentant, aisé, mais, comme le soulignent (Boyd & Crawford, 2011), l'accessibilité de la donnée ne rend pas automatiquement son utilisation éthique.

3.1 Les données personnelles

La CNIL définit une donnée personnelle comme suit [1] :

> Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.

1. http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/

Des discussions sont toujours en cours à la CNIL pour décider si, dans les faits, une donnée est ou non personnelle (cf. l'adresse IP qui a été intégrée récemment à cette définition), preuve s'il en est que le contour de cette notion est moins simple qu'il n'y paraît. A cette définition, on peut ajouter celle (toujours selon la CNIL) de **données sensibles** : *Les données sensibles sont celles qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou sont relatives à la santé ou à la vie sexuelle de celles-ci*. La collecte ou le traitement de données sensibles est interdit, sauf consentement explicite de l'intéressé (ou jugement spécifique impliquant une autorisation de la CNIL). D'une manière générale, le traitement de données personnelles (même non sensibles) est très strictement encadré par la loi dite *Informatique et Libertés* et ne doit, entre autres, *pas être réalisé pour d'autres finalités* que celles qui ont permis la collecte, sauf consentement éclairé de l'auteur [2]. Il est donc a priori illégal d'effectuer des traitements d'analyse sur un corpus collecté sur le web (ou pire, dans une boîte mail) pour peu qu'une information non-structurée [3] (un texte, un enregistrement) permette d'identifier un individu réel.

3.2 La question de l'anonymat

La première parade pour éviter cette identification est d'anonymiser les textes : plus de référence nominative [4], plus d'identification. De façon assez paradoxale, on demande au TAL d'outiller l'anonymisation, la reconnaissance d'entités nommées étant perçue comme la première étape (mais certes pas la seule, voir notamment (Amblard *et al.*, 2014) ou (Medlock, 2006)) pour détecter les éléments nominatifs dans un texte.

Une fois ces éléments identifiés, une solution populaire est la pseudonymisation, qui consiste à remplacer une entité par une autre entité de même type (un nom par un autre nom, un prénom par un autre prénom, etc.). Elle a le mérite de produire des contenus similaires à ceux d'origine, sur lesquels il sera possible d'entraîner des automates qui pourront ensuite être appliqués à de "vrais" textes. La pseudonymisation, permise par l'altération d'entités nommées en des éléments non-signifiants constitue ainsi pour (De Mazancourt *et al.*, 2014) un des niveaux d'anonymisation qui se situe, dans le cas des courriels entre l'anonymisation des méta-données et l'anonymisation dite *vraie*, définie comme l'impossibilité de restituer l'identité de l'auteur.

3.3 Impossibilité technique d'une anonymisation vraie

De nombreux chercheurs ont montré les limites d'une telle approche. Par exemple (Eshkol-Taravella *et al.*, 2014), démontre, lors du traitement du corpus oral ESLO (recueil d'interviews à Orléans en 1968), la présence d'éléments d'identification directs (*mon père a fondé le plus grand cabinet d'ophtalmologiste de la ville*) ou non directs (*le locuteur est patron de café au moment de l'enregistrement et il travaillait auparavant dans l'aviation militaire*) qui peuvent permettre, avec une connaissance raisonnable du contexte, de *réidentifier* les individus réels.

Récemment, dans un domaine proche, une étude (de Montjoye *et al.*, 2013) a montré qu'il suffisait de 4 points géolocalisés pour identifier de façon unique 95% des utilisateurs dans une base de plus d'un 1,5 million d'individus à partir des traces de connexion de leurs téléphones portables sur 6 mois. Sans aller toutefois jusqu'à la ré-identification il démontre qu'on peut lier de façon presque certaine une connexion isolée à tout le parcours du téléphone et donc de son utilisateur. De façon similaire, les outils de TAL, appliqués à des corpus suffisamment vastes, comme le sont ceux qui sont à disposition actuellement sur les forums ou réseaux sociaux, vont permettre d'établir des liens, de collecter un certain nombre d'indices sur les individus a priori anonymes. Ils vont permettre de déduire des textes que l'utilisateur X a une voiture, qu'il est allé l'année dernière aux Canaries et qu'il a acheté un smartphone de marque Y, simplement en analysant les "traces" qu'il laisse sur le Net. Ce sont ces corrélations dont sont friands les professionnels du marketing. Ces corrélations extraites des textes sont autant d'indices pour la réidentification. En faisant le parallèle avec les traces laissées par les téléphones portables, on imagine aisément qu'il est possible de collecter suffisamment d'indices à partir de textes pour connecter une portion significative de verbatims produits sur un réseau social, par exemple, même couverts par l'anonymat formel (masquage des identifants de toute sorte). Et dès lors, il suffira une fuite (par exemple) pour que ce parcours dans le réseau se mue en une masse d'informations personnelles et intimes sur des personnes réelles.

2. On retrouve cette notion de *consentement libre et éclairé* dans la relation soignant-malade, ainsi que stipulé dans l'Art. 16-3 du Code Civil
3. nous éliminons ici les méta-données structurées accompagnant les textes, pour lesquelles une étude linguistique n'est pas pertinente
4. au sens large, incluant tous les types d'identifants

4 Le cas du projet ODISAE

Si le TAL appliqué à des données à caractère personnel porte en soi le germe d'une violation potentielle de la vie privée, le linguiste ne doit pourtant pas s'abstenir de ce type de travail. En effet, l'étude de corpus de données à caractère personnel peut être particulièrement enrichissant sans que l'objectif soit de porter atteinte à la vie privée, même si elle peut en devenir un effet de bord. C'est le cas par exemple du projet ODISAE.

4.1 Présentation du projet

Le projet ODISAE, co-financé par BPIfrance et la Région Ile de France, labellisé par les Pôles de compétivité Cap Digital et Images et Réseaux dans le cadre du FUI-17 (Fonds Unifié Interministériel) réunit huit PME partenaires et un universitaire, le LINA. Le chef de file de ce projet est la société Eptica, éditeur de logiciel spécialisé dans la relation client. Les partenaires sont des éditeurs dans des domaines proches ou connexes (Cantoche, Kwaga, Jamespot, TokyWoky) ou organismes ayant un rôle de valideur de par une activité dans ce domaine (Aproged, Centre Départemental du Tourime de l'Aube et INSEE). L'objectif du projet est de fournir des outils logiciels innovants pour un centre de contact client en analysant le contenu des échanges réalisés entre le centre de contact et les utilisateurs. Le projet se focalise sur les échanges écrits (mail, chat, etc.). Il s'agit de considérer ces échanges comme un dialogue entre l'utilisateur et la marque (au sens large) et non pas comme une succession de messages déconnectés les uns des autres. On va par exemple tâcher de déclencher, avant qu'il ne soit trop tard, des actions lorsqu'une conversation se "passe mal" ou lorsque le client menace de changer de fournisseur (détection d'attrition). On cherche également à comprendre le degré d'adéquation de la FAQ, mise à disposition de ceux qui répondent aux clients, avec les problématiques qu'ils expriment au long de cet échange.

La première difficulté, pour un tel projet, est la faiblesse de l'état de l'art sur le sujet, faiblesse qui s'explique très simplement par un manque de corpus utilisables pour l'étude linguistique. En effet, le seul corpus d'e-mails à la disposition des chercheurs est le corpus Enron ((Klimt & Yang, 2004)) qui souffre de nombre de défauts pour une utilisation dans ce contexte, à commencer par son âge et ses conditions de réalisation. Il s'agit d'échanges datant de plus de dix ans, essentiellement entre collègues d'une même société (Enron). Or, en dix ans, l'usage de l'e-mail a considérablement changé. De plus, on ne s'adresse pas à ses collègues de la même façon qu'on s'adresse à une marque pour réclamer un remboursement.

Cette pénurie de corpus s'explique par le fait qu'un échange d'e-mails est une donnée éminemment personnelle, couverte de plus par des droits multiples comme le droit d'auteur et le droit à la correspondance privée. Dans le cadre du projet, la solution apportée pour permettre l'étude linguistique est décrite dans (De Mazancourt *et al.*, 2014). Elle situe sur deux plans :

1. sur un plan technique, les partenaires valideurs fournissent leurs corpus au consortium suite à une pseudonymisation effectuée dans les locaux du fournisseur de données. Aucune donnée *authentique* ne sort de chez ce partenaire, seul est fourni un corpus qui "ressemble" au corpus initial mais dont tous les noms et identifiants ont été remplacés par des entités similaires ;
2. sur un plan juridique, tous les membres du consortium sont soumis à un accord de confidentialité strict, empêchant la diffusion des corpus (ainsi pseudonymisés) en dehors du consortium.

Une phase d'anonymisation *vraie*, qui aurait impliqué la réécriture manuelle des e-mails afin d'en masquer toute information directe ou indirecte n'a pas été retenue lors du financement du projet. Elle aurait été particulièrement coûteuse.

4.2 Limites de la solution

Si la solution mise en place dans ODISAE permet au consortium de travailler et de produire les résultats attendus d'un point de vue opérationnel, elle se heurte à deux critiques du point scientifique. D'une part, le projet n'aura pas fait progresser la communauté sur l'aspect de la disponibilité d'un corpus d'e-mails puisque les données ne peuvent être diffusées à l'extérieur du consortium. D'autre part, les données restant confidentielles, il n'est pas possible de réfuter scientifiquement les conclusions qui auront pu être tirées de cette étude par les universitaires, la matière première permettant de valider ou d'invalider ces conclusions n'étant pas disponible.

Cette situation ne se limite pas aux mails. Nous avons évoqué précédemment le cas des données collectées sur les réseaux sociaux et on peut aussi citer les travaux sur des compte-rendus d'entretiens médicaux ((Amblard & Fort, 2014)) et nombre d'autres domaines encore pour lesquels les travaux en TAL ne peuvent être scientifiquement étayés par la publication des

corpus d'étude. Il n'est pas question de mettre en cause l'interdiction de cette diffusion mais bien de pointer la difficulté scientifique que cet interdit implique.

Cette mise en porte-à-faux des scientifiques porte peut-être en elle une piste de solution. En effet, si l'on excepte les applications de renseignement, les industriels ne sont pas a priori demandeurs de corpus de données à caractère nominatif. Ils sont demandeurs d'études linguistiques, de modélisations, voire de modèles effectifs pour des systèmes d'apprentissage. Leur tâche est d'industrialiser ces modélisations pour les mettre en œuvre dans des systèmes opérationnels. Or pour que les modèles puissent voir le jour, il faudrait que les scientifiques (et a priori seulement eux) aient accès aux corpus permettant de fabriquer ces modèles. Mais force est de constater qu'aujourd'hui, la fabrication de tels modèles dans un cadre industriel ne peut s'effectuer qu'au prix d'une activité scientifique affaiblie, voire sans étude scientifique du tout.

5 Résoudre le paradoxe

Pour faire du TAL avec des données personnelles, il faut que plusieurs questions soient résolues, en premier lieu délimiter ce qu'est une donnée personnelle. Une définition juridique existe, mais sa portée éthique n'est pas toujours claire et est de plus variable pour chacun. Le deuxième pas est d'outiller l'anonymisation vraie, qui est un objectif asymptotique et mouvant car les techniques de réidentification vont nécessairement évoluer. Enfin, fournir un cadre juridique à leur utilisation. La solution n'est pas uniquement technique, pas plus qu'elle n'est uniquement légale ou éthique, mais une combinaison de tous ces axes :

1. la technique doit assurer qu'un niveau suffisant de traitement a été réalisé sur les données pour rendre la réidentification la plus complexe possible ;
2. la licence associée aux données doit explicitement indiquer ce qui est possible ou non d'en faire afin de placer son utilisation dans un cadre juridique précis ;
3. enfin, la transparence du processus depuis la collecte des données jusqu'à leur fourniture doit permettre d'en assurer l'aspect éthique.

La Charte Ethique et Big Data fournit un cadre méthodologique permettant de décrire ces trois aspects, de fournir un descriptif qui accompagne la donnée tout au long de sa vie. On pourra alors, si le droit, les techniques ou le point de vue éthique évoluent, reconsidérer de façon éclairée l'usage que l'on peut faire d'un tel jeu de données.

Mais, on l'a vu, l'état actuel de la technologie ne permet pas de réutiliser une donnée pour le monde scientifique dès lors qu'elle possède un caractère personnel. Peut-être faut-il inventer un droit spécifique qui ferait de l'objet corpus un ensemble détaché des traces individualisables, de pouvoir le considérer comme l'on considère les résultats des cohortes en recherche médicale. Mais cela demanderait de résoudre un autre paradoxe : comment faire qu'un ensemble porte moins d'informations que la somme de ses parties ?

Remerciements

Le projet ODISAE est financé par BPIfrance et la Région Ile de France, dans cadre du 17e Fonds Unifié Interministériel (FUI). Le projet réunit l'Aproged, La Cantoche Productions, le Centre Départemental du Tourisme de l'Aube, Jamespot, Kwaga, Eptica, l'INSEE, le Laboratoire d'Informatique de Nantes-Atlantique (LINA) et TokyWoky.

Références

AMBLARD M. & FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. In *TALN - Traitement Automatique des Langues Naturelles*, p. 292–303, Marseille, France.

AMBLARD M., FORT K., MUSIOL M. & REBUSCHI M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

BOYD D. & CRAWFORD K. (2011). Six provocations for big data. In *A Decade in Internet Time : Symposium on the Dynamics of the Internet and Society*.

COUILLAULT A. & FORT K. (2013). Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasbourg, France. 4 pages.

DE MAZANCOURT H., COUILLAULT A. & RECOURCÉ G. (2014). L'anonymisation, pierre d'achoppement pour le traitement automatique des courriels. In *Journée d'Etude ATALA Ethique et TAL*, Paris, France.

DE MONTJOYE Y.-A., HIDALGO C., VERLEYSEN M. & BLONDEL V. (2013). Unique in the crowd : The privacy bounds of human mobility. *Sci. Rep.*, **3**.

ESHKOL-TARAVELLA I., KANAAN-CAILLOL L., BAUDE O., DUGUA C. & MAUREL D. (2014). Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO. In *Journée ATALA éthique et TAL*, Paris, France.

EUROPEAN COMMISSION (2014). The EU framework programme for research and innovation : How to complete your ethics self-assessment.

FORT K., ADDA G., SAGOT B., MARIANI J. & COUILLAULT A. (2014). Crowdsourcing for Language Resource Development : Criticisms About Amazon Mechanical Turk Overpowering Use. In Z. VETULANI & J. MARIANI, Eds., *Human Language Technology Challenges for Computer Science and Linguistics*, p. 303–314. Springer International Publishing.

KLIMT B. & YANG Y. (2004). Introducing the Enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*.

MEDLOCK B. (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).