Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales

Karim El Haff[*,***] Wissam Antoun[**], Florence Le Ber[*] Véronique Pitchon[***]

[*] Université de Strasbourg, ENGEES, CNRS, UMR 7357 ICube, F 67000 Strasbourg
kelhaff@unistra.fr, florence.le-ber@unistra.fr
[**] Inria-Paris, 75012 Paris, France
wissam.antoun@inria.fr
[***] Université de Strasbourg, CNRS, UMR 7044 Archimède, F 67000 Strasbourg
pitchon@unistra.fr

Résumé. Aujourd'hui, de nombreux projets se focalisent sur l'application des technologies linguistiques sur des corpus de médecine moderne surtout en matière de reconnaissance des entités nommées. Par ailleurs, les pharmacopées anciennes sont explorées avec une saisie manuelle des données par des spécialistes d'histoire et de biologie pour en retirer des connaissances. Ces analyses sont réalisées sans nécessairement passer par la reconnaissance des entités nommées, ce qui pourrait pourtant accélérer l'exploration des manuscrits. Par conséquent, nous proposons ici un mariage entre les deux pratiques par : (1) la création d'un ensemble de données de reconnaissance d'entités nommées pour les traductions anglaises de pharmacopées arabes médiévales et (2) l'entraînement et l'évaluation de modèles de langue pré-entraînés sur plusieurs domaines.

1 Introduction

Les progrès réalisés dans le traitement automatique du langage naturel (TAL) ou *Natural Language Processing* (NLP), une branche de l'intelligence artificielle qui permet aux ordinateurs d'analyser des textes écrits, parlés ou imagés, permettent d'extraire et de traiter les informations d'un corpus dans une langue humaine. Ce type de technologie peut être utilisé pour effectuer diverses tâches telles que la traduction automatique, l'exploration de textes, la reconnaissance d'entités nommées, la synthèse automatique de textes, la simplification automatique de textes, l'analyse de sentiments, les chatbots intelligents et d'autres applications qui pourront répondre aux besoins d'exploration de corpus. Les technologies du TAL s'appliquent dans de nombreux domaines d'intérêt majeur, tel que la médecine, où de nouveaux médicaments sont sans cesse recherchés.

Dans ce projet, nous nous focalisons sur une application du TAL dans le monde médical et historique, et plus précisément, la reconnaissance des entités nommées dans les pharmacopées de la civilisation arabe médiévale.

La période médiévale, surtout en Europe, est considérée comme une période sombre de l'histoire. De ce fait, la médecine médiévale est souvent négligée, étant perçue comme pleine