Traitement de la langue biomédicale au LIMSI

Christopher Norman[1] Cyril Grouin[1] Thomas Lavergne[2]
Aurélie Névéol[1] Pierre Zweigenbaum[1]
(1) LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
(2) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ

Nous proposons des démonstrations de trois outils développés par le LIMSI en traitement automatique des langues appliqué au domaine biomédical : la détection de concepts médicaux dans des textes courts, la catégorisation d'articles scientifiques pour l'assistance à l'écriture de revues systématiques, et l'anonymisation de textes cliniques.

ABSTRACT

Biomedical language processing at LIMSI

We propose demonstrations of three natural language processing tools developed at LIMSI for applications in the biomedical domain : medical concept detection in short texts, scientific paper categorization to assist systematic review authors, clinical text de-identification.

MOTS-CLÉS : domaine médical ; classification de textes ; extraction d'information ; désidentification ; revues systématiques.

KEYWORDS: medical domain; text classification; information extraction; de-identification; systematic reviews.

1 Détection de concepts dans des textes courts

Nous présentons un système qui effectue de la détection de concepts dans des textes courts (Zweigenbaum & Lavergne, 2017). Il est spécialisé dans la détermination de codes diagnostiques de la Classification internationale des maladies (CIM-10) de l'OMS pour des certificats de décès. Il combine deux techniques classiques : l'application d'un dictionnaire spécialisé et un apprentissage supervisé, entraîné sur plus de 300 000 exemples de diagnostics. La démonstration montre la détection de diagnostics dans des textes courts, ainsi que l'apport comparé de différentes méthodes d'hybridation entre dictionnaire et apprentissage supervisé :

— le calibrage du dictionnaire par apprentissage supervisé sur le corpus d'entraînement ;
— la prise en compte du dictionnaire comme attributs dans la classification ;
— la fusion des résultats du dictionnaire et de l'apprentissage supervisé.

Le système est démontré en français et en anglais. Sur les données de la campagne d'évaluation CLEF eHealth 2017 [1], la version française produit actuellement des résultats meilleurs que le meilleur système participant et la version anglaise est juste derrière les résultats du meilleur système participant.

1. https://sites.google.com/site/clefehealth2017/task-1

2 Assistance à l'écriture de revues systématiques

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Nous présentons un système de classification automatique d'articles présélectionnés à l'aide d'une requête soumise à un moteur de recherche (Norman *et al.*, 2017). Le système propose un ordonnancement des articles par ordre de pertinence par rapport aux critères d'inclusion définis par un corpus d'entraînement. La mise en œuvre ainsi que les résultats de cet outil de classification peuvent être exploités au travers de diverses interfaces graphiques. À titre d'exemple, nous présentons l'interface BibReview, utilisée dans le cadre du Yearbook of Medical Informatics (Névéol & Zweigenbaum, 2016) qui rapporte chaque année les résultats de revues de la littérature dans seize sous-domaines de l'informatique biomédicale. Cet outil permet de visualiser les articles de la préselection dans l'ordre proposé par le classifieur, de valider et de visualiser la sélection d'articles pour inclusion dans une revue de la littérature.

3 Désidentification de comptes rendus cliniques

MEDINA (MEDical INformation Anonymization) (Grouin, 2013) est un outil permettant de désidentifier (anonymiser) les informations potentiellement identifiantes contenues dans des comptes rendus cliniques. L'outil identifie les informations par type, puis remplace les informations identifiées par des pseudonymes (noms, prénoms, adresses, téléphones, etc.) et par un décalage des dates dans le passé (permettant de conserver les écarts temporels entre dates). Le résultat produit permet de bénéficier de textes représentatifs du domaine clinique sur lesquels réaliser des études, sans que la vie privée des patients ne soit mise en cause.

La démonstration proposée consiste à présenter les différentes étapes du système à base de règles et de lexiques pour anonymiser des documents cliniques.

Références

GROUIN C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, France.

GROUIN C., DELÉGER L., MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., GRAU B., ROSSET S. & ZWEIGENBAUM P. (2011). Extraction d'informations médicales au LIMSI. In *Démonstrations, TALN 2011*, Montpellier.

NORMAN C., LEEFLANG M., ZWEIGENBAUM P. & NÉVÉOL A. (2017). Tri automatique de la littérature pour les revues systématiques. In *TALN 2017*, Orléans, France : Association pour le Traitement Automatique des Langues.

NÉVÉOL A. & ZWEIGENBAUM P. (2016). Clinical natural language processing in 2015 : Leveraging the variety of texts of clinical interest. *Yearb Med Inform*, p. 234–239.

ZWEIGENBAUM P. & LAVERGNE T. (2017). Détection de concepts et granularité de l'annotation. In *TALN 2017*, Orléans, France : Association pour le Traitement Automatique des Langues.