Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Les grands modèles de langue biomédicaux préentraînés sur des données hors EHR sont moins performants en contexte multilingue réel

Alina Kramchaninova, Clara L. Oeste,
Narges Farokhshad, Lucas Sterckx
LynxCare Inc, Leuven, Belgium.
{alina.kramchaninova,clara.oeste,narges.farokhshad,lucas.sterckx}@lynx.care

RÉSUMÉ

Des travaux récents ont démontré que les grands modèles de langue (LLMs) sont capables de traiter des données biomédicales. Cependant, leur déploiement en zéro-shot dans les hôpitaux présente de nombreux défis. Les modèles sont souvent trop coûteux pour une inférence et un ajustement local ; leur capacité multilingue est inférieure par rapport à leur performance en anglais ; les ensembles de données de préentraînement, souvent issus de publications biomédicales, sont trop génériques pour une performance optimale, compte tenu de la complexité des scénarios cliniques présents dans les données de santé. Nous abordons ces défis et d'autres encore dans un cas d'usage multilingue réel à travers le développement d'un pipeline de normalisation de concepts de bout en bout. Son objectif principal est de convertir l'information issue de dossiers de santé non structurés (multilingues) en ontologies codifiées, permettant ainsi la détection de concepts au sein de l'historique médical d'un patient. Dans cet article, nous démontrons quantitativement l'importance de données réelles et spécifiques au domaine pour des applications cliniques à grande échelle.

ABSTRACT

Biomedical LLMs Pretrained on Non-EHR Data Underperform in Multilingual Real-World Settings

Recent works have demonstrated that Large Language Models (LLMs) are capable of processing biomedical data. However, the zero-shot deployment of such models in hospitals presents considerable challenges. The models are often too expensive for on-site inference and fine-tuning ; their multilingual capacity is subpar compared to their performance in English ; pretraining datasets, often in the form of biomedical publications, are too generic for optimal performance, considering the complex clinical scenarios that exist in healthcare data. We address these and other challenges in a multilingual real-world use case through the development of an end-to-end concept normalization pipeline. Its main goal is to convert the information from (multilingual) unstructured health records into codified ontologies, enabling concept detection within a patient's medical history. We quantitatively demonstrate the importance of real-world, domain-specific data for scalable clinical applications.

MOTS-CLÉS : normalisation de concepts cliniques, grands modèles de langue biomédicaux.

KEYWORDS: discontinuous NER, biomedical entity linking, clinical concept normalization, EHR.

Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM), pages 25–41.

1 Introduction

Biomedical data have been valuable for the training of domain-specific Large Language Models (LLMs). For instance, the PubMed corpus [1], which comprises over 30 million citations from journals and books, has been used for the pretraining of various LLMs (e.g., BioBERT (Lee *et al.*, 2020)), as well as the composition of datasets for specific tasks, such as question answering (Jin *et al.*, 2019). While models trained on clean, well-organised text have demonstrated the potential to outperform human experts on tasks like summarization (Van Veen *et al.*, 2024), they often struggle with real-world applications (Krishnamoorthy *et al.*, 2024; Gallifant *et al.*, 2024) such as obtaining Real-World Evidence (RWE) from Electronic Health Records (EHRs).

There has been substantial effort to compose domain-specific datasets. MIMIC-III (LSAEW & Pollard, 2016), for instance, consists of de-identified patient records and was used for the pretraining of LLMs such as ClinicalBERT (Huang *et al.*, 2019) and ClinicalT5 (Lehman & Johnson, 2023). However, there are significant challenges to using such models with real-world data, because such data are often noisy, grammatically incorrect, domain- and caresite-specific, and dynamic (e.g., drug names introduced or deprecated over time).

Leveraging the knowledge encoded in LLMs (Singhal *et al.*, 2023) can be crucial for clinical applications such as concept normalization. This task can be defined as the discovery of relevant concepts in unstructured, free-form text, and the subsequent mapping of these concepts to a large set of standardized medical concept names and identifiers in extensive dictionaries such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT [2]) or the Unified Medical Language System (UMLS [3]). The resulting databases of standardized evidence can further be used for predictive analytics and other clinical research (Kraljevic *et al.*, 2021).

To illustrate the difficulty of the concept normalization task, we conducted an experiment involving GPT-4, [4] which is not a medical LLM per se but can be used as one (Nassiri & Akhloufi, 2024). Previously, we tested BioGPT (Luo *et al.*, 2022), Llama3-OpenBioLLM-8B (Ankit Pal, 2024), and BioMistral-7B (Labrak *et al.*, 2024), but none could correctly respond to the following prompt :

```
List all of the medical concepts (and their respective CUIs) in the following text: “Legs: no infection or itching, but sensitive. Pitting oedema right > left”. The concepts can be flat, nested and discontinuous.
```

We note that response quality varies with prompt wording, and GPT-4’s output can be inconsistent. From the most accurate predictions presented in Table 1 we confirm the model’s potential as an end-to-end normalization system. However, we observe three fallacies : (1) the list of the generated concepts is not complete ; (2) the simplicity of the generated concepts could introduce erroneous evidence (e.g., "infection" being a linked concept instead of "no infection") ; (3) out of the generated concept-identifier combinations, only 57.14% were synonymous with the descriptions of the generated UMLS codes.

To achieve the desired output presented in Table 2, we introduce an in-domain end-to-end (multilingual) concept normalization pipeline. Section 2 tackles the works related to its main components, while the design and deployment of the pipeline is discussed in section 3. We test this pipeline against

1. pubmed.ncbi.nlm.nih.gov
2. www.snomed.org/what-is-snomed-ct
3. www.nlm.nih.gov/research/umls/index.html
4. openai.com/index/gpt-4

Concept	CUI	CUI Description
Legs	C0023216	Lower limb structure
Infection	C0009450	Infectious disease
Itching	C0021147	Incentives
Sensitive	C0020555	Hypertrichosis
Pitting oedema	C0031039	Pericardial effusion
Right	C0205090	Right
Left	C0205091	Left

TABLE 1 – The concepts and the CUIs generated by GPT-4 in response to the prompt provided in Section 1. The CUI descriptions are for the respective CUIs as per SNOMED CT.

EHR	(D)NER	EL	CUI Description
Legs : no infection or itching, but sensitive. Pitting oedema right > left.	No infection	C2712105	Absence of signs and symptoms of infection
	Infection	C0745283	Infectious process
	No itching	C1276050	No sensation of itching
	Itching	C5700180	Pruritic disorder
	Legs sensitive	C0578113	Tenderness in lower limb
	Pitting oedema right	C5230912	Edema of right lower limb
	Pitting oedema left	C5230911	Edema of left lower limb
	Pitting oedema	C0333243	Pitting edema
	Oedema	C0013604	Edema

TABLE 2 – Visualization of an end-to-end concept normalization system. The (D)NER and EL columns represent the subsequent steps in the pipeline, with the synthetic EHR as the input. The last column provides standard nominations (preferred names) for each code as per the SNOMED CT ontology.

open-source models in section 4 and discuss its strengths and limitations in sections 5 and 6.

2 Related Work

2.1 Clinical (Discontinuous) NER

Traditionally, Named Entity Recognition (NER) is treated as labeling a sequence of tokens, embedded by a pretrained LLM such as BERT (Devlin *et al.*, 2019). Such models often require substantial task-specific fine-tuning data, which motivates the exploration of the zero- and few-shot capabilities of LLMs (McInerney *et al.*, 2023; Košprdić *et al.*, 2023).

Although there exist a number of NER approaches on clinical tasks (Verma *et al.*, 2023; Yazdani *et al.*, 2022), most named entities are flat and contain minimal semantic information. Due to the complexity of the clinical narrative (Liu *et al.*, 2022), we argue that the most impactful concepts are the most detailed ones, often consisting of sub-concepts that are not adjacent to one another (discontinuous), such as "pitting oedema left" or "no itching" presented in Table 2. While Discontinuous NER (DNER) is explored in the research community (Li *et al.*, 2021), there is less focus on biomedical DNER in particular.

Finally, we must highlight that attribute detection of named entities is an important research topic (Van Es *et al.*, 2023), particularly for clinical data. Certain concepts must be recognized as negated or hypothetical (e.g., "*presumably* palpitations" or "diabetes *?*") to avoid false positives, i.e. patients being misinterpreted as having (had) the originally negated/hypothetical findings.

2.2 Clinical Entity Linking

Entity Linking (EL) is the task of mapping textual mentions of concepts to codes from a reference knowledge base. Most approaches to the LLM-based EL, just like the (D)NER task, use transfer learning, which involves fine-tuning the LLM on domain-specific data (Sung *et al.*, 2020; Liu *et al.*, 2021a).

Ambiguity in normalization (i.e., multiple ways of referring to the same concept) is a prominent problem that encompasses the treatment of synonyms in EL (Vretinaris *et al.*, 2021). For instance, cancer can appear as "systemic malignancy," "metastatic disease" and "secondary cancer," etc. This issue is a major bottleneck, since current approaches use only the nearest neighbor similarity of concept embeddings to rank candidates (Liu *et al.*, 2021a). Additionally, as EL often involves the task of measuring the similarity of input embeddings with often millions of other embeddings in a search space, and then retrieving the closest n candidates, improving the efficiency of time and memory usage remains a crucial topic (Ngo *et al.*, 2021).

Further challenges are observed when dealing with abbreviations and acronyms (Agrawal *et al.*, 2022), as well as multilingual concepts (Remy *et al.*, 2022). In Table 2, we observe that concepts in English are easily matched, regardless of British or American spelling conventions. However, this is not the case for similar words across languages, as the largest concept ontology (UMLS) contains 69.6% of its concepts in English and 10.7% of them in Spanish, while other languages account for a mere 2.9% or lower (Liu *et al.*, 2021b).

Several recent studies achieve strong EL performance without cross-lingual fine-tuning on language-specific corpora. For instance, Wajsbürt et al. (Wajsbürt, 2021) report a +20 F1 increase on the Quaero corpus using only French training data. Cross-lingual benchmarks for clinical entity linking (Alekseev *et al.*, 2022) demonstrated that targeted, language-specific benchmarks can substantially boost zero-shot performance across English, Dutch, and French. Earlier, the LREC-MultilingualBio workshop (Roller *et al.*, 2018) explored clinical text translation via medical ontologies and sentence templates, highlighting the value of in-language resources.

3 Methodology

In recent years, we observe an increase in the performance of medical LLMs along with an increase in the amount of data used for model pretraining and the size of medical language models, from the encoder-only BioBERT (Lee *et al.*, 2020) at 175M parameters to the autoregressive BioGPT (Luo *et al.*, 2022) at 1.5B parameters, and 540B parameters of the MedPalm model (Singhal *et al.*, 2023). Oftentimes, state-of-the-art models are not publicly available, and hospitals in Europe are reluctant to partner and share data with external providers, partly due to ethical and regulatory (Ong *et al.*, 2024), as well as financial considerations (Dubas-Jakóbczyk *et al.*, 2024).

Our goal was therefore to create a small, scalable pipeline that does not require high-end GPUs for

on-prem deployment. While we cannot reveal all the details involving the training of the proprietary pipeline, we can share the steps undertaken for the creation of each component.

3.1 Clinical Data and Annotations

The lack of clear guidelines and consistent annotations can severely harm the performance of LLMs (Sylolypavan *et al.*, 2023). So, to extract high-quality information from clinical text, we developed our proprietary annotation scheme. Our detailed annotation guidelines and the raw Dutch EHR sentences remain confidential to comply with hospital agreements. To support reproducibility, we provide a high-level overview here and share synthetic examples in the appendix. It encompasses over two dozen NER labels, including various descriptive modifiers and attributes, such as negations, being separate NER labels. We equally allow for overlapping (nested) annotations to obtain the most detailed codified concepts. We also take the potential relationships between the entities into account. For instance, in the example below we annotate "dependent edema" and "edema" as Disorders (D), and "left leg" and "leg" as anatomical parts (A) to then further obtain `C5230911` [5]— "Edema of left lower limb" by combining "edema" with "left leg," and `C0577685` [6]—"Gravitational edema of leg" by combining "dependent edema" with "leg."

$$\left\lceil \text{Dependent} \left\lceil \text{edema} \right\rceil^{\text{D}} \right\rceil^{\text{D}} \text{of the} \left\lceil \text{left} \left\lceil \text{leg} \right\rceil^{\text{A}} \right\rceil^{\text{A}}$$

The data used for fine-tuning and validation consists of ca. 350 human-annotated patient records from various clinical subdomains. Upon segmentation and deduplication, the training set accounts for 10,341 annotated sentences and 112,881 concepts. Of these, 64,595 are unigrams, 28,250 are bigrams, and 20,036 concepts consist of three or more words. It is important to note that these concepts are not unique. The high number of unigrams results from the model learning to predict nested and discontinuous concepts, often using single-word entities like medication units or simple terms representing multiple concepts (e.g., "no" in Table 2). However, such entities are not forwarded to the entity linker, as they lack clinical relevance and are unlikely to be linked to a CUI.

3.2 Custom Discontinuous NER (C-DNER)

Building on the Clinical (Discontinuous) NER methods surveyed in Section 2.1, we develop a Custom Discontinuous NER (C-DNER) component : an encoder–decoder model (250 M parameters) pretrained on generic multilingual text and then fine-tuned on our in-domain Dutch clinical sentences. Upon fine-tuning, C-DNER returns a list of flat, nested, and discontinuous entity spans in a structured format (e.g., JSON), rather than token-level BIO tags. The model size was determined with the consideration for future on-prem deployment, as we strove for a lightweight end-to-end concept normalization pipeline.

We use transfer learning (using a multilingual base model for task-specific fine-tuning) to enhance the multilingual potential of our pipeline. This base model that we used for DNER is an open-source multilingual model pretrained on generic data, and fine-tuned it on hospital data (see subsection 3.1) exclusively in Dutch. While our annotation effort and hospital partnerships have centered on

5. purl.bioontology.org/ontology/SNOMEDCT/816182002
6. purl.bioontology.org/ontology/SNOMEDCT/300981003

	EHR			Mantra			Quaero			E3C		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Mistral + SapBERT	**78.78**	59.11	67.54	60.21	32.21	41.97	51.59	21.00	29.85	52.9	50.00	51.41
Mistral + C-EL	78.35	62.14	69.31	66.49	35.01	45.87	54.69	23.44	32.86	54.92	48.62	51.58
C-DNER + SapBERT	77.66	64.22	70.30	**73.08**	37.54	**49.60**	49.13	22.47	32.63	64.5	58.99	61.62
C-DNER + C-EL	78.31	**70.13**	**73.99**	68.06	36.69	47.68	52.46	24.91	33.78	**64.66**	**59.22**	**61.82**
QuickUMLS	64.94	68.05	66.46	53.69	**44.54**	48.69	**57.46**	**33.36**	**42.61**	47.77	55.76	51.46

TABLE 3 – Comparative results of end-to-end systems on 4 datasets. EHR : Electronic Health Records (hospital) data in Dutch. Other open-source datasets are Mantra (in Dutch), Quaero (in French), and E3C (in English). F1 is arguably the most important score as it integrates both Precision (P) and Recall (R) in one metric.

Dutch EHRs, we are extending our pipeline to additional languages as new hospital datasets in those languages become available. We must also mention that this model does not treat attribute extraction as a separate task but extracts flat, discontinuous and nested concepts that often include the attribute keyword (e.g., "presumably" or "potentially") inside the larger concept span.

In the next sections we refer to this component as Custom Discontinuous NER (C-DNER).

3.3 Clinical Entity Linking

Once we have used C-DNER to detect the relevant clinical entities, these concept become the input for our custom Entity Linker (EL). This is an encoder-only model with ca. 278M parameters, fine-tuned on a dataset that contains roughly 1.5M entries, of which ca. 30% in English (RxNorm [7] extension), and 70% in Dutch (UMLS 2023, and in-house annotations).

To select the top linking candidate for each given concept (named entity), we use an in-house multilingual knowledge base containing around 15M references which originate from public UMLS release, extended with additional vocabularies and translations of SNOMED, and RxNorm. We retrieve the 10 most similar concepts based on cosine similarity, and then re-sort the generated candidates based on predefined parameters stored in UMLS for each concept like the rank of vocabulary, the semantic type and the term preference.

In the next sections we refer to this component as Custom Entity Linker (C-EL).

3.4 Reproducibility Statement

While proprietary patient data cannot be publicly released for privacy reasons, we summarize here the key finetuning resources : (1) C-DNER finetuning : 10,341 Dutch clinical sentences from 350 de-identified records. (2) C-EL finetuning : 1.5 M concept–CUI pairs (30 % English RxNorm-extended ; 70 % Dutch UMLS-extended). (3) Notebook and synthetic examples are provided in the appendix.

7. www.nlm.nih.gov/research/umls/rxnorm/index.html

4 Results

The performance of the proprietary pipeline is compared against two generic, open-source end-to-end systems. As our baseline, we use QuickUMLS (Soldaini & Goharian, 2016) that operates by the principle of approximate string matching of the concepts from the UMLS ontology onto the input data. Due to the aforementioned challenges in the EHR data, we hypothesize that this baseline will result in lower scores on clinical narratives.

The second pipeline consists of two open-source models for the (D)NER and EL tasks. Out of the available multilingual LLMs capable of processing medical data in a single-GPU setup, we selected the 4-bit quantized version of the Mistral 7B Instruct model (Jiang *et al.*, 2023) for the (D)NER task. We found that even though the model was pretrained solely on English data, it could extract the necessary information from text in other languages if the prompt itself was composed in English. Our choice was based on an empirical evaluation of the model's biomedical counterpart BioMistral-7B (Labrak *et al.*, 2024), as we found the latter to be more prone to hallucinating in experiments, despite it being pretrained in multiple languages, among which Dutch and French.

Mistral 7B Instruct also exhibited output inconsistency. To achieve the highest-recall scenario, we repeated the prompt 10 times, which is notably impractical for production use. For the EL task, we used the multilingual SapBERT trained with the 2020AB release of the UMLS knowledge base [8].

To obtain objective results, we further performed an ablation study by including the performance of two other pipelines : C-DNER with SapBERT, and Mistral with C-EL. We emphasize that the open-source models were not fine-tuned on EHR data to underline the importance of in-domain data, which renders smaller models competitive for real-world applications. We also do not fine-tune any of the selected models on the open datasets selected for these experiments to objectively evaluate the out-of-the-box (OOTB) performance on multilingual biomedical text. With this paper, we include a notebook in Appendix A that showcases the working of an open-source pipeline.

Further, we selected four datasets in three languages : English, Dutch, and French. First, the EHR dataset in Dutch consisted of 35 patient records in total, of which 24 and 11 belonged to the oncology and cardiology domains, respectively. These records were then split into sentences with 1,194 to-be-normalized concepts in total. We translated [9] some segments presented in Appendix A from Dutch to English to showcase a sample. As we hypothesized that models fine-tuned on EHR data would be capable of decent performance on any biomedical data, the second dataset we selected was the Dutch (test) subset of the Mantra GSC corpus (Kors *et al.*, 2015), annotated for both NER and EL tasks. It consists of sentences harvested from the European Medicines Agency (EMA) documents that contain information on marketed medications, with 362 concepts in total. Similarly, the EMA subset of the Quaero medical corpus (Névéol *et al.*, 2014) lists 1,970 concepts in French. The English test subset of the E3C corpus (Magnini *et al.*, 2021) consists of 2,389 concepts from a number of clinical records (which are semantically the closest to our EHR test set in Dutch). All corpora include UMLS CUI codes to evaluate the EL task. We selected one public corpus per language (Mantra – Dutch, Quaero – French, E3C – English) alongside our private Dutch EHR set to evaluate robustness across both linguistic and genre shifts. Limiting each language to a single benchmark avoids overlap between datasets and allows us to measure how well models trained in one clinical setting generalize to others.

For (D)NER, we did not assess the model's label assignment performance for two reasons : (a)

8. huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR
9. Certain medical concepts, particularly abbreviations and acronyms, might have been mistranslated by DeepL

annotation schemes differ across datasets; (b) our primary focus is on the quality of the identified named entities for code assignment, not their NER labels. The metrics we used for the evaluation of the end-to-end concept normalization systems are precision, recall and F1. We do not include any relaxed metrics for this experiment, as we aim to evaluate the (multilingual) OOTB performance of LLMs on EHR data, rather than their potential. The evaluation scores for all models and datasets are presented in Table 3.

5 Discussion

We observe that while all pipelines are capable of processing the data in three languages to various degrees, our custom pipeline outperforms the rest on clinical narratives, namely, the EHR and E3C datasets. This is a particularly interesting observation, since the E3C test subset is in English, while our model was fine-tuned only on Dutch data, which highlights the importance of (sub)domain-tailored solutions. The performance of the pipelines that combine generic and custom LLMs equally demonstrate the importance of each component in the pipeline. For instance, the highest-scoring pipeline on the Mantra dataset combines a custom DNER model with the SapBERT model, outperforming the Mistral + SapBERT pipeline's F1 by nearly 8 percentage points. We observe a similar performance on the E3C test set.

With regard to the non-EHR test sets, we equally observe the strong performance of custom components. However, we emphasize that the annotations of the Quaero and Mantra datasets were tailored to the default ranking of the UMLS thesaurus (Névéol *et al.*, 2014; Kors *et al.*, 2015). Therefore, this explains the strong performance of the QuickUMLS pipeline, particularly on the Quaero dataset, as opposed to the custom entity linker that was fine-tuned according to the guidelines that specified the preferred UMLS terms among the many candidate concepts.

The difference in custom annotation and ranking standards is equally reflected in the EHR dataset. For instance, we observe that Mistral + SapBERT outperform the custom pipeline by a small margin for the precision metric. We attribute this to Mistral's tendency to generate simple, flat concepts (predominantly unigrams) that are further easily mapped to reference codes, while our custom DNER model produces a hierarchical tree of nested concepts.

This observation is supported by the precision scores for the Quaero test set (74% of which consists of unigrams), as the DNER + SapBERT pipeline scores lower than Mistral + SapBERT. Additionally, we must note that these scores are calculated only for a subset of entities that are currently listed in medical ontologies. Other non-codified entities (e.g., those in languages other than English, full medication prescriptions and (discontinuous) measurements) are absent from the experiments, as there are currently no open benchmarks.

Finally, we underline that these results are often solely an indication of performance, rather than a representative evaluation thereof. Clinical model (pre-)training can essentially never be complete due to the shifts in the data, such as new treatment trends. But, most importantly, general medical annotations might not reflect the desired output of normalization systems, as expected in applied subdomains. Oftentimes, the final evaluation scores of clinical concept normalization pipelines change as per medical experts' validation, as they filter or adjust hierarchically the reference knowledge bases (ontologies), and adjust the ranking of preferred terms. We identify additional challenges related to the clinical normalization task in the following section.

6 Challenges

6.1 Discontinuous NER

Human annotations can help us evaluate discontinuous NER internally, but there is a notable lack of open benchmarks, specifically for the biomedical/clinical (sub)domain(s). Moreover, clinical ontologies are incomplete, and often lack codes for longer, more detailed concepts.

For instance, not all negated concepts can be standardized, unlike "no infection" and "no itching" in Table 2. That is why we highlight the importance of equally detecting nested (overlapping) concepts, for instance, "pitting oedema right," "pitting oedema," and "oedema." If the longest span ("pitting oedema right") is absent from the ontology of choice, the next concept in the hierarchical structure ("pitting oedema") becomes the next linking candidate.

6.2 Clinical Entity Linking

Evaluation strategies remain a limitation for clinical EL. In most research, for training and evaluation, entity linkers use either identical or similar datasets, albeit with different degrees of formality (Zhang *et al.*, 2022).

The lack of subdomain-specific (e.g., oncology, cardiology, etc.) benchmarks and fine-tuning data is particularly noticeable when trying to navigate synonymous concepts. For instance, SNOMED CT provides two UMLS concept identifiers for the concept "metastatic neoplasm" — `C2939420` [10] and `C0013930` [11], the latter referring to the concept "tumor embolus." However, the latter entry does not list `C2939420` as its potential synonym, which raises the question whether it is a context or ontology-related phenomenon.

6.3 Multilingualism

The lack of cross-lingual data (training data and evaluation benchmarks) is particularly noticeable in the clinical domain. Although transfer learning can offer multilingual performance, the generalization capabilities of LLMs usually increase with the similarities of languages (Pires *et al.*, 2019).

Adding machine-translated (MT) data can enable multilingual capability, preserving original annotations and ensuring accurate translation of medical jargon is challenging. For instance, when we translate [12] the term "longmetastasen" (a compound noun) from Dutch to English, we observe that the preferred translation is "lung metastases" (a compound noun) rather than "pulmonary metastases" (an adjective-noun phrase) as generated within surrounding context "Geen long- en niermetastasen" ("No pulmonary or renal metastases"). "Lung metastasis" is then potentially linked to the code `C0153676` [13], the preferred name of which is "Metastatic malignant neoplasm to lung." While we do not expect generic MT systems to list "neoplasm" as a synonym to "metastasis," we must remark that clinical ontologies do. [14]

10. purl.bioontology.org/ontology/SNOMEDCT/14799000
11. purl.bioontology.org/ontology/SNOMEDCT/252986008
12. www.deepl.com/en/translator
13. purl.bioontology.org/ontology/SNOMEDCT/94391008
14. Not the term "neoplasm" on its own but with its descriptive modifiers "metastatic," "malignant."

7 Conclusion

In this paper, we showed that bespoke in-domain solutions are superior to generic (biomedical) LLMs in their application to real-world data, and specified a number of challenges that hospitals can struggle with when crafting (multilingual) pipelines specifically applied to the concept normalization task.

We underline the importance of smaller, scalable models, often tailored to certain subdomains (e.g., oncology, cardiology, psychiatry, etc.) and hope that the effort on multilingual alignment of biomedical datasets will continue to democratize research in healthcare worldwide.

Références

AGRAWAL M., HEGSELMANN S., LANG H., KIM Y. & SONTAG D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 1998–2022.

ALEKSEEV A., MIFTAHUTDINOV Z., TUTUBALINA E., SHELMANOV A., IVANOV V., KOKH V., NESTEROV A., AVETISIAN M., CHERTOK A. & NIKOLENKO S. (2022). Medical crossing : a cross-lingual evaluation of clinical entity linking. In *Proceedings of the thirteenth language resources and evaluation conference*, p. 4212–4220.

ANKIT PAL M. S. (2024). Openbiollms : Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics.

DUBAS-JAKÓBCZYK K., NDAYISHIMIYE C., SZETELA P. & SOWADA C. (2024). Hospitals' financial performance across european countries : a scoping review protocol. *BMJ open*, **14**(1), e077880.

GALLIFANT J., CHEN S., MOREIRA P., MUNCH N., GAO M., POND J., CELI L. A., AERTS H., HARTVIGSEN T. & BITTERMAN D. (2024). Language models are surprisingly fragile to drug names in biomedical benchmarks. *arXiv preprint arXiv :2406.12066*.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv e-prints*, p. arXiv–1904.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv e-prints*, p. arXiv–2310.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). Pubmedqa : A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577.

KORS J. A., CLEMATIDE S., AKHONDI S. A., VAN MULLIGEN E. M. & REBHOLZ-SCHUHMANN D. (2015). A multilingual gold-standard corpus for biomedical concept recognition : the mantra gsc. *Journal of the American Medical Informatics Association*, **22**(5), 948–956.

KOŠPRDIĆ M., PRODANOVIĆ N., LJAJIĆ A., BAŠARAGIN B. & MILOŠEVIĆ N. (2023). From zero to hero : Harnessing transformers for biomedical named entity recognition in zero-and few-shot contexts. *arXiv e-prints*, p. arXiv–2305.

KRALJEVIC Z., SHEK A., BEAN D., BENDAYAN R., TEO J. & DOBSON R. (2021). Medgpt : Medical concept prediction from clinical narratives. *arXiv e-prints*, p. arXiv–2107.

KRISHNAMOORTHY S., SINGH A. & TAFRESHI S. (2024). Llm-based section identifiers excel on open source but stumble in real world applications. *arXiv e-prints*, p. arXiv–2404.

LABRAK Y., BAZOGE A., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). Biomistral : A collection of open-source pretrained large language models for medical domains. *arXiv e-prints*, p. arXiv–2402.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LEHMAN E. & JOHNSON A. (2023). Clinical-t5 : Large language models built using mimic clinical text. *PhysioNet*.

LI F., LIN Z., ZHANG M. & JI D. (2021). A span-based model for joint overlapped and discontinuous named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 4814–4828.

LIU F., SHAREGHI E., MENG Z., BASALDELLA M. & COLLIER N. (2021a). 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* : Association for Computational Linguistics.

LIU F., VULIĆ I., KORHONEN A. & COLLIER N. (2021b). Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 565–574.

LIU J., JI D., LI J., XIE D., TENG C., ZHAO L. & LI F. (2022). Toe : A grid-tagging discontinuous ner model enhanced by embedding tag/word relations and more fine-grained tags. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**, 177–187.

LSAEW J. & POLLARD T. (2016). Data descriptor : Mimiciii, a freely accessible critical care database. *Thromb Haemost*, **76**(2), 258–262.

LUO R., SUN L., XIA Y., QIN T., ZHANG S., POON H. & LIU T.-Y. (2022). Biogpt : generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, **23**(6), bbac409.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2021). The e3c project : European clinical case corpus. *Language*, **1**(L2), L3.

MCINERNEY D., YOUNG G., VAN DE MEENT J.-W. & WALLACE B. (2023). CHiLL : Zero-shot custom interpretable feature extraction from clinical notes with large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 8477–8494, Singapore : Association for Computational Linguistics.

NASSIRI K. & AKHLOUFI M. A. (2024). Recent advances in large language models for healthcare. *BioMedInformatics*, **4**(2), 1097–1143.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

NGO D.-H., KEMP M., TRURAN D., KOOPMAN B. & METKE-JIMENEZ A. (2021). Semantic search for large scale clinical ontologies. In *AMIA Annual Symposium Proceedings*, volume 2021, p. 910 : American Medical Informatics Association.

ONG J. C. L., CHANG S. Y.-H., WILLIAM W., BUTTE A. J., SHAH N. H., CHEW L. S. T., LIU N., DOSHI-VELEZ F., LU W., SAVULESCU J. *et al.* (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, **6**(6), e428–e432.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual bert ? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4996–5001.

REMY F., DE JAEGER P. & DEMUYNCK K. (2022). Taming large lexicons : translating clinical text using medical ontologies and sentence templates. In *EmP 2022 : the Engineers meet Practicians conference*.

ROLLER R., KITTNER M., WEISSENBORN D. & LESER U. (2018). Cross-lingual candidate search for biomedical concept normalization. *MultilingualBIO : Multilingual Biomedical Text Processing*, p.1̃6.

SINGHAL K., AZIZI S., TU T., MAHDAVI S. S., WEI J., CHUNG H. W., SCALES N., TANWANI A., COLE-LEWIS H., PFOHL S. *et al.* (2023). Publisher correction : Large language models encode clinical knowledge. *Nature*, **620**(7973), 19–19.

SOLDAINI L. & GOHARIAN N. (2016). Quickumls : a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*, p. 1–4.

SUNG M., JEON H., LEE J. & KANG J. (2020). Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3641–3650.

SYLOLYPAVAN A., SLEEMAN D., WU H. & SIM M. (2023). The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, **6**(1), 26.

VAN ES B., RETEIG L. C., TAN S. C., SCHRAAGEN M., HEMKER M. M., ARENDS S. R., RIOS M. A. & HAITJEMA S. (2023). Negation detection in dutch clinical texts : an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, **24**(1), 10.

VAN VEEN D., VAN UDEN C., BLANKEMEIER L., DELBROUCK J.-B., AALI A., BLUETHGEN C., PAREEK A., POLACIN M., REIS E. P., SEEHOFNEROVÁ A. *et al.* (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, **30**(4), 1134–1142.

VERMA H., BERGLER S. & TAHAEI N. (2023). Comparing and combining some popular ner approaches on biomedical tasks. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

VRETINARIS A., LEI C., EFTHYMIOU V., QIN X. & ÖZCAN F. (2021). Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 international conference on management of data*, p. 2310–2318.

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. Thèse de doctorat, Sorbonne Université.

YAZDANI A., PROIOS D., ROUHIZADEH H. & TEODORO D. (2022). Efficient joint learning for clinical named entity recognition and relation extraction using fourier networks : A use case in adverse drug events. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, p. 212–223.

ZHANG S., CHENG H., VASHISHTH S., WONG C., XIAO J., LIU X., NAUMANN T., GAO J. & POON H. (2022). Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 868–880.

A Appendix

Anamnesis

Increase in abdominal circumference, ankle oedema and weight.

Aorta / A. Pulmonalis The aortic root and aortic ascendens are not dilated.

Discussion : above mentioned patient, known with DM2, hypertension, total AV block for which DDD pacemaker, subsequently for recovery left ventricular ejection fraction recovered after CRTP, three-vessel lesion for which 2-2023 stent LAD with residual lesion prox Cx, CTO RCA with collaterals in which on SPECT 11-2022 old collateral infarction.

Patient known to have CABG ('4), NSTEMI ('12) wc PCI , AL 1x DES + POBA RCx, AF (CHA2DS2-VASc 7), PHT (RVSP ca 50mmHg), DM2, HT, COPD G3 (home 1L O2), OSAS, TIA, CVA, gout, hyperthyreoidia wv strumazol, anaemia with intestinal angiodysplasia, cholecystectomy, obesity, panic attacks.

ECG on admission Sinus rhythm with 2 :1 AV block with ventricular frequency, 34/min, intermediate axis, Q in III. QRS 150ms aspecific widening.

No palpable resistances, black, bready discharge to glove.

IVM no reflow papaverine ic with recovery flow and recovery ST-T segments Angioseal for haemostasis after sheath, removal Vneous line inserted for inotropics.

Note : accepted for PCI proximal LAD 06-2017 : asthma cardiac obv 3VD 2014 : cataract 02-2019 : dec cordis o.b.v. deterioration LVF.

X-thorax on admission substantial corfigure with rounded sinus pleura on right and possibly some enhanced vascularity.

Notebook Biomedical LLMs Pretrained on Non-EHR Data Underperform in Multilingual Real-World Settings

EL inference: https://github.com/cambridgeltl/sapbert/blob/main/inference/inference_on_snomed.ipynb

[ ]:
```
!pip install ctransformers
```

[ ]:
```
from ctransformers import AutoModelForCausalLM
from transformers import AutoTokenizer, AutoModel
import ast
import numpy as np
import torch
from tqdm import tqdm
from scipy.spatial.distance import cdist
```

0.1 Named Entity Recognition

[ ]:
```
LLM = AutoModelForCausalLM.from_pretrained("TheBloke/Mistral-7B-Instruct-v0.
↪1-GGUF", model_file="mistral-7b-instruct-v0.1.Q4_K_M.gguf",␣
↪model_type="mistral", gpu_layers=50)
```

[ ]:
```
def generate(prompt, num_tries, llm) -> list:
    """
        Prompts LLM to extract medical concepts from clinical note.

        Args:
            prompt (str): Input prompt
            num_tries (int): number of tries
            llm (AutoModelForCausalLM): LLM

        Returns:
            response (list): list of medical concepts
    """
    response = []
    for _ in range(num_tries):
        output = llm(prompt)
        try:
            output = ast.literal_eval(output)
            if isinstance(output, dict) \
```

```
                and len(output) == 1 \
                and 'concepts' in output \
                and isinstance(output['concepts'], list) and all(isinstance(x,␣
 ↪str) for x in output['concepts']):
                    if len(output['concepts']) > len(response):
                        response = output['concepts']
        except:
            continue
    return response
```

[ ]:
```
sample_note_fr = "Jambes : pas d'infection ni de gonflement. Oedème de Quincke␣
 ↪gauche > droite"

# In this prompt we do not specify that we want flat, nested and discontinuous␣
 ↪concepts as it confuses the model
prompt = f"""
List all the medical concepts of this text in a json with 'concepts' as key:

{sample_note_fr}

Answer
"""
```

[ ]:
```
entities = generate(prompt, 10, LLM)
print(entities)
```

```
['Jambes', 'infection', 'gonflement', 'Oedème de Quincke', 'gauche', 'droit']
```

1 Entity Linking

We use the multilingual SapBERT model to embed the extracted entities and measure the similarity between the entries in a reference ontology.

[ ]:
```
tokenizer = AutoTokenizer.from_pretrained("cambridgeltl/
 ↪SapBERT-UMLS-2020AB-all-lang-from-XLMR")
model = AutoModel.from_pretrained("cambridgeltl/
 ↪SapBERT-UMLS-2020AB-all-lang-from-XLMR")
```

[ ]:
```
# codes as per snomed ct to be verified at http://purl.bioontology.org/ontology/
 ↪SNOMEDCT/{code}

candidates = [('Edema of left lower limb', '816182002'),
 ('left', '7771000'),
 ('right', '24028007'),
 ('Angioedema', '41291007'),
 ('Quincke\'s edema', '41291007'),
 ('Angioneurotic edema', '41291007'),
```

```
 ('Infectious process', '441862004'),
 ('infection', '441862004'),
 ('Absence of signs and symptoms of infection', '397680002'),
 ('No infection', '397680002'),
 ('swelling', '442672001'),
 ('edema', '79654002')]
```

[ ]:
```
all_names = [p[0] for p in candidates]
all_ids = [p[1] for p in candidates]
```

[ ]:
```
bs = 1
all_reps = []
for i in tqdm(np.arange(0, len(all_names), bs)):
    toks = tokenizer.batch_encode_plus(all_names[i:i+bs],
                                       padding="max_length",
                                       max_length=25,
                                       truncation=True,
                                       return_tensors="pt")

    output = model(**toks)
    cls_rep = output[0][:,0,:]

    all_reps.append(cls_rep.cpu().detach().numpy())
all_reps_emb = np.concatenate(all_reps, axis=0)
```

[ ]:
```
def link(entity, model, tokenizer):
    """
      Links medical concepts to entries from a medical ontology.

      Args:
          entity (str): Clinical named entity extracted in the previous step
          model (AutoModel): EL
          tokenizer (AutoTokenizer): Tokenizer

      Returns:
          candidate (str): codified candidate term
    """
    query_toks = tokenizer.batch_encode_plus([entity],
                    padding="max_length",
                    max_length=25,
                    truncation=True,
                    return_tensors="pt")
    query_output = model(**query_toks)
    query_cls_rep = query_output[0][:,0,:]
    dist = cdist(query_cls_rep.cpu().detach().numpy(), all_reps_emb)
    nn_index = np.argmin(dist)
    candidate = candidates[nn_index]
```

```
    return candidate
```

[ ]:
```
for entity in entities:
    prediction = link(entity, model, tokenizer)
    print(f"{entity}: {prediction}")
```

```
Jambes: ('edema', '79654002')
infection: ('infection', '441862004')
gonflement: ('swelling', '442672001')
Oedème de Quincke: ("Quincke's edema", '41291007')
gauche: ('left', '7771000')
droit: ('right', '24028007')
```