KOPTIENT Anaïs - M2 LTTAC
Présentation pour le cours de
Linguistique et TAC

Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers

Cyril Grouin, Arnaud Rosier, Olivier Dameron, Pierre Zweigenbaum

STRUCTURE

- Contexte et état de l'art
- Matériel
- Méthodes
- Résultats
- Discussion et conclusion

CONTEXTE ET ÉTAT DE L'ART

Contexte

- Surtout aux États-Unis car il existe la Protected Health Information : liste de 18 éléments qu'il faut anonymiser.
- En France, ça n'existe pas, donc les auteurs reprennent des éléments de la liste de la PHI : noms, prénoms, dates et âges.
- 2007 : compétition sur l'anonymisation des comptes-rendus d'hôpitaux. Les meilleurs ont obtenu une F-mesure >98%.

CONTEXTE ET ÉTAT DE L'ART

Les systèmes déjà existants

- DE-ID : identifie les éléments à anonymiser grâce à un dictionnaire et des expressions régulières. Rappel = 96,7% et Précision = 74,9%.
- MeDS : utilise aussi expressions régulières et listes. Rappel sur les noms propres = 99,47%, Rappel sur les autres éléments = 96,93% ; Précision = 92%.

MATÉRIEL

Corpus de comptes-rendus

- Utilisation des unités fonctionnelles (UF) d'un centre cardio-pneumologique.

- Extraction des comptes-rendus avec création d'un fichier contenant le texte et d'un fichier XML qui contient les métadonnées (données d'identification du patient et de la provenance du document) => pour les anonymiser à la source.

MATÉRIEL

L'anonymiseur DE-ID

- Composé d'un ensemble de scripts PERL
- 3 étapes d'anonymisation:
  - utilisation de dictionnaires (médicaux et de langue) et de listes d'entités nommées.
  - utilisation de listes de déclencheurs.
  - utilisation d'expressions régulières
- Création (automatique) d'un identifiant unique pour chaque mot anonymisé.

MATÉRIEL

Résultats de DE-ID

- Noms de cliniciens : 99,5% de Rappel et 72,5% de Précision.
- Dates avec années : 76,1% de Rappel et 71,3% de Précision.
- Globalement : 96,7% de Rappel et 74,9% de Précision.
- Faux-positif = faible (19,72 f-p pour 100 000 mots), mais élevé pour les dates (7,769 f-p pour 100 000 mots) et pour les noms (1,494 f-p pour 100 000 mots).

MÉTHODES

Anonymisation à la source

- 1ere passe d'anonymisation lors de la collecte des textes. Simple mais permet de supprimer des informations essentielles : nom, prénom et date de naissance du patient.
- Algorithme :
  - remplace le nom, prénom et nom marital du patient par <Nom_patient>, <Prénom_patient>, <Nom_marital>
  - remplace la date de naissance par <Date_de_naissance>

MÉTHODES

Conversion partielle en français de DE-ID

- Dictionnaires et listes : récupération de dictionnaires et listes d'entités nommées sur le site de l'ABU. Problème : toponymes désaccentués, liste des prénoms contient des noms anglo-saxons.
- Listes de déclencheurs : traduction des listes de DE-ID et éventuelle complétion.
- Expressions régulières : les plus difficiles à adapter car elles dépendent de la manière dont le logiciel a été conçu.

MÉTHODES

Conception d'un nouvel anonymiseur : medina

- MEDical INformation Anonymization.
- D'abord anonymisation des noms, prénoms, dates et villes :
  - Expressions régulières pour les entités numériques,
  - Dictionnaires et listes pour les entités nommées.
- Puis seconde anonymisation fondée sur le voisinage de ce qui avait déjà été anonymisé.
- Limite : les données ne proviennent que d'un seul hôpital.

MÉTHODES

Mode d'évaluation

- Comparaison des balises attendues avec celles produites.
  - Rappel (nombre de balisages corrects sur le nombre de balisages attendus).
  - Précision (nombre de balisages corrects sur le nombre de balisages produits).
- Pour comparer : ont relevé chaque balise avec leur contexte immédiat droite et gauche (quatre caractères)

MÉTHODES

Mode d'évaluation et ses limites

- Prénom composé : le système retournera deux balises => l'évaluation considérera que c'est un échec alors que non.
- Patronymes qui sont des prénoms : l'anonymiseur pourrait considérer que le nom de famille est un prénom, donc il le balisera comme prénom => l'évaluation considérera que c'est faux alors qu'il a bien été anonymisé.

RÉSULTATS

Tableau 1 : Comparaison des balisages des différents outils d'anonymisation.

	Rappel	**Précision**	**Corrects**	**Ramenés**	**Attendus**
Premier niveau	0,91	1,00	73	73	80
DE-ID francisé	0,63	0,25	103	402	163
Medina	0,85	0,91	139	152	163

RÉSULTATS

Les erreurs de Médina

Tableau 2 : Exemples de résultats de Medina. Note : les informations identifiantes de ces exemples ont été modifiées manuellement avant d'être copiées dans le tableau.

Étape	Énoncé
Anonymisation réussie	
Entrée	*J'ai examiné en consultation Madame <Nom marital patient> Michèle, née le 13.1.1943, âgée de 62 ans, pour le contrôle annuel de son stimulateur double chambre.*
Sortie	*J'ai examiné en consultation Madame <nom /> <prenom />, née le <date />, âgée de <age />, pour le contrôle annuel de son stimulateur double chambre.*
Sur-anonymisation	
Entrée	*(le pace maker est actuellement réglé en VVI à une fréquence de 52/mn).*
Sortie	*(le <prenom /> maker est actuellement réglé en VVI à une fréquence de 52/mn).*
Sur- (*Pace Maker*) et sous- (*PECRESSE*) anonymisation	
Entrée	*Le Pace Maker avait été contrôlé au mois de janvier par le PR PECRESSE et ne montrait pas de signe d'usure.*
Sortie	*Le <prenom /> <nom /> avait été contrôlé au mois de <date /> par le PR PECRESSE et ne montrait pas de signe d'usure.*
Autres anonymisations réussies, hors évaluation	
Entrée	*un épisode d'IVG justifiant d'un traitement par Lasilix et la réduction de la posologie de Soprol à 1/jour.*
Sortie	*un épisode d'IVG justifiant d'un traitement par <medicament /> et la réduction de la posologie de <medicament /> à 1/jour*

DISCUSSION & CONCLUSION

- Pas de sur-anonymisation avec l'anonymisation à la source
- Francisation de DE-ID : pas pu reproduire autant de ressources qu'il y avait dans la version anglaise. Il aurait fallut changer le coeur du système, mais trop long à faire.
- Seul un petit échantillon du corpus a été anonymisé manuellement.
- Pour augmenter le rappel de Medina : utiliser des ressources complémentaires.

MERCI POUR VOTRE ATTENTION

Vous pouvez poser vos questions