Déterminer la difficulté de termes médicaux en contexte

Anaïs Koptient, Natalia Grabar

CNRS, Univ Lille, UMR 8163 - STL, F-59000 Lille, France
prénom.nom@univ-lille.fr

1 Introduction

Les documents du domaine médical contiennent des termes techniques, comme *hémopéricarde*, *iridomalacie* ou *blépharite*, qui ont une sémantique opaque. Les patients ont souvent des difficultés à comprendre et à utiliser de tels termes. Pourtant, la compréhension de ces notions est cruciale pour les patients car elle est liée à leur santé et bien-être. La détection des termes et séquences qui peuvent présenter des difficultés de compréhension pour les patients est donc une problématique importante, avec l'objectif de leur simplification par la suite. La plupart des travaux portant sur la détection de termes difficiles consistent en l'utilisation des classifieurs avec comme descripteurs la fréquence, les propriétés sémantiques, lexicales, morphologiques et psycholinguistes des termes et leur contexte (Yimam et al., 2018).

2 Approche

Nous utilisons 100 cas cliniques choisis aléatoirement au sein du corpus CAS (Grabar et al., 2018). Les cas cliniques sont des documents médicaux semblables aux comptes-rendus d'hospitalisation. Ils décrivent les antécédents des patients, la raison de leur visite à l'hôpital, les procédures et traitements effectués et l'issue. Les cas cliniques traitent de sujets et spécialités différents. Les cas sont analysés syntaxiquement avec l'analyseur Cordial (Laurent et al., 2009) pour les segmenter en groupes syntaxiques. Nous obtenons ainsi 13 918 groupes. Les cas sont ensuite annotés manuellement par huit annotateurs (niveau master mais sans connaissances médicales en particulier) avec objectif de marquer si les groupes syntaxiques sont compris ou pas. Au total, 7 225 groupes syntaxiques sont compris par les annotateurs et 2 129 groupes ne sont pas compris. Ce petit corpus constitue nos données de référence.

Nous utilisons une méthode supervisée avec le classifieur Random Forest tel qu'implémenté dans Scikit-Learn (Pedregosa et al., 2011) pour déterminer la difficulté des groupes syntaxiques. Nous utilisons des descripteurs internes (nombre de lettres, de syllabes, de phonèmes, fréquence, etc.) aux mots et des descripteurs externes (contextes de droite et gauche) pour l'entraînement et l'évaluation. Nous obtenons un modèle bi-classe. Nous effectuons plusieurs expériences : exploitation de descripteurs externes uniquement, de descripteurs internes uniquement et combinaison des descripteurs externes et internes. Chaque expérience est évaluée par validation croisée à dix plis, ce qui permet de calculer la précision, le rappel et la F-mesure. Nous effectuons également un test d'ablation des descripteurs pour mieux évaluer leur importance individuelle pour la tâche.