

Translating UMLS Concepts to Improve Medical Entity Linking in French: A SapBERT-Based Approach

Amaury FIERENS^a, Alexandre ENGLEBERT^{a,b}, and Sébastien JODOGNE^a

^a*ICTEAM, Louvain School of Engineering, UCLouvain, Belgium*

^b*Cliniques Universitaires Saint-Luc & NMSK Lab, IREC, UCLouvain, Belgium*

ORCID ID: Amaury Fierens <https://orcid.org/0009-0003-3668-8804>

Alexandre Englebert <https://orcid.org/0000-0002-9021-7806>

Sébastien Jodogne <https://orcid.org/0000-0001-6685-7398>

Abstract. Medical Entity Linking (MEL) refers to the task of automatically detecting and codifying medical concepts, which is an important preprocessing step for exploiting unstructured medical reports. SapBERT is an efficient MEL method for the English language. Unfortunately, its cross-language counterpart underperforms in French due to the lack of training data. To address this limitation, this paper explores the possibility of training SapBERT on machine-generated French translations of the UMLS Metathesaurus. Evaluation on the QuaeroFrenchMed benchmark demonstrates that this approach outperforms Cross-lingual SapBERT for SNOMED-CT MEL on the QuaeroFrenchMed benchmark.

Keywords. Medical Entity Linking, SNOMED-CT, SapBERT, Cross-language

1. Introduction

Hospitals and health centers generate vast amounts of unstructured textual data, such as medical reports, which are difficult to process and analyze. Consequently, Medical Entity Linking (MEL) focuses on automatically detecting and codifying medical concepts, transforming unstructured information into structured data. Over the last few years, efficient models like NeighBERT [1] or SapBERT [2] have been developed. While these methods work well for English, a significant challenge in applying them to other languages is their reliance on English data for training. In this paper, we explore the possibility of leveraging Machine Translation (MT) to improve the cross-lingual performance of the SapBERT approach. Specifically, MT is first used to translate the entire UMLS Metathesaurus into the target language, after which a SapBERT model is trained on this translation to identify the SNOMED-CT concepts.

2. Methods

SapBERT is a self-supervised machine learning model trained to align biomedical entity representations [2]. SapBERT uses synonyms from the multiple nomenclatures available in the UMLS Metathesaurus. It leverages contrastive learning techniques to

bring similar terms closer together and push dissimilar terms further apart in the vector space of embeddings generated by a BERT-like encoder. The original SapBERT model was trained on English synonyms. It has since been adapted to exploit synonyms from multiple languages, resulting in the Cross-lingual SapBERT model. While the latter model performs well on French biomedical tasks, its performance still lags behind its English counterpart due to the limited availability of French data in the UMLS Metathesaurus. To bypass this limitation, we first translated the entire English UMLS Metathesaurus into French using a state-of-the-art MT model (Opus-mt-en-fr from the Marian framework [3]), then trained SapBERT models on the translated corpus.

3. Results

Our models were evaluated on the development set of QuaeroFrenchMed [4]. The annotated UMLS concepts were converted to the SNOMED-CT US ontology using the mapping provided by the UMLS 2023AB distribution, excluding entities without a direct mapping. The embeddings generated by the different SapBERT models were then matched against the vector representations of the entities of the ontology using an approximate nearest neighbor search. Based on the resulting top predicted concepts, we computed the accuracy at top-1 and top-5 positions. Results show that the best French SapBERT model outperforms Cross-lingual SapBERT, with a 3.5% improvement in top-1 accuracy and 2.2% in top-5 accuracy. The PubMedBERT initialization delivers the best performance among the investigated BERT-like encoders, even though it is an English-based model. Results are summarized in the following table:

Model	MT Model	Encoder	Top-1 accuracy	Top-5 accuracy
Cross-lingual SapBERT	/	XLNet-RoBERTa base	81.09	92.84
French SapBERT (<i>ours</i>)	Opus-mt-en-fr	PubMedBERT	84.66	95.02
	Opus-mt-en-fr	CamemBERT-bio	75.50	91.35
	Opus-mt-en-fr	DrBERT	72.01	91.18

4. Discussion and conclusion

This paper demonstrates that combining SapBERT with Machine Translation offers a promising solution to the challenges of processing unstructured medical data in French. Future research will explore the usability of French SapBERT in clinical environments, its applicability to ICD-11 codification, and its extension to other languages.

References

- [1] Singh A, Krishnamoorthy S, Ortega JE. NeighBERT: Medical Entity Linking Using Relation-Induced Dense Retrieval. *J Healthc Inform Res.* 2024 Jun;8(2):353–69.
- [2] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2021. p. 4228–38.
- [3] Tiedemann J, Aulamo M, Bakshandaeva D, Boggia M, Grönroos SA, Nieminen T, et al. Democratizing neural machine translation with OPUS-MT. *Lang Resources & Evaluation.* 2024 Jun;58(2):713–55.
- [4] Névél A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French medical corpus: A resource for medical entity recognition and normalization. *Proc of BioTextMining Work.* 2014;24–30.