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.

3 Résultats

Nous obtenons la F-mesure de 0,854 (descripteurs externes), 0,822 (descripteurs internes) et 0,858 (combinaison des deux). Ces deux ensembles de descripteurs fournissent donc des informations complémentaires. De plus, nous voyons que le contexte est un indicateur important pour la détection de passages difficiles à comprendre : les descripteurs externes (seuls ou avec descripteurs internes) fournissent de meilleurs scores. Le test d'ablation indique que la fréquence dans la langue générale (Lexique3 (New et al., 2001)), le nombre de syllabes et la cohérence phonèmes/orthographe sont des descripteurs importants : la F-mesure descend alors à 0,809 et 0,811, respectivement. Les séquences détectées comme difficiles à comprendre pourront être simplifiées par la suite.

Remerciements. Ce travail s'inscrit dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR (ANR-17-CE19-0016-01).

Références

Grabar, N., V. Claveau, et C. Dalloux (2018). Cas : French corpus with clinical cases. In *LOUHI 2018*, Bruxelles, Belgique, pp. 1–12.

Laurent, D., S. Nègre, et P. Séguéla (2009). L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.

New, B., C. Pallier, L. Ferrand, et R. Matos (2001). Une base de donnÃ©es lexicales du franÃ§ais contemporain sur internet : Lexique//a lexical database for contemporary french : Lexique. *Annee Psychologique - ANNEE PSYCHOL 101*, 447–462.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Yimam, S. M., C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, et M. Zampieri (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, pp. 66–78. Association for Computational Linguistics.

Summary

Medical documents contain terms with specific meaning, which usually present understanding problems for patients. For a better simplification of medical documents, it is important to first detect difficult to understand sequences. We propose to address this issue as classification problem. Our objective is to predict sequences difficult to understand within syntactically parsed documents. We use internal and external features, and their combination. The results obtained show F-mesure values over 0.8. Ablation test indicates that among the most relevant features we can find frequency in a general-language corpus, phoneme/spelling coherence, and number of syllables.