Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Compléter des annotations humaines par des données synthétiques pour l'alignement d'entités biomédicales

Adam Remaki[1] Christel Gérardin[1,2] Eulàlia Farré-Maduell[3] Martin Krallinger[3]
Xavier Tannier[1]
(1) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Limics, 75006 Paris, France
(2) Service de médecine interne, Hôpital Tenon, Assistance Publique - Hôpitaux de Paris, Paris, France
(3) Barcelona Supercomputing Center, Barcelona, Spain
adam.remaki@etu.sorbonne-universite.fr,

RÉSUMÉ

Nous présentons **SynCABEL**, une méthode visant à réduire la dépendance aux annotations manuelles nécessaires à l'apprentissage supervisé de l'alignement d'entités biomédicales, en les complétant par des exemples synthétiques. SynCABEL exploite des LLMs pour générer des exemples d'entraînement riches en contexte couvrant l'ensemble des concepts candidats d'une base de connaissances cible, offrant ainsi une supervision plus large. En utilisant des modèles génératifs récents et une inférence guidée, notre approche établit de nouveaux états de l'art sur trois jeux de données de référence : MedMentions (anglais), QUAERO (français) et SPACCC (espagnol). En faisant varier la quantité de données annotées manuellement disponibles, SynCABEL atteint des performances comparables à un entraînement entièrement supervisé tout en réduisant jusqu'à 60% le volume d'annotations humaines nécessaires. Enfin, nous introduisons un protocole d'évaluation fondé sur un *LLM-as-a-judge*, qui montre que SynCABEL augmente la proportion de prédictions cliniquement valides.

ABSTRACT

SynCABEL : Synthetic Contextualized Augmentation for Biomedical Entity Linking

We present **SynCABEL**, a framework addressing a key bottleneck in supervised biomedical entity linking (BEL) : the scarcity of expert-annotated training data. SynCABEL leverages large language models to generate context-rich synthetic training examples for all candidate concepts in a target knowledge base, providing broad supervision without manual annotation. We demonstrate that SynCABEL, when combined with decoder-only models and guided inference establish new state-of-the-art results across three widely used multilingual benchmarks : MedMentions (English), QUAERO (French), and SPACCC (Spanish). By varying the amount of available manually annotated data, we show that SynCABEL achieves performance comparable to fully supervised training while reducing human annotation requirements by up to 60%. Finally, we introduce an LLM-as-a-judge evaluation protocol which demonstrates that SynCABEL increases the rate of clinically valid predictions.

MOTS-CLÉS : Alignement d'Entités Biomédicales, Augmentation de Données, Fouille de Textes.

KEYWORDS: Biomedical Entity Linking, Data Augmentation, Health Data Mining.

ARTICLE ACCEPTÉ À : AI4Health@IJCAI-ECAI 2026 : AI and Health Special Track of the 35th International Joint Conference on Artificial Intelligence, Bremen, Germany, August 15-21, 2026.
URL : https://2026.ijcai.org/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 6–6.