Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

SIMI : un système de suggestion de littérature médicale

Pierre Jourlin
LIA, Avignon Université, Avignon, France
pierre.jourlin@univ-avignon.fr

Résumé

Nous faisons la démonstration de SIMI, un système de suggestion de littérature médicale entièrement automatisé. À partir d'une description d'un cas clinique en français, SIMI extrait les termes médicaux présents en résolvant simultanément les éventuelles ambiguïtés. Il traduit alors les termes en anglais et construit une requête de recherche documentaire qui comprend les éventuels synonymes et hyponymes des termes originaux. Cette requête permet de retrouver et catégoriser les documents pertinents issus d'une base de plusieurs dizaines de millions de notices bibliographiques multilingues français-anglais. Ce système a été développé dans le cadre d'un transfert technologique associant une université, une société d'accélération de transfert technologique et une société qui commercialise une solution de téléexpertise médicale.

Abstract

SIMI: A recommender system of medical literature

We present SIMI, a fully automated recommender system of medical literature. SIMI extracts the medical terms from a clinical case report in French, while simultaneously resolving any ambiguities. Then, it translates them into English and constructs a search query that includes possible synonyms and hyponyms of the original terms. This query allows the retrieval and categorization of relevant documents from a database of several tens of millions of multilingual bibliographic records in French and English. SIMI was developed as part of a technology transfer project involving a university, an organization for the acceleration of technology transfer and a company that markets a telemedicine system.

Mots-clés : Recherche documentaire ; désambiguïsation ; ontologie de termes médicaux
Keywords: Document retrieval; disambiguation; ontology of medical terms

1 Du cas clinique à la requête

Nous reprenons les méthodes de désambiguïsation et d'extraction de termes médicaux dans un cas clinique décrit en français qui ont été brièvement présentées dans (Murata et coll., 2021) et qui s'appuient sur un cadre plus théorique décrit dans (Jourlin, 2022). L'objet de cette démonstration est de présenter une application industrielle qui met à profit ces méthodes.

Dans ce contexte, un lien a été établi entre une université, une société d'accélération de transfert technologique et une entreprise qui propose une plateforme de partage entre médecins. Cette dernière permet d'échanger au sujet de patients complexes et solliciter des avis surspécialisés auprès des médecins experts indépendants, mais utilisateurs de la plateforme.

L'objectif de ce développement logiciel est de permettre une recherche instantanée entre les cas présentés sur la plateforme et ceux déjà publiés dans la littérature scientifique, afin de fournir un

support scientifique aux experts décideurs, d'une antériorité à ce type pathologie dans la littérature scientifique.

2 Lexiques et ontologies

La démonstration s'appuie sur un lexique issu du MeSH bilingue français-anglais, maintenu par l'INSERM. Le lexique a été ensuite été contextualisé grâce à l'application SIDRES (Murata et coll., 2021), ce qui permet d'ignorer les usages non médicaux de ces termes, notamment dans les expressions courantes de la langue française (« au **sein** de mon établissement », « au **cœur** du problème », etc.). La contextualisation permet également de :

- Faire la distinction entre mention d'un symptôme actuel et mention d'un antécédent.
- Faire la distinction entre mention d'une présence et mention d'une absence d'un symptôme actuel ou d'un antécédent.
- Associer des termes à certaines expressions, par exemple l'empan de texte « patiente âgée de 30 ans » donnera lieu à l'extraction de deux termes, un représentant la catégorie d'âge « adulte/adult » et l'autre la catégorie « femme/female ».

La même fonction logicielle est utilisée pour d'un côté, extraire les termes originaux de la requête à partir de la description clinique et pour d'un autre côté, extraire les termes médicaux présents dans les titres, les résumés et le cas échéant le texte complet des articles scientifiques. Les termes MeSH présents dans les notices bibliographiques sont en revanche utilisés avec une seule normalisation de casse.

3 Notices bibliographiques

Pour les besoins de la démonstration, nous avons téléchargé et indexé les notices bibliographiques de Pubmed, DOAJ et EuroPMC. Le lexique contextualisé non compressé a une taille de 289 Mo. Il contient plus de 300 000 termes médicaux rangés dans 37 catégories sémantiques distinctes. Un peu plus de 28 millions d'articles postérieurs à 1990 constituent un index de notices bibliographiques d'une taille compressée inférieure à 4 Go. Néanmoins, le fonctionnement en temps réel du logiciel nécessite dans son stade actuel de développement une architecture matérielle de type serveur dédié avec un minimum de 90 Go de mémoire vive disponible.

4 Interface

Une interface web permet d'entrer une description clinique. Pour les besoins de la démonstration, nous mettrons à disposition des descriptions cliniques réellement rédigées par des médecins, s'adressant à des confrères/consœurs. Après soumission, le système renvoie une liste de termes extraits et une première liste de documents potentiellement pertinents. L'interface permet en outre un raffinement de la requête, en suggérant à l'utilisateur d'ajouter des termes sélectionnés automatiquement sur la base de plusieurs critères et en permettant la suppression ou la modification des termes de la requête initiale.

Remerciements

L'auteur tient à remercier Julieta Murata et Rémy Carrette pour leur contribution déterminante au développement de ce projet durant l'année 2021, mais aussi Marguerite Leenhardt et Guillaume Gouvernet de la SATT Sud Est, ainsi que Emilie Mercadal et David Bensoussan de la société ROFIM, pour leur confiance et leur investissement dans le projet.

Références

MURATA J., CARRETTE R. & JOURLIN P. (2021). SIDRES : A Novel Annotation Tool For The Automatic Detection of Semantic Entities. Actes de *Traitement Automatique des Langues Naturelles*, Lille, France. pp.15-17. ⟨hal-03265913⟩

JOURLIN P. (2022) Disambiguation for the Classification of Lexical Items. France, Patent n° : EP3937059A1. ⟨hal-03598242⟩