Cross-Lingual Entity Linking Using GPT Models in Radiology Abstracts

Mariana Dias(✉) and Carla Teixeira Lopes

INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal
{up201606486,ctl}@fe.up.pt

Abstract. Entity linking is an important task in medical natural language processing (NLP) for converting unstructured text into structured data for clinical analysis and semantic interoperability. However, in lower-resource languages, this task is challenging due to the limited availability of domain-specific resources. This paper explores a translation-based cross-lingual entity linking approach using GPT models, GPT-3.5 and GPT-4o, for zero-shot machine translation and entity linking with in-context learning. We evaluate our approach using a Portuguese-English parallel dataset of radiology abstracts. Our results show that chunk-level machine translation outperforms sentence-level translation. Moreover, our translation-based approach to cross-lingual entity linking of UMLS concepts outperformed the multilingual encoder method baseline. However, the in-context learning entity linking approach did not outperform a translation-based approach with a dictionary-based entity linking method.

Keywords: Medical Entity Linking · Large Language Models · GPT

1 Introduction

Entity linking is an important task in medical natural language processing (NLP), especially in clinical settings where large volumes of unstructured text require analysis and interpretation. By linking entity mentions in these documents to standardized concepts in medical terminologies, we can transform unstructured textual documents into a structured format more suitable for clinical analysis and decision support and ensure semantic interoperability [16].

Most medical ontologies and vocabularies are primarily available in English, limiting their use in lower-resource languages for various NLP tasks, including entity linking. Current state-of-the-art approaches use transformer models for cross-lingual entity linking. These models leverage multilingual encoders to align entity mentions across languages, often requiring pre-training and fine-tuning on large-scale data.

Recent advances in Large Language Models (LLM), particularly generative pre-trained transformer (GPT) models, have unlocked new possibilities to solve NLP tasks that do not require in-domain or task-specific training [4]. To our knowledge, no prior research has explored cross-lingual entity linking using GPT models in a translation-based framework. This paper aims to analyze the potential of GPT models for cross-lingual entity linking through machine translation,

Dias, M., Lopes, C.T. (2025). Cross-Lingual Entity Linking Using GPT Models in Radiology Abstracts. In: Grabis, J., Vos, T.E.J., Escalona, M.J., Pastor, O. (eds) Research Challenges in Information Science. RCIS 2025. Lecture Notes in Business Information Processing, vol 548. Springer, Cham. https://doi.org/10.1007/978-3-031-92471-2_2

J. Grabis et al. (Eds.): RCIS 2025, LNBIP 548, pp. 20–37, 2025.
https://doi.org/10.1007/978-3-031-92471-2_2

and in-context learning for entity recognition, alignment with an ontology, and projection. In this context, alignment refers to associating entity mentions in the translated text with standardized entities from an ontology, and projection involves transferring the linked entities back to the original language. Our experiment focuses on a radiology dataset [3], linking entity mentions to a standardized radiology ontology, RadLex.

We aim to address the following research questions:

1. How does the granularity of prompt context impact GPT models' performance in machine translation of radiology-related data?
2. Do larger, more advanced GPT models achieve better results than smaller ones in the entity linking task?
3. Do GPT models outperform other approaches for entity linking?

2 Cross-Lingual Entity Linking

Given a textual document D in a source language L_S, the goal of the cross-lingual entity linking task is to identify entity mentions m_1, ..., m_n within D and link each m_i mention to an entity $E_j \in KB$, where KB is a knowledge base in the target language L_T containing a set of entities $\{E_1, ..., E_m\}$, like the Unified Medical Language System (UMLS). The UMLS metathesaurus [1] is a well-known biomedical knowledge base that integrates various vocabularies.

Most works formulate the entity linking task as a multi-class classification or ranking problem [18]. Recent approaches use transformer-based models to build dense entity representations and compute relevance scores for entity candidates. Botha et al. [2] developed a bi-encoder model with mention and entity encoders initialized from pre-trained multilingual BERT models. Their method embeds mention-entity pairs in a shared vector space to retrieve entity candidates. Their approach outperformed others on the TR2016$^{\text{hard}}$ dataset, including Upadhyay et al.'s [19] FastText-based method.

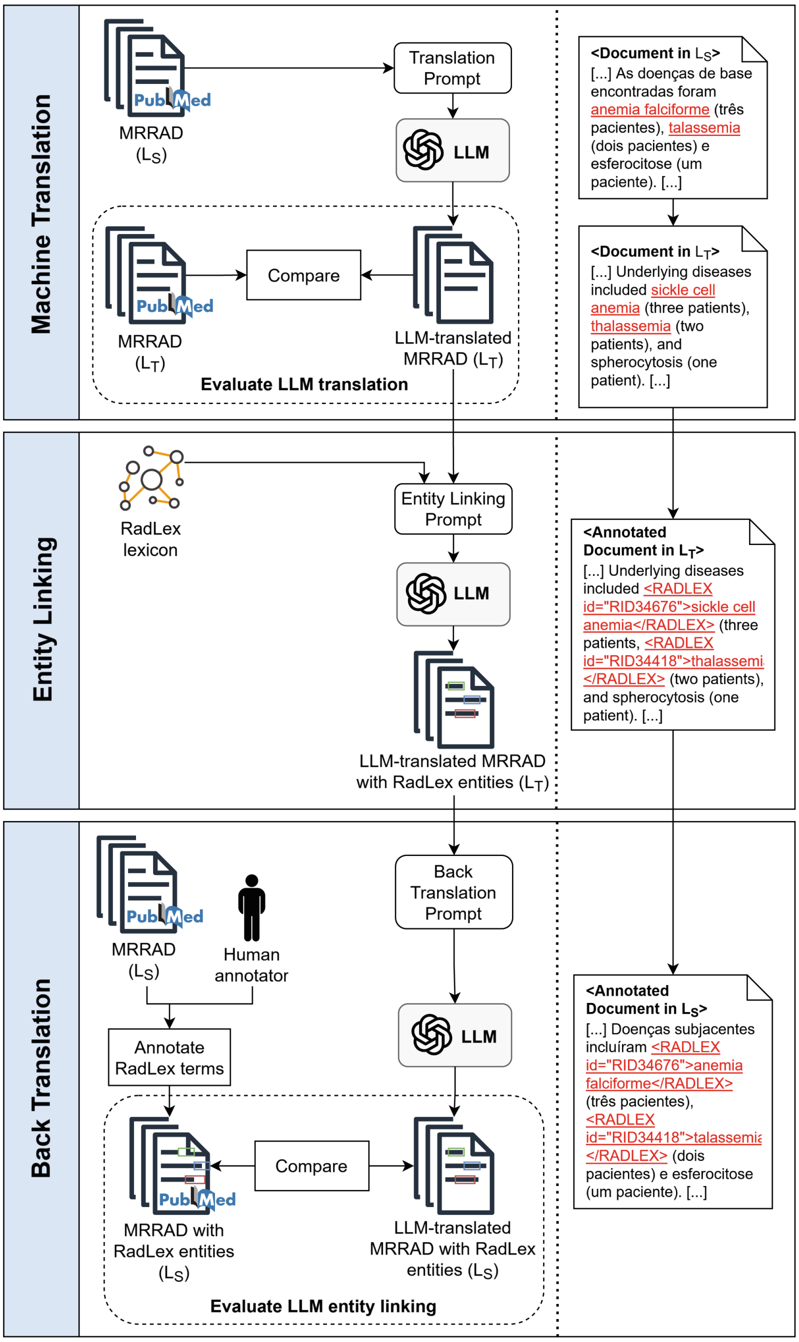
In the biomedical domain, Liu et al. [10] developed cross-lingual variations of SapBERT [9], a biomedical BERT-based model fine-tuned on UMLS synonyms. Their approach, leveraging multilingual encoders MBERT and XLMR, outperformed monolingual models on the cross-lingual biomedical entity linking benchmark (XL-BEL)[1] in lower-resource languages linguistically distant from Romance and Germanic languages.

Recent research has explored the use of GPT models for entity linking, particularly through in-context learning. Shlyk et al. [17] created a retrieval-augmented entity linking approach for biomedical concepts using in-context learning prompts. Groza et al. [6] evaluated GPT models for linking phenotype concepts through in-context learning. Both studies reported competitive performance on benchmark datasets. Other approaches, such as Ding et al.'s [5], leverage prompt engineering and instruction tuning to improve entity linking performance.

[1] https://paperswithcode.com/dataset/xl-bel.

3 Methodology

The cross-lingual entity linking pipeline consists of three phases: (1) translating a document D in a radiology dataset from the source language L_S (Portuguese) to the target language L_T (English), (2) recognizing entity mentions m_i and aligning them to terms E in the RadLex ontology, and (3) back-translating the annotated document D_a to L_S (Portuguese). Figure 1 provides an overview of the system's architecture with examples of outputs from each phase.



<Document in L_S>
[...] As doenças de base encontradas foram anemia falciforme (três pacientes), talassemia (dois pacientes) e esferocitose (um paciente). [...]

<Document in L_T>
[...] Underlying diseases included sickle cell anemia (three patients), thalassemia (two patients), and spherocytosis (one patient). [...]

<Annotated Document in L_T>
[...] Underlying diseases included <RADLEX id="RID34676">sickle cell anemia</RADLEX> (three patients, <RADLEX id="RID34418">thalassemi </RADLEX> (two patients), and spherocytosis (one patient). [...]

<Annotated Document in L_S>
[...] Doenças subjacentes incluíram <RADLEX id="RID34676">anemia falciforme</RADLEX> (três pacientes), <RADLEX id="RID34418">talassemi </RADLEX> (dois pacientes) e esferocitose (um paciente). [...]

Fig. 1. Architecture of an entity linking system for RadLex entities in the MRRAD dataset.

Our pipeline employs a domain-specific ontology and parallel corpus for evaluation. The knowledge base used is the RadLex[2] lexicon, developed by the Radiological Society of North America (RSNA). The RadLex ontology consists of 46,761 classes, of which 1,323 are linked to UMLS concepts. We evaluated our approach with the Multilingual Radiology Research Articles Dataset[3] (MRRAD) [3], a Portuguese-English parallel corpus that contains 34 PubMed abstracts related to radiology. Table 1 summarizes the dataset statistics.

Table 1. MRRAD dataset statistics: number of documents, average number of sentences per document, and average number of words per document.

Language	# Documents	Avg. # Sentences/Doc	Avg. # Words/Doc
Portuguese	34	123.6	2,947.2
English	34	151.7	2,908.4

Our goal with this study is to assess the feasibility of a three-stage LLM-based translation approach for cross-lingual entity linking. To achieve this, we compared the performance of two proprietary models from OpenAI, GPT-3.5 and GPT-4o, using zero-shot machine translation, in-context learning entity linking with pre-filtered ontology terms, and different prompting strategies. As a baseline, we included a system that combines GPT-based machine translation and back translation with dictionary lookup for entity linking and projection.

3.1 Machine Translation

For machine translation, we proposed two task-specific prompts with different granularities: sentence-level and word-chunk fitted to the LLM's context window. In both prompts, past queries and responses are retained to maintain context. The sentence-level prompt uses a full sentence as input, while the chunk-level prompt uses word chunks obtained by tokenizing the text with OpenAI's tiktoken[4] tokenizer and splitting it based on the LLM's context window as the threshold for maximum chunk size. Moreover, the chunk-level prompt uses a format that differentiates between the first and the subsequent chunks. We present the machine translation prompting approaches in Fig. 2.

[2] https://www.rsna.org/radlex/.

[3] https://github.com/lasigeBioTM/MRRAD.

[4] https://github.com/openai/tiktoken.

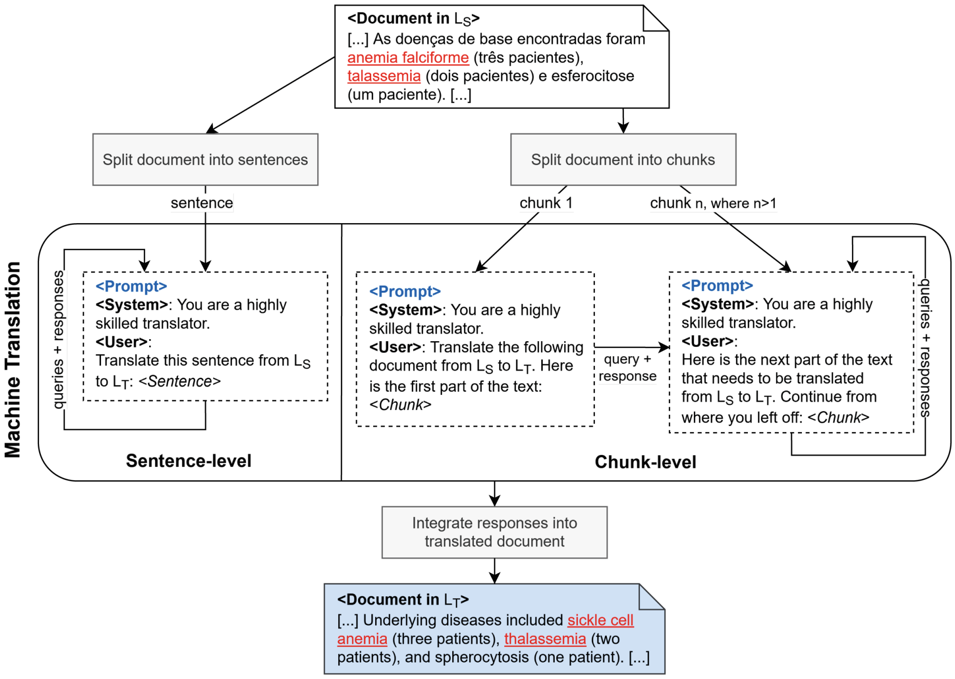

<Document in L_S>
[...] As doenças de base encontradas foram anemia falciforme (três pacientes), talassemia (dois pacientes) e esferocitose (um paciente). [...]
Split document into sentences
Split document into chunks
sentence
chunk 1
chunk n, where n>1
Machine Translation
queries + responses
<Prompt>
<System>: You are a highly skilled translator.
<User>:
Translate this sentence from L_S to L_T: <Sentence>
Sentence-level
<Prompt>
<System>: You are a highly skilled translator.
<User>: Translate the following document from L_S to L_T. Here is the first part of the text: <Chunk>
query + response
<Prompt>
<System>: You are a highly skilled translator.
<User>:
Here is the next part of the text that needs to be translated from L_S to L_T. Continue from where you left off: <Chunk>
queries + responses
Chunk-level
Integrate responses into translated document
<Document in L_T>
[...] Underlying diseases included sickle cell anemia (three patients), thalassemia (two patients), and spherocytosis (one patient). [...]

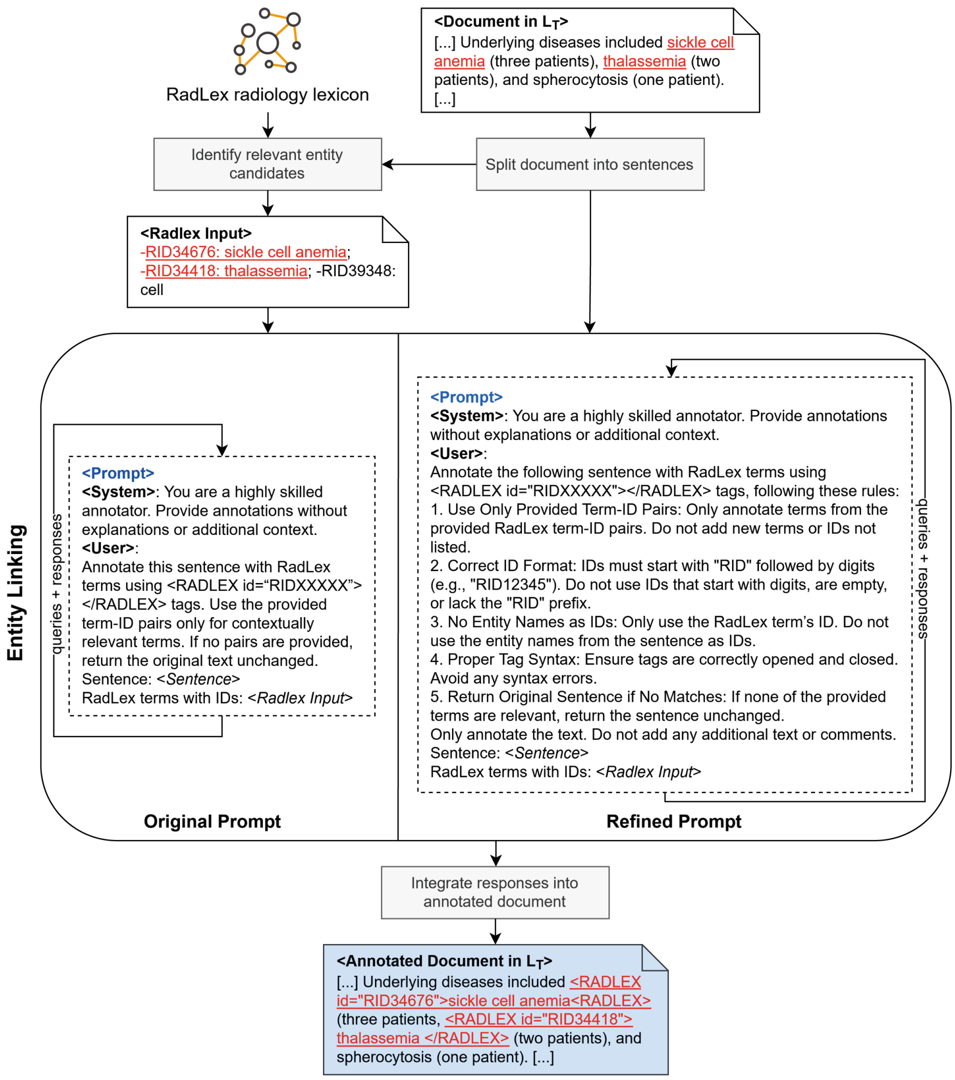
Fig. 2. Machine translation prompt experiment with example.

3.2 Entity Linking

Following the work of Hu et al. [7] of prompt engineering for clinical named entity recognition, we designed two prompting approaches for entity linking: an original prompt with a task description, format specification, and context, and a subsequently refined prompt. Figure 3 illustrates the entity linking prompting process, including the used prompts.

The original prompt follows a structured format that includes a task description, a format specification, and context to guide entity recognition and alignment with RadLex terms. We instruct models to use HTML tags to annotate entity mentions and their linked RadLex entities. The input consists of a sentence from a translated radiology abstract, supplemented with a list of relevant RadLex terms and their identifiers. We generate a list of candidate RadLex terms for each sentence using a dictionary lookup approach. We identify relevant RadLex terms and synonyms while filtering out shorter terms, retaining only those longer than three characters.

The subsequent refined prompt provides more detailed formatting instructions based on an analysis of the results from the initial prompt. It instructs the models to use valid tag syntax and identifiers by addressing common errors identified with the initial prompt. We formulated and refined five rules using ChatGPT: 1) use only provided term-id pairs to reduce hallucinations of non-existent RadLex terms or identifiers, 2) enforce identifier formatting, 3) prohibit entity names as identifiers, 4) ensure proper tag syntax to mitigate improperly closed tags, and 5) instruct the model to return the original sentence if no entities are recognized.


RadLex radiology lexicon
Identify relevant entity candidates
Split document into sentences
Entity Linking
queries + responses
queries + responses
Original Prompt
Refined Prompt
Integrate responses into annotated document

<Document in L_T>
[...] Underlying diseases included sickle cell anemia (three patients), thalassemia (two patients), and spherocytosis (one patient). [...]

<Radlex Input>
-RID34676: sickle cell anemia; -RID34418: thalassemia; -RID39348: cell

<Prompt>
<System>: You are a highly skilled annotator. Provide annotations without explanations or additional context.
<User>:
Annotate this sentence with RadLex terms using <RADLEX id="RIDXXXXX"></RADLEX> tags. Use the provided term-ID pairs only for contextually relevant terms. If no pairs are provided, return the original text unchanged.
Sentence: <*Sentence*>
RadLex terms with IDs: <*Radlex Input*>

<Prompt>
<System>: You are a highly skilled annotator. Provide annotations without explanations or additional context.
<User>:
Annotate the following sentence with RadLex terms using <RADLEX id="RIDXXXXX"></RADLEX> tags, following these rules:
1. Use Only Provided Term-ID Pairs: Only annotate terms from the provided RadLex term-ID pairs. Do not add new terms or IDs not listed.
2. Correct ID Format: IDs must start with "RID" followed by digits (e.g., "RID12345"). Do not use IDs that start with digits, are empty, or lack the "RID" prefix.
3. No Entity Names as IDs: Only use the RadLex term's ID. Do not use the entity names from the sentence as IDs.
4. Proper Tag Syntax: Ensure tags are correctly opened and closed. Avoid any syntax errors.
5. Return Original Sentence if No Matches: If none of the provided terms are relevant, return the sentence unchanged.
Only annotate the text. Do not add any additional text or comments.
Sentence: <*Sentence*>
RadLex terms with IDs: <*Radlex Input*>

<Annotated Document in L_T>
[...] Underlying diseases included <RADLEX id="RID34676">sickle cell anemia<RADLEX> (three patients, <RADLEX id="RID34418"> thalassemia </RADLEX> (two patients), and spherocytosis (one patient). [...]

Fig. 3. Entity linking prompting experiment with example.

To evaluate the entity linking phase, we use two approaches as baselines: a dictionary-based approach through the NCBO annotator and a multilingual encoder-based method. The NCBO annotator [8], developed by the National Center for Biomedical Ontology (NCBO), annotates biomedical documents by matching terms to a dictionary built from ontologies hosted in BioPortal[5]. We integrated this approach into our pipeline by replacing the GPT-based entity linking stage with the NCBO Annotator while maintaining the machine translation and back translation steps. We accessed the BioPortal REST API[6] through the Annotator endpoint with default parameter settings[7]. For the multilingual

[5] https://bioportal.bioontology.org/ontologies.
[6] https://data.bioontology.org/annotator.
[7] For more information, consult the documentation at https://data.bioontology.org/documentation.

encoder baseline, we used SapBERT-UMLS-2020AB-all-lang-from-XLMR[8] [10], a SapBERT model trained on UMLS, to generate dense embeddings to represent RadLex entities. We generated candidate entity mentions using an n-gram approach and performed entity linking by computing similarity scores between mention embeddings and RadLex entity embeddings. We linked mentions to RadLex terms when the similarity score exceeded a threshold of 0.9. Figure 4 demonstrates the pipeline for the baseline approaches.

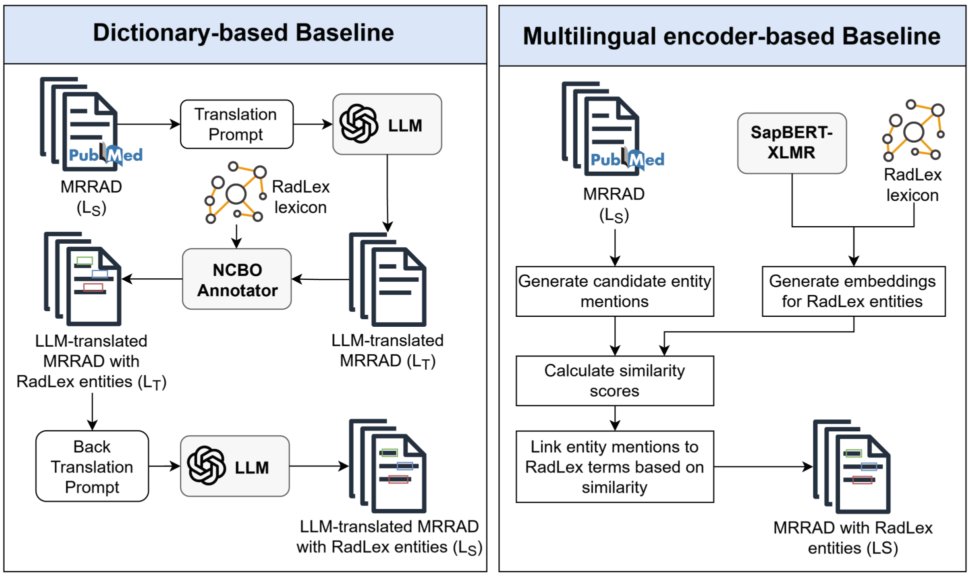

Dictionary-based Baseline
MRRAD (L_S)
Translation Prompt
LLM
RadLex lexicon
NCBO Annotator
LLM-translated MRRAD with RadLex entities (L_T)
LLM-translated MRRAD (L_T)
Back Translation Prompt
LLM
LLM-translated MRRAD with RadLex entities (L_S)
Multilingual encoder-based Baseline
MRRAD (L_S)
SapBERT-XLMR
RadLex lexicon
Generate candidate entity mentions
Generate embeddings for RadLex entities
Calculate similarity scores
Link entity mentions to RadLex terms based on similarity
MRRAD with RadLex entities (LS)

Fig. 4. Cross-lingual entity linking baseline pipeline.

3.3 Back Translation

We performed back translation on the annotated text containing RadLex entities, projecting the tags from the target language L_T to the source language L_S through the translation process. To maintain the integrity of the HTML-like tags that identify entity mentions, we did not use a chunk-based approach, as used in the machine translation phase, to prevent cutting off entities. Instead, we used a sentence-level back translation prompt with an additional instruction to preserve HTML tags in the output to ensure that the structure of the original text is maintained while incorporating the linked entities. Initially, we used the same prompt as for machine translation and refined it based on results from experiments on a few documents. For the final prompt, we consulted ChatGPT for suggestions on potential prompts that could decrease errors. Figure 5 shows the back translation prompting process.

[8] https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR.

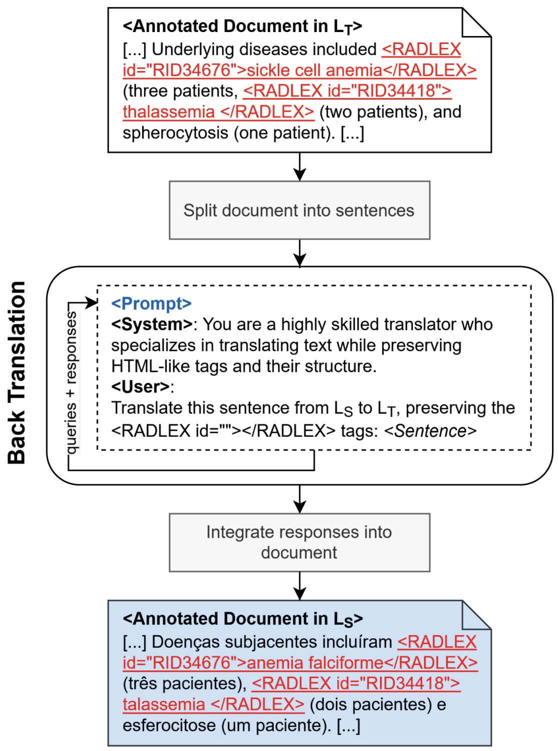

<Annotated Document in L_T>
[...] Underlying diseases included <RADLEX id="RID34676">sickle cell anemia</RADLEX> (three patients, <RADLEX id="RID34418"> thalassemia </RADLEX> (two patients), and spherocytosis (one patient). [...]
Split document into sentences
Back Translation
queries + responses
<Prompt>
<System>: You are a highly skilled translator who specializes in translating text while preserving HTML-like tags and their structure.
<User>:
Translate this sentence from L_S to L_T, preserving the <RADLEX id=""></RADLEX> tags: <Sentence>
Integrate responses into document
<Annotated Document in L_S>
[...] Doenças subjacentes incluíram <RADLEX id="RID34676">anemia falciforme</RADLEX> (três pacientes), <RADLEX id="RID34418"> talassemia </RADLEX> (dois pacientes) e esferocitose (um paciente). [...]

Fig. 5. Back translation prompting experiment with example.

To evaluate the performance of our approach in cross-lingual entity linking, we manually created a gold standard for the MRRAD dataset by annotating each document with RadLex entity mentions. We performed the annotation using the Protégé ontology editor with the Knowtator plugin[9]. During the annotation process, we focused on ontology classes that represent biomedical terms to ensure linking to relevant radiological concepts in diverse medical contexts. The result-

Table 2. Classes selected for gold standard annotation with total of descendent nested subclasses.

Class	Name	# Descendant Classes
RID3	Anatomical Entity	38,165
RID34785	Clinical Finding	2,230
RID5	Imaging Observation	1,133
RID50606	Imaging Specialty	86
RID7479	Non-anatomical Substance	392
RID34861	Object	403
RID1559	Procedure	610
RID39128	Process	35

[9] https://github.com/UCDenver-ccp/Knowtator-2.0.

ing gold standard dataset contains 4,327 linked entities, averaging 127 linked entities per document. Table 2 presents the selected classes and the number of nested subclasses for annotation.

4 Results

We divided the experiments into three phases: document-level machine translation quality evaluation in Sect. 4.1, entity linking error analysis in Sect. 4.2, and cross-lingual entity linking evaluation in Sect. 4.3.

4.1 Document-Level Machine Translation

We report the results of executing the machine translation prompts described in Sect. 3.1 in Table 3. To evaluate the translation quality, we used measures that assess lexical precision, BLEU [11] and ChrF++ [12] with SacreBLEU[10] [13], and neural metrics that evaluate semantic accuracy, COMETkiwi [15] (wmt22-COMETkiwi-da[11]) and COMET-22 [14] (wmt22-COMET-da[12]).

Table 3. Machine translation performance of GPT models on MRRAD dataset.

System	BLEU	ChrF++	COMETkiwi	COMET-22
Prompt S				
GPT-3.5	50.85	**88.29**	58.73	88.11
GPT-4o	**52.27**	81.06	60.86	88.42
Prompt C				
GPT-3.5	36.50	66.40	61.68	88.48
GPT-4o	33.81	64.43	**62.34**	**88.50**

The chunk-level prompt (Prompt C) performs better than the sentence-level prompt (Prompt S) with neural-based COMET measures. However, it performs worse using lexical-based measures like BLEU and ChrF++. We performed a qualitative analysis of machine translation outputs generated using different prompts and GPT models to better understand why the lexical-based measures declined in performance with GPT-4o, while the neural-based metrics improved. We present examples of translations that illustrate that lexical-based measures are likely more sensitive to exact word matches and less adaptable to variations in vocabulary and phrasing than neural-based metrics. Listing 1.1 demonstrates an example of machine translation outputs with the sentence-level prompt where the GPT-4o translation had more variations in vocabulary and phrasing.

[10] https://github.com/mjpost/sacrebleu.

[11] https://huggingface.co/Unbabel/wmt22-cometkiwi-da.

[12] https://huggingface.co/Unbabel/wmt22-comet-da.

Listing 1.1: Machine translation outputs using Prompt S.

Original sentence:
RESUMO OBJETIVO: Descrever a distribuição dos escores de cálcio coronariano numa população de homens brasileiros brancos assintomáticos submetidos à avaliação pela tomografia ultra-rápida.
Reference translation:
ABSTRACT OBJETIVE: To describe the distribution of coronary artery calcium scores in a population of asymptomatic white Brazilian men undergoing assessment with ultrafast computed tomography.
Translation with GPT-3.5:
SUMMARY OBJECTIVE: To describe the distribution of coronary calcium scores in a population of asymptomatic Brazilian white men undergoing evaluation by ultrafast computed tomography.
Translation with GPT-4o:
OBJECTIVE SUMMARY: Describe the distribution of coronary calcium scores in a population of asymptomatic white Brazilian men assessed using ultra-fast tomography.

Listing 1.2 shows an example of machine translation outputs with the chunk-level prompt where GPT-4o preserves the full citation of the PEPI program and maintains the original sentence flow, whereas GPT-3.5 omits the citation and splits the sentence into two.

Listing 1.2: Machine translation outputs using Prompt C.

Original sentence:
Os pacientes foram divididos randomicamente utilizando o programa PEPI {[COMPUTER PROGRAMS FOR EPIDEMIOLOGISTS (PEPI)] by J.H. Abramson and Paul M. Gahlinger. Version 4.04x}, em dois grupos com sorteio de envelope selado, as seringas contendo a droga eram preparadas por um pesquisador que não fosse avaliar o paciente.
Reference translation:
The patients were randomly divided into two groups using a raffle with sealed envelops and the computer program PEPI (Computer Programs for Epidemiologists by J.H. Abramson and Paul M. Gahlinger. Version 4.04x). The syringes containing the drug were prepared by a researcher that would not evaluate the patients.
Translation with GPT-3.5:
The patients were randomly divided into two groups using the PEPI program, with sealed envelope randomization. The syringes containing the drug were prepared by a researcher who did not evaluate the patient.
Translation with GPT-4o:
The patients were randomly divided using the PEPI program (COMPUTER PROGRAMS FOR EPIDEMIOLOGISTS (PEPI) by J.H. Abramson and Paul M. Gahlinger. Version 4.04x) into two groups with sealed envelope allocation, and the syringes containing the drug were prepared by a researcher who would not evaluate the patient.

Based on the previous analysis, we have decided to prioritize neural-based measures. Thus, we conducted statistical tests to evaluate the machine translation quality difference between the two prompts and the two GPT models using the COMET-based metrics. As the data was paired, we initially considered conducting paired t-tests. However, upon assessing normality and outliers assumptions, we found that the reference-based COMET-22 and reference-free COMETkiwi metrics did not meet normal distribution requirements. Therefore, we used the Wilcoxon signed-rank test as a non-parametric alternative to the paired t-test. We used this test to compare 1) the mean difference between the two prompts for each model, and 2) the difference between the two models using the same prompts. Table 4 presents the Wilcoxon signed-rank test results.

Table 4. p-values of Wilcoxon signed-rank test pairwise comparisons.

Comparison	COMETkiwi	COMET-22
GPT-3.5$_{PromptS}$ < GPT-3.5$_{PromptC}$	<.001	<.01
GPT-4o$_{PromptS}$ < GPT-4o$_{PromptC}$	<.01	.289
GPT-3.5$_{PromptS}$ < GPT-4o$_{PromptS}$	<.001	<.01
GPT-3.5$_{PromptC}$ < GPT-4o$_{PromptC}$	.163	.361

For both COMETkiwi and COMET-22, GPT-4o significantly outperforms GPT-3.5, indicating an advantage of GPT-4o over GPT-3.5 when using Prompt S. However, there are no significant differences for either metric with Prompt C. For the comparison between prompts with the same models, there are significant differences for both metrics with GPT-3.5. This suggests that, for the GPT 3.5 model, Prompt C produces higher COMETkiwi and COMET-22 scores than Prompt S. Since Prompt C demonstrates superior performance in machine translation, we will use the documents translated with this prompt for the entity linking task.

4.2 Entity Linking Error Analysis

To assess the performance of our prompting strategies in the entity linking task, we analyzed errors in the entity linking and back translation phases, focusing on hallucinations and their impact on entity linking performance. In this context, we consider hallucinations as invalid RadLex identifiers generated by the GPT models. To understand the nature of the hallucinations, we analyzed and categorized the misrepresented RadLex identifiers that caused them. We classified common linking errors into five types: missing, no prefix, numeric, invalid, and textual. Table 5 provides definitions and examples for each error type.

Table 5. Entity linking error typology with examples.

Error type	Description	Example
Missing	Entity mentions without a RadLex identifier	"Os registros médicos de 3.101 <RADLEX id=""></RADLEX> vítimas [...]"
No prefix	Numeric identifiers that are RadLex terms but lack the "RID" prefix	"[...] quantificação do <RADLEX id="11800">cálcio</RADLEX>"
Numeric	Numeric identifiers that are not RadLex terms	"Considerado significativo quando alpha <RADLEX id="12345">0,05</RADLEX>"
Textual	Entity mentions with textual identifier	"Esta <RADLEX id="disorder"> condição</RADLEX> é rara [...]"
Invalid	Identifiers that follows a valid format but are not RadLex terms	"[...] aspectos clínicos e radiográficos <RADLEX id="RID12940">correspondentes </RADLEX>"

We compared the performance of the dictionary-based baseline and GPT models using different entity linking prompts across both phases. Table 6 shows the frequency of RadLex entity mentions identified and associated hallucination rates.

Regarding the baseline, using the NCBO Annotator for entity linking and the GPT models for machine and back translation resulted in a higher hallucination rate with GPT-3.5, suggesting that GPT-4o is slightly more reliable in generating accurate RadLex identifiers. The GPT models exhibited lower hallucination rates in the entity linking stage compared to the back translation phase, with GPT-4o achieving the lowest hallucination rates near 0%. In the back translation stage, GPT-4o still maintained low hallucination rates of 1.22%-1.42%. The GPT-3.5 model identified more total and unique terms with the refined entity linking prompt than with the initial prompt and had a slightly lower hallucination rate. We also observed a reduction in hallucination rates with the refined prompt compared to the original prompt in all observations.

We analyzed the distribution of the classified RadLex identifier errors across different approaches, including the dictionary-based baseline, and GPT models with different prompts in the entity linking and back translation phases, as illustrated in Fig. 6.

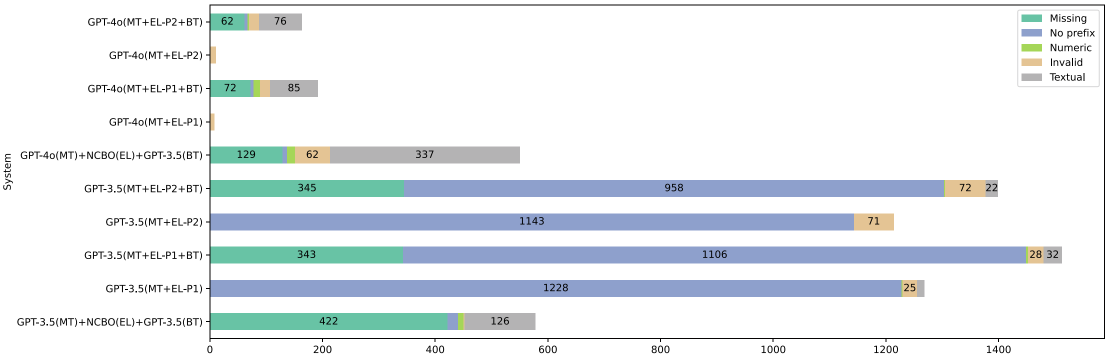


Fig. 6. Distribution of RadLex identifier errors.

Consistent with the earlier analysis, where entity linking approaches exhibited lower hallucination rates, the methods in the entity linking phase had significantly fewer errors, particularly missing and textual errors. This suggests that these error types were likely introduced during the back translation process.

The NCBO baseline with the GPT-4o model generally produced fewer missing errors compared to the baseline with GPT-3.5. However, it showed a higher occurrence of invalid and textual errors. The refined entity linking prompt seems to have greatly decreased the occurrence of no prefix errors in the two GPT models, although it led to an increase in invalid errors with the GPT-3.5 model. Overall, the GPT-4o model outperformed the GPT-3.5 model in minimizing missing, no prefix, and invalid errors but exhibits a slightly higher frequency of textual errors.

Table 6. Overview of frequency of RadLex terms identified across all experiments and hallucination rates.

System	**Total**	**Unique**	**Hallucination Rate (%)**
Ground truth	4,327	927	-
GPT-3.5$_{MT}$+NCBO$_{EL}$+GPT-3.5$_{BT}$	6,641	3,316	8.70
GPT-4o$_{MT}$+NCBO$_{EL}$+GPT-4o$_{BT}$	7,136	3,713	7.71
GPT-3.5$_{MT+EL-P1}$	13,824	5,161	9.17
GPT-3.5$_{MT+EL-P2}$	14,235	5,215	8.53
GPT-4o$_{MT+EL-P1}$	13,032	3,984	**0.06**
GPT-4o$_{MT+EL-P2}$	13,036	3,999	0.08
GPT-3.5$_{MT+EL-P1+BT}$	12,806	4,876	11.81
GPT-3.5$_{MT+EL-P2+BT}$	13,207	4,907	10.59
GPT-4o$_{MT+EL-P1+BT}$	13,523	4,105	1.42
GPT-4o$_{MT+EL-P2+BT}$	13,336	4,105	1.22

MT: machine translation, EL-P1: original entity linking prompt, EL-P2: refined entity linking prompt, BT: back translation.

We proceeded to perform a qualitative analysis of textual errors generated by the GPT-3.5 and GPT-4o models to understand if there is a correlation between the back translation prompt and the induction of hallucinations. In Listing 1.3, we provide an example of an output that demonstrates how all observed textual errors originated from sentences that did not contain RadLex entity mentions during the entity linking step, but were incorrectly annotated with the back translation prompt.

Listing 1.3: Entity linking output with textual error during back translation stage.

Original sentence:
Recentemente, desenvolvemos um sistema de visão computacional, o qual denominamos SIStema para a Detecção e a quantificação de Enfisema Pulmonar (SISDEP).
Machine Translation output:
We recently developed a computer vision system, named Pulmonary Emphysema Detection and Quantification System (SISDEP).
Entity Linking output:
We recently developed a computer vision system, named Pulmonary Emphysema Detection and Quantification System (SISDEP).
Back Translation output:
Desenvolvemos recentemente um sistema de visão computacional, chamado <RADLEX id="Pulmonary_Emphysema">Sistema de Detecção e Quantificação de Enfisema Pulmonar</RADLEX> (SISDEP).

4.3 Cross-Lingual Entity Linking Evaluation

To assess the effectiveness of our approach in biomedical cross-lingual entity linking, we focused on a subset of the RadLex ontology that contains standardized UMLS concepts. We used document-level precision (P), recall (R), and F1-score (F1) metrics, which are calculated based on the overlap between predicted entity mentions and their corresponding RadLex terms in each document. To evaluate the impact of hallucinations, we perform post-processing on the back translation step to filter out invalid RadLex entity links. Table 7 presents the document-level evaluation results, comparing the performance of our approach to the dictionary-based and multilingual encoder baselines.

The translation-based approach to entity linking outperformed the multilingual encoder-based SapBERT-XLMR model. Our approach performed the best with the dictionary-based NCBO annotator for entity linking compared to the in-context learning approach of the GPT models. For the dictionary-based baseline, the choice of GPT model had a minimal impact on performance.

Applying post-processing to filter out hallucinations enables a more accurate assessment of the approaches' performance, as any observed decrease in precision is more likely to reflect the approach's limitations in identifying relevant entities, rather than errors caused by hallucinated identifiers. The entity linking prompt strategies had no performance impact with the GPT-4o model. In contrast, for the GPT-3.5 model, the refined entity linking prompt resulted in a slight improvement in recall, with no effect on precision, suggesting that the refined prompt was more effective at identifying relevant UMLS concepts from the RadLex ontology.

Table 7. Document-level entity linking evaluation results with UMLS terms.

System	**P**	**R**	**F1**
SapBERT-XLMR$_{base}$	0.15	0.17	0.14
GPT-3.5$_{MT}$			
+ NCBO$_{EL}$+GPT-3.5$_{BT-NoPP}$	0.29	**0.78**	0.40
+ GPT-3.5$_{EL-P1+BT-NoPP}$	0.15	0.73	0.24
+ GPT-3.5$_{EL-P2+BT-NoPP}$	0.17	0.76	0.26
+ NCBO$_{EL}$+GPT-3.5$_{BT-WithPP}$	**0.38**	**0.78**	**0.49**
+ GPT-3.5$_{EL-P1+BT-WithPP}$	0.32	0.73	0.42
+ GPT-3.5$_{EL-P2+BT-WithPP}$	0.31	0.76	0.42
GPT-4o$_{MT}$			
+ NCBO$_{EL}$+GPT-4o$_{BT-NoPP}$	0.24	**0.78**	0.34
+ GPT-4o$_{EL-P1+BT-NoPP}$	0.28	0.76	0.38
+ GPT-4o$_{EL-P2+BT-NoPP}$	0.28	**0.78**	0.39
+ NCBO$_{EL}$+GPT-4o$_{BT-WithPP}$	0.37	0.78	0.48
+ GPT-4o$_{EL-P1+BT-WithPP}$	0.33	0.76	0.43
+ GPT-4o$_{EL-P2+BT-WithPP}$	0.33	0.78	0.44

MT: machine translation, BT: back translation, EL-P1: original entity linking prompt, EL-P2: refined entity linking prompt, NoPP: without post-processing, WithPP: with post-processing.

5 Discussion

The machine translation evaluation demonstrated that prompt choice had a significant impact on GPT-3.5's performance, whereas GPT-4o was less influenced by prompt variations. This finding addresses RQ1, confirming that the granularity of a prompt's context impacts model performance. We also concluded that GPT-4o significantly outperformed GPT-3.5, indicating that, in the machine translation phase, a larger model achieved the best results, addressing RQ2.

In the cross-lingual entity evaluation, we did not find any major differences in performance between GPT-3.5 and GPT-4o or between the different entity linking prompt strategies. This suggests that, relating to RQ2, increasing model size did not lead to improvements in entity linking performance unlike in machine translation.

Regarding our translation-based entity linking approach, the dictionary-based approach achieved comparable or superior F1-scores in comparison to the in-context learning method with GPT models. In response to RQ3, GPT models did not outperform the dictionary-based approaches for entity linking, indicating that GPT models did not significantly enhance the contextual recognition of relevant entities.

6 Conclusion

In this study, we explored the use of GPT models, specifically GPT-3.5 and GPT-4o, in the cross-lingual entity linking task using a translation-based approach that consists of three phases: machine translation, entity linking, and back translation. We explored different prompting strategies and entity linking approaches, including a dictionary-based method and in-context learning.

In the machine translation phase, our results showed that chunk-level machine translation outperformed sentence-level translation in the MRRAD dataset. During the entity linking phase, our error analysis revealed that the GPT-4o model had a near 0% hallucination rate. In the back translation phase, when evaluating cross-lingual entity linking with UMLS terms in the RadLex ontology, our approach outperformed the baseline multilingual encoder-based method. However, the in-context learning entity linking approach did not outperform the dictionary-based method.

Overall, our translation-based approach to cross-lingual entity linking shows potential as a viable method, but its effectiveness should be further evaluated on a wider range of datasets to assess its robustness. While post-processing helped mitigate hallucinations, it could not overcome the limitations of GPT models in accurately linking entities.

For future work, it would be interesting to explore other LLMs beyond GPT models and implement a knowledge retriever for ambiguous entities that could further enrich the prompt context and improve model performance.

Acknowledgments. This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021–2026, within project HfPT, with reference 41.

References

1. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32** (2004). https://doi.org/10.1093/nar/gkh061
2. Botha, J.A., Shan, Z., Gillick, D.: Entity linking in 100 languages. In: EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2020). https://doi.org/10.18653/v1/2020.emnlp-main.630
3. Campos, L., Pedro, V., Couto, F.: Impact of translation on named-entity recognition in radiology texts. Database: J. Biol. Databases Curation **2017** (2017). https://doi.org/10.1093/database/bax064
4. Ding, B., et al.: Is GPT-3 a good data annotator? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1 (2023). https://doi.org/10.18653/v1/2023.acl-long.626

5. Ding, Y., Poudel, A., Zeng, Q., Weninger, T., Veeramani, B., Bhattacharya, S.: EntGPT: linking generative large language models with knowledge bases (2024). https://doi.org/10.48550/arXiv.2402.06738
6. Groza, T., et al.: An evaluation of GPT models for phenotype concept recognition. BMC Med. Inform. Decis. Making **24** (2024). https://doi.org/10.1186/s12911-024-02439-w
7. Hu, Y., et al.: Improving large language models for clinical named entity recognition via prompt engineering. J. Am. Med. Inform. Assoc. **31** (2024). https://doi.org/10.1093/jamia/ocad259
8. Jonquet, C., Shah, N.H., Youn, C.H., Musen, M.A., Callendar, C., Storey, M.A.: NCBO annotator: semantic annotation of biomedical data. In: 8th International Semantic Web Conference, Poster and Demo Session (ISWC 2009), Washington, DC, USA (2009). https://hal.science/hal-04276274
9. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference (2021). https://doi.org/10.18653/v1/2021.naacl-main.334
10. Liu, F., Vulić, I., Korhonen, A., Collier, N.: Learning domain-specialised representations for cross-lingual biomedical entity linking. In: ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, vol. 2 (2021). https://doi.org/10.18653/v1/2021.acl-short.72
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 311–318. Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1073083.1073135
12. Popovic, M.: CHRF ++: words helping character n-grams. In: WMT 2017 - 2nd Conference on Machine Translation, Proceedings (2017). https://doi.org/10.18653/v1/w17-4770
13. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191. Association for Computational Linguistics, Brussels (2018). https://www.aclweb.org/anthology/W18-6319
14. Rei, R., et al.: COMET-22: unbabel-IST 2022 submission for the metrics shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 578–585. Association for Computational Linguistics, Abu Dhabi (2022). https://aclanthology.org/2022.wmt-1.52/
15. Rei, R., et al.: COMETKIWI: IST-unbabel 2022 submission for the quality estimation shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 634–645. Association for Computational Linguistics, Abu Dhabi (2022). https://aclanthology.org/2022.wmt-1.60/
16. Seinen, T.M., et al.: Use of unstructured text in prognostic clinical prediction models: a systematic review (2022). https://doi.org/10.1093/jamia/ocac058
17. Shlyk, D., Groza, T., Mesiti, M., Montanelli, S., Cavalleri, E.: REAL: a retrieval-augmented entity linking approach for biomedical concept recognition. In: Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., Tsujii, J. (eds.) Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pp. 380–389. Association for Computational Linguistics, Bangkok (2024). https://doi.org/10.18653/v1/2024.bionlp-1.29

18. Tsai, C.T., Upadhyay, S., Roth, D.: Multilingual Entity Linking. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-74901-8
19. Upadhyay, S., Gupta, N., Roth, D.: Joint multilingual supervision for cross-lingual entity linking. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018 (2018). https://doi.org/10.18653/v1/d18-1270