Dé-identification d'un corpus clinique pour le traitement automatique du français
De-identification of a corpus of clinical documents for French bioNLP

Aurélie Névéol[a], Cyril Grouin[a], Stéfan Darmoni[b,c], Pierre Zweigenbaum[a]

[a] *LIMSI-CNRS UPR 3251, Rue John von Neumann, 91400 Orsay, France*
[b] *Équipe CISMeF, CHU de Rouen, 1, rue de Germont, 76000 Rouen, France*
[c] *GCSIS, LITIS EA 4051, Équipe TIBS, Université de Rouen, France*

Abstract

Background*: Much clinical information is contained in the text of patient Electronic Health Records and is not directly accessible for automatic computation. Natural Language Processing (NLP) techniques have been successfully developed to extract information from text and convert it to machine-readable representations. While many tools are available for processing English text, resources for other languages are lacking.* ***Objective****: To create a de-identified corpus of clinical text in French in order to enable the development and test of clinical NLP tools.* ***Methods:*** *A large number of patient records are obtained from a major French hospital. State-of-the-art text processing techniques are applied for de-identification and conversion of the records into a high-quality machine readable text format.* ***Results****: 100 records have been manually validated for further processing. Improvements were made to the MEDINA de-identification tool to ensure quality mass processing of the remaining records.* ***Conclusion****: The release of a high-quality de-identified corpus of clinical text in French is forthcoming to support the development and test of clinical NLP tools in French.*

Keywords :

Algorithms; Confidentiality; Information Retrieval and Storage/methods; Electronic Health Records; France; Natural Language Processing

Introduction

Dans le domaine biomédical, les informations cliniques et institutionnelles sont contenues dans le texte de publications scientifiques ou de dossiers patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, des méthodes de Traitement Automatique de Langue Naturelle (TALN) ont été développées avec succès afin d'extraire des informations pertinentes des textes libres et de les convertir en représentations structurées exploitables par l'homme et par la machine. Cependant, peu d'outils exploitant ces méthodes sont actuellement disponibles pour le Traitement Automatique de la langue biomédicale en français. Il manque en particulier un étiqueteur syntaxique ou extracteur de concepts UMLS® (Unified Medical Language System®). La disponibilité d'un étiqueteur syntaxique dédié aux textes du domaine biomédical en français permettrait d'améliorer les performances d'outils ciblant d'autres taches plus complexes qui peuvent exploiter les étiquettes syntaxiques, comme par exemple la reconnaissance d'entités nommées, l'extraction de concepts et l'indexation automatique. Les méthodes de TALN mises en œuvre par l'état de l'art reposent sur la création de règles ou sur l'application d'algorithmes d'apprentissage. Dans les deux cas, il est crucial de disposer d'un corpus annoté avec le type d'entités qui devront à terme être reconnues automatiquement par les outils: étiquettes morpho-syntaxiques, entités nommées, etc. Ce travail a pour but de développer un corpus clinique du français mis à disposition de la communauté scientifique.

Matériel et Méthodes

Dossiers patients issus d'un hôpital public français

Depuis l'avènement du dossier patient informatisé (DPI), la recherche d'information à l'intérieur du DPI est devenue une préoccupation nationale [1]. Afin de développer et tester des algorithmes de recherche d'information dédiés, un corpus comprenant 1 000 dossiers patients avec un historique d'au moins 50 séjours hospitaliers a été créé en 2009 avec l'approbation de la CNIL. Ce corpus comprend 130 000 documents en langue naturelle décrivant certains aspects du parcours médical du patient: compte-rendu de séjour, compte-rendu d'examen, compte rendu de consultation… Les documents sont au format .doc. Le nom des patients tels que connus dans le système hospitalier ont été remplacés par "XX XX", les dates de naissance par jj/mm, laissant l'année de naissance intacte et les numéros de dossier sont supprimés. Dans un premier temps, nous avons sélectionné un échantillon aléatoire de 100 documents afin de valider une chaîne de pré-traitement et d'anonymisation.

Anonymisation et pré-traitement des documents

En l'absence de cadre légal précis en France pour la dé-identification de données médicales destinées à être diffusées à des fins de recherche, nous nous appuyons sur les critères dé-

finis par le HIPAA (Health Insurance Portability and Accountability Act) aux États-Unis [2], qui sont pris en compte par les outils de dé-identification développés pour l'anglais [3]. L'outil MEDINA [4] a été appliqué sur le corpus converti au format texte (à l'aide des outils antidoc et catdoc) afin d'identifier les portions à dé-identifier. Le balisage de MEDINA a été validé manuellement afin de garantir l'anonymisation des documents. Les portions sensibles sont ensuite ré-identifiées, c'est-à-dire remplacées par des informations équivalentes préservant l'intégrité du texte en langue naturelle tout en garantissant la confidentialité des données médicales.

La figure 1 illustre le travail réalisé sur un document fictif (les dates et les noms du document original ont été modifiés).

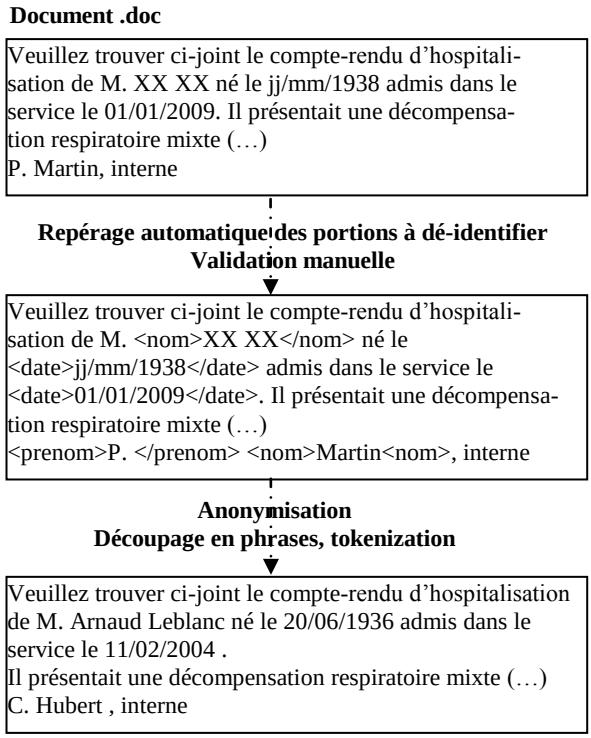

Figure 1- Pré-traitement d'un extrait de document fictif

Un pré-traitement supplémentaire s'est révélé nécessaire afin de formater le texte dans le cas où le document d'origine contenait des lignes segmentées et non des paragraphes de texte. Finalement, une segmentation en phrases a été effectuée à l'aide de l'outil OpenNLP, ainsi qu'une tokénisation simple.

Résultats

Le tableau 1 présente les performances du repérage des portions à anonymiser dans les documents, ainsi que le typage de ces portions: noms, adresse, ville, etc. La validation manuelle est utilisée comme référence. Différentes configurations de MEDINA ont été évaluées: *(i)* le module à base de règles, *(ii)* le module d'apprentissage (Conditional Random Fields - CRF) entrainé sur des données issues d'un hôpital autre que celui du corpus de test et *(iii)* le module d'apprentissage entrainé sur les données du corpus de test (au moyen d'une validation croisée répétée 10 fois – VC10).

Tableau 1 – Performance de la dé-identification automatique

	Précision	Rappel	Fmesure (%)
Médina-règles			
Repérage	86,20	82,50	84,60
Typage	84,60	80,40	82,40
CRF-autre			
Repérage	92,86	79,80	85,84
Typage	52,92	42,79	47,32
CRF-test VC10			
Repérage	**99,06**	**93,44**	**96,17**
Typage	**95,93**	**87,57**	**91,56**

Discussion et Conclusion

Dé-identification: les CRF renvoient les meilleurs résultats pour le repérage ainsi que pour le typage si un corpus d'entrainement suffisamment similaire au corpus de test est disponible. Les règles ont des performances stables par rapport aux tests effectués sur d'autres corpus et ont servi de base à la création de la référence. Ce travail a également permis d'étendre la fonctionnalité de MEDINA qui consiste à remplacer les portions à dé-identifier par un texte équivalent, préservant la cohérence du document en langue naturelle.

Perspectives: Nous poursuivons ce travail par la dé-identification automatique d'une plus grande portion du corpus et la création d'annotations morpho-syntaxiques et cliniques. Le corpus final sera mis à disposition de la communauté.

Remerciements

Ce travail a bénéficié en partie du financement ANR TecSan RAVEL (2012-2015). Les auteurs remercient P. Massari et B. Dahamna pour leur aide technique pour obtenir le corpus.

References

[1] Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, Campillo B, Jouini W, Grosjean J, Massari P, Griffon N, Dupuch M, Tayalati F, Dugas E, Balvet A, Grabar N, Pereira S, Frandji B, Darmoni S, Cuggia M. RAVEL: Retrieval And Visualization in ELectronic health records. *Stud Health Technol Inform.*2012;180:194-8.

[2] HIPAA Administrative Simplification Regulation Text 2006. § 164.514. Accessed on April 1, 2013 at: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf

[3] Ferrández Ó, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Generalizability and comparison of automatic clinical text de-identification methods and resources. *AMIA Annu Symp Proc*. 2012;2012:199-208.

[4] Grouin C, Zweigenbaum P. Une approche à plusieurs étapes pour anonymiser des documents médicaux. *In: RSTI-RIA.* 2011. 25(4):525-549. Hermès-Lavoisier