Overview of MedProcNER Task on Medical Procedure Detection and Entity Linking at BioASQ 2023

Automatic detection, normalization and indexing of clinical procedures in clinical texts shared task: resources, methods and evaluation

Salvador Lima-López[1,*], Eulàlia Farré-Maduell[1], Luis Gascó[1], Anastasios Nentidis[2,3], Anastasia Krithara[2,3], Georgios Katsimpras[2,3], Georgios Paliouras[2,3] and Martin Krallinger[1]

[1]*Barcelona Supercomputing Center, Spain*
[2]*National Center for Scientific Research "Demokritos", Athens, Greece*
[3]*Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract

Recent advances in NLP techniques, the use of large language models and Transformers are showing promising results for processing clinical content. The development of tools for automatic recognition of medical concepts, variables, and clinical expressions is key for the semantic analysis of clinical records, semantic search engines and the generation of structured data representations.

Despite the importance of medical procedures for management, diagnosis prevention and prognosis, there are few comprehensive resources for medical procedure extraction and normalization. In order to foster the development of procedure mention detection and entity linking systems, we have released the MedProcNER (Medical Procedures Name Entity Recognition) corpus, a high quality, manually annotated collection of 1000 clinical case reports written in Spanish. The corpus has been exhaustively labeled by physicians following detailed annotation guidelines and quality control measurements. Additionally, a multilingual Silver Standard corpus has also been generated for English, Italian, French, Portuguese, Romanian, Dutch, Swedish and Czech, to provide a clinical NLP resource for research in these languages. A total of 9 teams from 8 different countries have participated in the MedProcNER track of BioASQ 2023 (part of CLEF 2023), using mostly Transformers architectures and models like RoBERTA, BioMBERT, ALBERT, Longformers or SapBERT. MedProcNER was structured into three sub-tracks: a) Clinical Procedure Entity Recognition task, b) Clinical Procedure Normalization task to SNOMED CT and c) Clinical Procedure-based Document Indexing task. The MedProcNER corpus, guidelines, and resources (including cross-mappings to MeSH and ICD-10) are freely available at: https://zenodo.org/record/7929830

Keywords

named entity recognition, entity linking, document indexing, clinical procedures, clinical NLP, SNOMED CT

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
salvador.limalopez@gmail.com (S. Lima-López)
0000-0002-7384-1877 (S. Lima-López); 0000-0002-7384-1877 (E. Farré-Maduell); 0000-0002-4976-9879 (L. Gascó); 0000-0002-2646-8782 (M. Krallinger)

1. Introduction

The recent use and exploitation of large language models (LLMs) and transformer-based technologies have resulted in considerable improvements in clinical NLP and health data processing systems, in particular for content in English, but also increasingly for other languages. Efficient semantic annotation strategies/named entity recognition of key clinical concepts, such as diseases, medications or adverse events are critical for medical text mining applications, including question answering, information extraction, predictive modeling or even generative artificial intelligence.

Procedures constitute an essential part of medical practice. From establishing a dialogue with the patient to prescribing a diet or performing a surgical procedure, procedures explain the decisions taken to solve specific problems. Medical texts are very rich in procedures, and mentions of procedures are very heterogeneous.

Thus, there is a clear need to foster the development of concept recognition systems for medical procedures to characterize existing treatment options, diagnostic procedures and to analyze key aspects of therapeutic or preventive techniques associated with patient care. Procedures are also relevant for clinical coding of electronic health records.

Despite underpinning clinical practice and healthcare, few attempts have focused on the automatic detection and normalization of medical procedures from clinical texts. Patel et al. [1] propose a corpus of different types of medical records annotated with 11 medical entities based on Unified Medical Language System (UMLS) classification [2], which includes procedures and also closely related entities such as medical devices. Relationships between anatomical entities and procedures were also annotated. The NEREL-BIO corpus [3] is a collection of abstracts from biomedical articles in English and Russian for nested named entity recognition (NER) with a total of 37 labels that distinguish between scientific, medical and laboratory procedures. In French, corpora such as MERLOT [4] or APcNER [5] also consider procedures within their annotation scheme. The former is a collection of de-identified clinical notes with 12 different labels, entity attributes for assertions and temporality and 37 relation types. The latter is a smaller corpus made up of clinical records annotated with 5 labels, including one for diagnostic procedures and another for therapeutic procedures. Unfortunately, none of these two corpora are freely available. Use of APcNER can be requested for research purposes. SemClinBr [6] is a corpus in Brazilian Portuguese that consists of de-identified clinical records from over 12 specialties obtained from multiple hospitals in Brazil that also uses the semantic types defined by UMLS. Similarly to APcNER, it distinguishes between diagnostic and therapeutic procedures. Additionally, it includes annotations for negation detection and abbreviation disambiguation. Finally, the Spanish corpus CT-EBM-SP [7] contains 1200 clinical trial documents annotated with four entity types obtained from UMLS: disorder, anatomical part, chemicals and procedures.

A common downside of all these corpora is that, despite the inspiration on UMLS for their annotation scheme, they do not normalize the annotated entities. Normalization, also referred to as entity linking, consists on matching each of the entities in a text to a knowledge source (such as a controlled vocabulary or ontology), so that different surface forms of a concept can be agglutinated. Proper entity linking or concept harmonization is a key step for efficient exploitation and analysis of the extracted information by predictive modelling approaches or advanced semantic search technologies.

Some widely-used knowledge sources that structure clinical procedures are ICD-10 (International Classification of Diseases, Tenth Revision) and SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms). ICD-10-PCS (International Classification of Diseases, Tenth Revision, Procedure Coding System) is specifically used by public and private health providers for billing, and also to obtain statistics on medical procedures. In contrast, SNOMED CT has been devised to structure medical records and its concepts adapt more closely to the writing of health professionals. As such, it is often the ontology of choice for the normalization of patients' records. SNOMED CT specifically contains the category PROCEDURE, but offers other possibilities when required (REGIME/THERAPY, MEDICINAL PRODUCT, and other).

To advance technology and research related to medical procedures, we present the MedProcNER Gold Standard corpus. The corpus builds on the DisTEMIST disease corpus [8, 9] by presenting manually curated annotations and SNOMED CT normalization for clinical procedures in the same collection of 1,000 documents from a variety of clinical specialties. In order to maximize the utilization and effectiveness of the MedProcNER corpus, it was employed for a collaborative task as part of the BioASQ and CLEF 2023 evaluation initiative. This paper presents an overview of the data and results of the MedProcNER Shared Task. It is structured as follows: Section 2 introduces the shared task, including its sub-tasks and evaluation methods. Next, Section 3 describes the MedProcNER corpus and other associated resources, while Section 4 presents the participation results and proposed methodologies. Finally, Section 5 concludes the paper with a discussion of some of the most interesting aspects, learned lessons, future work and more.

2. Task Description

2.1. Shared Task Description

The MedProcNER shared task challenges participants to create automatic systems that can extract different aspects of information about clinical procedures. More specifically, these aspects (explained in Section 2.2) are clinical procedure recognition, clinical procedure normalization and clinical procedure-based document indexing. Figure 1 gives a visual overview of the shared task and its setting.

To develop their systems, participants were asked to use the MedProcNER corpus, a Gold Standard dataset of 1,000 clinical case reports manually annotated by multiple clinical experts with clinical procedures with its mentions normalized to SNOMED CT codes. The MedProcNER is also sometimes referred to as ProcTEMIST due to its relation with the DisTEMIST corpus (same documents, different labels). Section 3.1 provides more detail about the corpus, its content and annotation process.

The participants' predictions were evaluated against the manual annotations done by clinical experts. Each team was allowed to submit up to 5 runs. The evaluation process and metrics is detailed in Section 2.3.

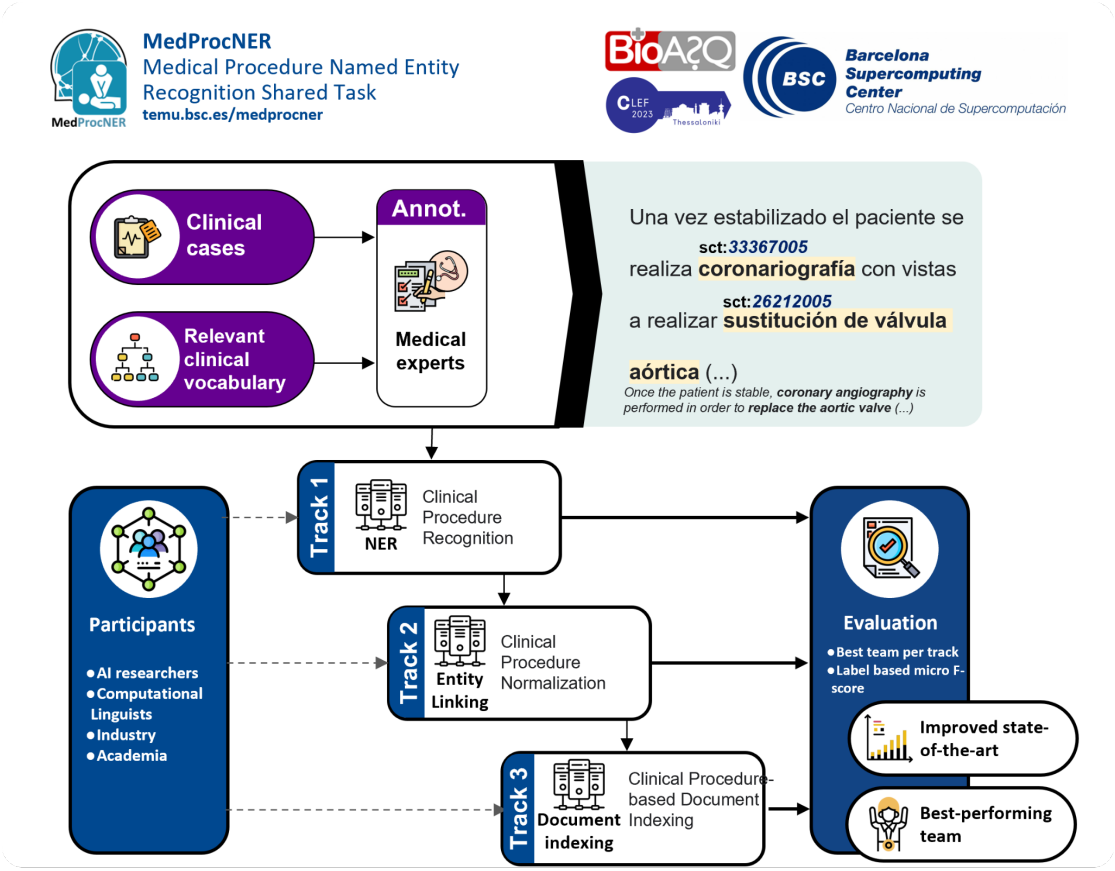

Figure 1: Visual overview of the MedProcNER Shared Task, originally used for task dissemination.

2.2. Sub-tasks

MedProcNER proposes three different sub-tasks, each focused on a different part of the information extraction pipeline:

- **Clinical Procedure Recognition:** This is a named entity recognition (NER) task where participants are challenged to automatically detect mentions of clinical procedures in a corpus of clinical case reports in Spanish.
- **Clinical Procedure Normalization:** In this entity linking (EL) task, participants must create systems that are able to assign SNOMED CT codes to the mentions retrieved in the previous sub-task.
- **Clinical Procedure-based Document Indexing:** This is a semantic indexing challenge in which participants automatically assign clinical procedure SNOMED CT codes to the full clinical case report texts so that they can be indexed. In contrast to the previous sub-task, participants do not need to rely on any previous systems, making this an independent sub-task.

2.3. Evaluation

All three MedProcNER sub-tasks are evaluated using micro-averaged precision, recall and F1-score. Micro-average calculations use the aggregated amount of true positives, false positives and false negatives over the entire test set. These metrics are calculated as follows:

$$\text{Precision (P)} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

$$\text{Recall (R)} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\text{F1 score (F1)} = \frac{2 * (P * R)}{(P + R)}$$

It is worth emphasizing that entity linking systems are not evaluated individually, but rather on an end-to-end fashion. This means that we did not give a complete list of mentions to be normalized and participants had to rely on their predictions from the named entity recognition step. As a result, the scores may not properly reflect the full performance of the systems. In turn, this type of evaluation provides a broader assessment of complete systems that is closer to how they would perform in a real-world application.

As part of the task, an official MedProcNER evaluation library was released and is available on GitHub[1].

2.4. Baseline

To provide a baseline for participants to compare their systems to, we created a system that uses a vocabulary transfer approach from the training to the test set. For the named entity recognition sub-task, we created a dictionary with all annotations from the training set. Next, the system uses a Levenshtein lexical lookup approach with a sliding window of varying length to try and find each dictionary entry in the test set texts. For the Entity Linking sub-task, we assign to each mention found in the test set the same code it had in the training set. Finally, for the Indexing sub-task we just group all codes used in the previous sub-task.

The baseline system's results are shown in Table 1.

Table 1
Results of the MedProcNER vocabulary transfer baseline system for each of the sub-tasks

Sub-task	P	R	F1
Entity Recognition	0.4731	0.4758	0.4744
Entity Linking	0.4599	0.4627	0.4613
Document Indexing	0.6285	0.508	0.5619

[1] https://github.com/TeMU-BSC/medprocner_evaluation_library

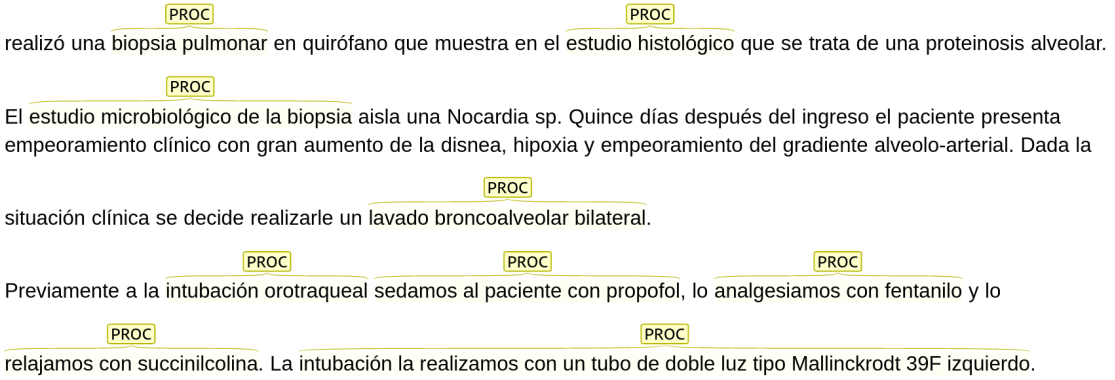

Figure 2: Excerpt from the MedProcNER corpus. Translation with annotated entities in italics: "[...] A *pulmonary biopsy* performed in the operating room reveals, in the *histological study*, that it is a case of alveolar proteinosis. The *microbiological study of the biopsy* isolates a Nocardia sp. [...] Given the clinical situation, a *bilateral bronchoalveolar lavage* is decided. Prior to *orotracheal intubation*, the patient is *sedated with propofol*, *provided analgesia with fentanyl*, and *muscle relaxation is achieved using succinylcholine*. We perform *intubation using a left Mallinckrodt 39F double-lumen tube*."

3. Corpus and Resources

This section explains all the different resources released for the MedProcNER shared task, including the MedProcNER corpus, its annotation guidelines, gazetteer for normalization and additional data.

3.1. MedProcNER Gold Standard corpus

Corpus overview. The MedProcNER Gold Standard corpus is a collection of 1,000 clinical case reports in Spanish from various medical specialties including, amongst others, oncology, odontology, urology and psychiatry. Clinical experts have manually annotated and then normalized with SNOMED CT mentions of procedures in the corpus. The documents were originally selected for a collection of documents known as the SPACCC corpus[2] and have been also used for the corpus and shared task on diseases DisTEMIST [8]. The MedProcNER corpus is publicly available on Zenodo[3].

Figure 3.1 shows an example of an annotated document with multiple procedure mentions, showcasing the complexity of some of the mentions included in the corpus.

MedProcNER is the first Gold Standard in Spanish containing manually annotated corpus of procedures in clinical documents, with the mentions manually mapped to SNOMED CT. Together with the DisTEMIST [8] and LivingNER [10] corpora and guidelines, it is part of an effort to promote the development and accessibility of annotated resources for clinical information extraction in Spanish validated by clinical experts. Other resources part of this initiative include PharmaCoNER [11], CANTEMIST [12] and MEDDOPROF [13].

[2]https://zenodo.org/record/2560316
[3]https://doi.org/10.5281/zenodo.7817745

Document selection. The case reports in the corpus were obtained from SciELO (Scientific Electronic Library Online)[4], an electronic library that contains publications from scientific journals of Latin America, South Africa and Spain. After their retrieval and pre-processing (which included extracting the appropriate sections and removing embedded figure references or citations), a set of 1,000 documents was manually selected by a practicing oncologist to ensure that they were relevant, content-rich and varied.

Corpus annotation. The MedProcNER corpus was originally annotated and standardized by two clinical experts from a Spanish tertiary hospital. The annotated mentions and their normalization were post-processed and revised afterwards by a third physician. The annotation process was performed concurrently with the DisTEMIST corpus. For annotations we used the brat tool [14].

Annotation and normalization guidelines were specifically created for this task. As with the DisTEMIST corpus, annotation involved discussions between physicians, particularly regarding complex mentions. This, together with multiple rounds of inter-annotator agreement (IAA) through parallel annotation of a section of the corpus, resulted in an iterative refinement of the guidelines. The total IAA score (computed as the pairwise agreement between two independent annotators) is of 81.2.

The MedProcNER annotation guidelines are further discussed in Section 3.2.

Corpus format. The MedProcNER text documents are released in plain text format with UTF-8 encoding, with the annotations presented in two different stand-off versions. The first version includes the original annotation files as outputted by brat [14]. These are .ann files, one for each text file, where each line represents an annotation, including its label, its start and end position and its associated text. The second version is a single tab-separated file (.tsv) which includes all annotations in the corpus. Similarly to the .ann files, this version includes one annotation per row with an additional field for the corresponding filename.

The normalization data are offered in .tsv format. It includes the same columns as the annotation data, with an extra column for codes. Composite mentions with more than one associated code are concatenated by the symbol "+". In addition to the assigned code, we provide four more columns with metadata given by the annotators that might be useful for normalization systems. The first is the semantic relation between the mention and the assigned code, which can be either EXACT (the code is a perfect match for the mention), NARROW (the code is more general than the mention) or NOCODE (the annotators were not able to find an appropriate code in the ontology for the mention). The other three are boolean values describing a) whether the mention includes an abbreviation; b) whether the mention is composite; and c) whether the annotators needed to read the mention in context to assign a correct code.

As for the indexing data, it is also offered in .tsv format. It is made up of one row per file. Each row has two columns: the filename and a list of codes that describe said file. As in the normalization data, different codes are separated using the symbol "+".

Corpus statistics. The MedProcNER/ProcTEMIST corpus includes 1,000 documents, which amount to 16,678 sentences and 350,764 tokens. MedProcNER uses the same splits as the DisTEMIST corpus, with the training and test set containing 750 and 250 documents, respectively.

The annotated corpus includes one label: PROCEDIMIENTO ("procedure" in Spanish), with a

[4] http://www.scielo.org

total of 14,684 mentions (7,179 unique).

For the normalization and indexing task, only a subset of 250 normalized documents taken from the training set was released. The rest of the annotated training data will be released as post-workshop material.

3.2. MedProcNER Annotation Guidelines

The MedProcNER guidelines describe how to annotate or label clinical procedure mentions in medical documents in Spanish, as well as how to map or associate them to their corresponding SNOMED CT codes. They were created *de novo* by clinical experts and iteratively refined after multiple rounds of parallel annotation.

The first version of the guidelines includes 31 pages and a total of 60 rules divided into different types (general, positive, negative and special). The guidelines also include a discussion of the task's importance and use cases, basic information about the task and annotation process, a description of different procedure types, a comparison with similar clinical entity types, and indications and resources for the annotators.

Difficulties related to annotation and normalization of procedures are: use of descriptive language; presence of acronyms and abbreviations; multiple parts (e.g. anatomical entities, techniques, instruments, materials); and ambiguous wording.

The rules within the guidelines describe what we consider a medical procedure and restrict how to annotate these mentions. Generally, we could classify the procedures annotated in the corpus as follows:

1. Medical exploration and inspection methods that require little or no instrumentation:*pulmonary auscultation; abdominal palpation; neurological exam*
2. Imaging and laboratory tests: *brain MRI; chest CT with contrast; blood count; EKG*)
3. Drug administration (excludes specific drug names, e.g. amoxicillin): *antibiotic therapy; corticosteroids; beta blockers.*
4. Administration of blood; plasma; serum; bolus and continuous medication pumps: *transfusion of 2 packed red blood cells; fluid therapy.*
5. Simplified surgical treatments: *transsphenoidal hypophysectomy; insertion of testicular prosthesis.*
6. Complex surgical descriptions: *placement under general anesthesia of an octopolar electrode (Octrode) in posterior spinal cords at the level of T9-T12; enucleation under general anesthesia using an oral approach through the upper vestibule.*

Importantly, the corpus does not include administrative procedures such as "discharge", "referral" or "admission".

The MedProcNER guidelines are available in Zenodo[5].

3.3. MedProcNER Gazetteer

The MedProcNER gazetteer has been built on the basis of the 31/10/2022 version of the Spanish edition of SNOMED CT. This version of SNOMED CT is composed of more than 300,000 concepts

[5]https://doi.org/10.5281/zenodo.7817666

organized in 19 different hierarchies including "procedure", "substance" and "regime/therapy". To simplify the entity linking and indexing task, we compiled a reduced subset of the terminology with a smaller set of concepts to which the mentions can be mapped. For the selection of concepts in the gazetteer, the hierarchy "procedure" was selected, as well as additional sub-hierarchies that were also detected as procedure related during the corpus annotation effort. It is important to note that codes associated with mentions in the test set were not utilized in generating the gazetteer. Consequently, any test set mentions that lacked a corresponding code in the gazetteer were not considered for the shared task evaluation.

This gazetteer is meant to be used as a reference for procedure normalization so that only a small subset of concepts is considered instead of the entire SNOMED vocabulary. It consists of 234,674 lexical entries, out of which 130,219 are considered main terms. Within these entries, there are 130,219 unique codes originating from 19 hierarchies. The hierarchy with the highest frequency is "procedure" (94,133 entries), followed by "substance" (40,846 entries), "clinical drug" (30,063 entries), and "medicinal product form" (18,390 entries). To generate the gazetteer, *Gaznomed*[6] repository was utilized to extract a separate file through concept tabulations from the RF2 files, which SNOMED CT employs for publishing its terminologies. During the generation process, any lexical entries with ambiguous meanings were excluded from the gazetteer.

3.4. Additional data

In addition to the Gold Standard corpus, annotation guidelines and normalization gazetteer, there are other additional resources that have been released as part of the task. All resources are available in Zenodo together with the Gold Standard corpus[7].

Multilingual Silver Standard

Following the popularity of the LivingNER and DisTEMIST Multilingual Silver Standard [8], we decided to release a multilingual version of the MedProcNER corpus. The aim of this multilingual version is to provide an approximation of the Spanish manual annotation in languages that currently do not have any annotated data to extract and evaluate this type of information. Despite the shortcomings of the Silver Standard, it may be used in different ways by researchers who wish to develop systems in their own language by training intermediate systems with acceptable results or even manually correct the generated data.

The MedProcNER Multilingual Silver Standard was available initially in six different languages: English, French, Italian, Portuguese, Romanian and Catalan. Later versions also included Dutch, Czech and Swedish. The translations and annotations were created using the same methodology described in the DisTEMIST overview paper [8]. Firstly, we took the SPACCC text files that were already translated for DisTEMIST. These were originally translated from Spanish to the target languages using different neural machine translation systems with manual checks performed to ensure translation quality. In parallel, a separate translation was done for a list of annotations without context. The translation was done using existing machine translation tools, such as the SoftCatala API[8] for Catalan.

[6] https://github.com/luisgasco/gaznomed
[7] https://doi.org/10.5281/zenodo.7817666
[8] https://www.softcatala.org/traductor/

Table 2
Overview of the teams that participated in MedProcNER. In the Affiliation column, A/I stands for academic or industry institution. In the Tasks column, R stands for entity recognition, L for entity linking and I for document indexing.

Team Name	Affiliation	Tasks	Reference
Onto-NLP	Ontotext, USA [I]	R/L	[17]
NLP-CIC-WFU	CIC IPN, México / Wake Forest University, USA [A]	R	-
Vicomtech	Vicomtech, Spain [I]	R/L/I	[18]
saheelmayekar	Freelance	R	-
Samy Ateia	Universität Regensburg, Germany [A]	R/L/I	[19]
SINAI	Universidad de Jaén, Spain [A]	R/L	[20]
Fusion	Sofia University, Bulgary [A]	R/L	[21]
BIT.UA	IEETA, University of Aveiro, Portugal [A]	R/I/L	[22]
KFU NLP Team	Kazan Federal University, Russia [A]	R/L/I	-

Next, the translated annotations were transferred to the translated text files using a look-up system. This look-up takes into account individual annotations, their translations and also a lemmatized version of the entities (obtained using spaCy[9]). In order to minimize the number of false positives and negatives, we only looked up in each document the annotations present in that file instead of all annotations in the corpus. Significantly, transferred annotations carry over the SNOMED CT code originally assigned to the Spanish annotation in the Gold Standard corpus.

Normalization cross-mapping

In previous editions of BioASQ, two MESINESP tasks [15, 16] were conducted to address the indexing of Spanish biomedical documents using the MeSH terminology and its Spanish version DeCS. In order to further promote the improvement of document indexing using these terminologies, and to enhance the reusability of MedProcNER data, we have generated cross-mappings from the SNOMED CT normalized mentions in the corpus to MeSH and ICD-10. The cross-mappings were performed through the UMLS Meta-thesaurus. This process establishes valuable connections between different medical vocabularies, facilitating the integration of multiple terminologies and enabling more effective document indexing and retrieval in the biomedical texts.

4. Results

4.1. Participation Overview

Out of 47 total registered teams, 9 different teams submitted at least one run of their predictions. In terms of sub-task participation, all 9 teams participated in the entity recognition sub-task, 7 participated in the entity linking sub-task, and 4 participated in the document indexing sub-task. In summary, a total of 68 runs were submitted. Table 2 shows a complete list of all participant teams.

[9] https://spacy.io/

Table 3
Results of MedProcNER Entity Recognition sub-task. The best result is bolded, and the second-best is underlined.

Team Name	Run name	P	R	F1
Onto-NLP	run1-bsc-bio-ehr-es-pharmaconer-voting	0.7397	0.4374	0.5497
	run1-bsc-bio-ehr-pharmaconer-voting-filtered	0.7425	0.4374	0.5505
	run1-pharmaconer_filtered_with_exact_match	0.3296	0.6104	0.428
NLP-CIC-WFU	Hard4BIO_RoBERTa	0.7132	0.6507	0.6805
	Hard4BIO_RoBERTa_postprocessing	0.7188	0.654	0.6849
	Lazy4BIO_RoBERTa_postprocessing	0.6301	0.6002	0.6148
Vicomtech	run1-xlm_roberta_large_dpa_e105	0.8054	0.7535	0.7786
	run2-roberta_bio_es_dpa_e119	0.7679	0.7629	0.7653
	run3-longformer_base_4096_bne_es	0.7478	0.7588	0.7533
saheelmayekar	predicted_data	0.3975	0.535	0.4561
Samy Ateia	run1-gpt3.5-turbo	0.523	0.2106	0.3002
	run2-gpt-4	0.6355	0.3874	0.4814
SINAI	run1-fine-tuned-roberta	0.7631	0.7505	0.7568
	run2-lstmcrf-512	0.7786	0.7043	0.7396
	run3-fulltext-GRU	0.7396	0.711	0.725
	run4-fulltext-LSTM	0.7538	0.7353	0.7444
	run5-lstm-BIO	0.7705	0.7049	0.7362
Fusion	run1-BioMBERT-NumberTagOnly	0.6948	0.6599	0.6769
	run2-BioMBERT-FullPrep	0.6894	0.6599	0.6743
	run3-XLM-RoBERTA-Clinical	0.7047	0.6916	0.6981
	run4-Spanish-RoBERTa	0.7165	0.7143	0.7154
	run5-Adapted-ALBERT	0.6928	0.6264	0.658
BIT.UA	run0-lc-dense-5-wVal	<u>0.8015</u>	<u>0.7878</u>	<u>0.7946</u>
	run1-lc-dense-5-full	0.7954	**0.7894**	0.7924
	run2-lc-bilstm-all-wVal	0.7941	0.7823	0.7881
	run3-PlanTL-dense-bilstm-all-wVal	0.7978	0.787	0.7923
	run4-everything	**0.8095**	<u>0.7878</u>	**0.7985**
KFU NLP Team	predicted_task1	0.7192	0.7403	0.7296

4.2. System Results

The complete results for the entity recognition, linking and document indexing are shown in tables 3, 4 and 5, respectively. The top-scoring results for each sub-task were:

- *MedProcNER Entity Recognition sub-task.* The BIT.UA team achieved the highest F1-score, 0.7985, highest precision (0.8095) and highest recall (0.7984) with their transformer-based solution that uses masked CRF and data augmentation. Teams Vicomtech and SINAI also obtained F1-scores over 0.75 using RoBERTa models.
- *MedProcNER Entity Linking sub-task.* The highest F1-score (0.5707), precision (0.5902) and

recall (0.5580) were obtained by Vicomtech, who used semantic search techniques, based on transformer models and cross-encoders (sapBERT). Teams SINAI and Fusion were also above 0.5 F1-score.

- *MedProcNER Document Indexing sub-task.* The Vicomtech team also obtained the highest F1-score (0.6242), precision (0.6371) and recall (0.6295), with the KFU NLP Team coming in second place (0.4927 F1-score).

Table 4
Results of MedProcNER Entity Linking sub-task. The best result is bolded, and the second-best is underlined.

Team Name	Run name	P	R	F1
Onto-NLP	run1-cantemist-top1	0.2642	0.4895	0.3432
	run1-ehr-top1	0.263	0.4873	0.3416
	run1-pharmaconer-top1	0.2742	0.508	0.3562
	run1-pharmaconer-voter	0.2723	0.5044	0.3536
Vicomtech	run1-xlm_roberta_large_dpa_e105_sapbert	**0.5902**	0.5525	**0.5707**
	run2-roberta_bio_es_dpa_e119_sapbert	0.5665	0.5627	0.5646
	run3-roberta_bio_es_dpa_e119_sapbert_condition	0.5662	0.5625	0.5643
	run4-roberta_bio_es_dpa_e119_sapbert_cross_encoder	0.5248	0.5213	0.523
	run5-longformer_base_4096_bne_es_sapbert	0.5498	**0.558**	0.5539
Samy Ateia	run1-gpt-3.5-turbo	0.4051	0.0749	0.1264
	run2-gpt-4	0.4304	0.1282	0.1976
SINAI	run1-fine-tuned-roberta	0.531	0.5224	0.5267
	run2-lstmcrf-512	0.5455	0.4936	0.5183
	run3-fulltext-GRU	0.5079	0.4884	0.498
	run4-fulltext-LSTM	0.5173	0.5047	0.5109
	run5-lstm-BIO	0.5352	0.4898	0.5115
Fusion	run1-BioMBERT-NumberTagOnly_XLMRSapBERT	0.5432	0.516	0.5293
	run2-BioMBERT-FullPrep_XLMRSapBERT	0.5332	0.5105	0.5216
	run3-XLM-RoBERTA-XLMRSapBERT	0.5332	0.5235	0.5283
	run4-Spanish-RoBERTa_predictions	0.5377	0.5362	0.5369
	run5-Adapted-ALBERT_predictions	0.5461	0.4939	0.5187
BIT.UA	run0-lc-dense-5-wVal	0.318	0.3126	0.3153
	run1-lc-dense-5-full	0.3143	0.3121	0.3132
	run2-lc-bilstm-all-wVal	0.3133	0.3087	0.311
	run3-PlanTL-dense-bilstm-all-wVal	0.3188	0.3145	0.3166
	run4-everything	0.3211	0.3126	0.3168
KFU NLP Team	predicted_task2	0.3917	0.4033	0.3974

4.3. Methodologies

The methodologies presented by the MedProcNER participants are a good showcase of some of the trends in Natural Language Processing and Information Extraction in the last few years, as

Table 5
Results of MedProcNER Indexing sub-task. The best result is bolded, and the second-best is in italics.

Team Name	Run name	P	R	F1
Vicomtech	run1_roberta_bio_es_dpa_e119_sapbert	0.6182	**0.6295**	0.6238
	run2_roberta_bio_es_dpa_e119_sapbert_cross_encoder	0.5885	0.5917	0.5901
	run3_longformer_base_4096_bne_es_sapbert	0.6039	*0.6288*	0.6161
	run4_xlm_roberta_large_dpa_e105_sapbert	**0.6371**	0.6109	*0.6239*
	run5_roberta_bio_es_dpa_e119_sapbert_condition	*0.619*	**0.6295**	**0.6242**
Samy Ateia	run1-gpt3.5-turbo	0.506	0.1083	0.1785
	run2-gpt-4	0.5266	0.1811	0.2695
BIT.UA	run0-lc-dense-5-wVal	0.3517	0.3619	0.3567
	run1-lc-dense-5-full	0.3475	0.3612	0.3542
	run2-lc-bilstm-all-wVal	0.3484	0.3593	0.3537
	run3-PlanTL-dense-bilstm-all-wVal	0.3544	0.3654	0.3598
	run4-everything	0.3551	0.3619	0.3585
KFU NLP Team	predicted_task3	0.4805	0.5054	0.4927

well as of some newest methods.

As is common in the current NLP landscape, most of the participant teams used Transformer-based models [23] and large language models. Out of nine participant teams in the NER sub-task, all nine of them used Transformers as their architecture of choice for their systems. Teams Ontotext, NLP-CIC-WFU, Vicomtech, SINAI and Fusion all used a RoBERTA [24] model in at least one of their runs, making it the most popular architecture in the task. Other architectures used include BioMBERT [25] (team Fusion), ALBERT [25] (also by team Fusion), Longformers [26] (team Vicomtech) or SapBERT [27] (team KFU NLP).

One participant, Samy Ateia, used Generative Pre-trained Transformer (GPT) models for all three sub-tasks. Specifically, they used the GPT 3.5-turbo and GPT 4 [28] models fine-tuned for the task using in-context few shot learning. The final performance of these systems was not particularly strong on any of the three sub-tasks, with the recall values being particularly weak. However, putting the submission into a wider context, the system probably required less training than other approaches, particularly due to its fine-tuning using a few-shot learning approach. Nevertheless, this advantage is somewhat counterbalanced by high computational cost of GPT models.

Some participants also went beyond model training and fine-tuning and tried to make the most out of the task data. On the one hand, team NLP-CIC-WFU used different pre-processing techniques to convert the training data into the BIO format (hard and lazy conversion). Then, they fine-tuned a RoBERTA model [24] and applied rule-based post-processing for subwords and possible tokenization issues. The hard preprocessing approach shows an improvement in the final F1-score of 0.07 points over the lazy approach. Interestingly, team BIT.UA (who obtained the highest results for the NER task) used some data augmentation techniques for their transformer-based solution with masked CRF.

The entity linking subtask encompassed a diverse range of systems, employing both super-

vised and unsupervised techniques. The majority of participants opted for approaches centered around semantic search and/or textual similarity to perform the normalization task. Notably, the top-performing team, team Vicomtech, employed SapBERT representations in conjunction with various strategies, including cross-encoder architectures. This SapBERT-XLMR model [29], a multilingual system designed to effectively represent biomedical concepts in vector form, was also used by the Fusion team using it in an unsupervised manner. The KFU team took a distinct approach by training a model using the Synonym Marginalization loss function proposed by Sung et al. [30]. Additionally, they incorporated UniPELT adapters, which facilitated a more efficient model-fitting process.

It is worth noting that the normalization system employing autoregressive models achieved limited results, specially in terms of recall. This outcome might suggest that the large language model is not effectively capturing the relationship between the mentions and the SNOMED terminology codes.

No novel systems were proposed for the Document Indexing sub-task. All participants repurposed their systems from previous sub-tasks.

5. Discussion

Comments on the annotation and normalization of clinical procedures.

Procedures constitute a complex category, which encompasses a wide diversity of clinically-associated methods. This is illustrated in the MedProcNER corpus, which contains mentions of procedures used in a wide variety of medical and surgical specialties. While some of the challenges, such as abbreviations (*Endosc. biop. ovary*), acronyms (*brain MRI*) and composite mentions (*transfusion of platelets and red blood cells*), are common to other entities when processing biomedical texts, particularly EHRs written in busy clinical environments, others are very specific to procedures.

As mentioned in Subsection 3.2, procedure mentions can range from very simple (*chest x-ray*) to extremely elaborate (*a small amount of pus was evacuated within the synovial sheath by means of a drain with opening of the A1 and A5 pulleys and irrigation system*). This, combined with the diversity of very specialized procedures in different specialties, resulted in frequent challenges during annotation and normalization.

Complex surgical descriptions (e.g. *placement under general anesthesia of an octopolar electrode (Octrode) in posterior spinal cords at the level of T9-T12; enucleation under general anesthesia using an oral approach through the upper vestibule*) are especially difficult. Despite restrictive rules established in the annotation guidelines, it is often difficult to determine where to start and finish a mention due to their descriptive nature and multiple sub-parts. We believe these type of mentions merit further work and exploration, able to capture the overall meaning while also considering the distinct parts within each mention.

Normalization also proved to be very challenging, sometimes caused by the concepts themselves and others due to the structure and coverage of SNOMED CT. SNOMED CT is a very rich ontology that is under permanent growth and development. Consequently, some branches being very rich in concepts while others might lack clinically relevant terms. Some SNOMED-related complications found during the manual concept normalization were:

1. **Missing concepts.** Despite the ontology's inclusion of very specific concepts (e.g. involving catheterization (for instance, *326919008 |Endoscopic catheterization of pancreatic duct and bile duct systems (procedure)|*), some more basic ones (such as the Brucella Coombs test) were harder to find or missing.
2. **Lack of generic/parent codes.** In some cases, we were able to find hyper-specific entries for a given procedure but not a more generic one. For instance, when looking for the generic concept "osteosynthesis", we found only children concepts such as *466257003 |Ultrasonic osteosynthesis system (physical object)|*.
3. **Similar descriptors.** A source of heterogeneity in concept IDs for the same clinical concept is the apparent duplication of codes who share almost equal descriptors. For instance, *55162003 |Tooth extraction (procedure)|* and *173291009 |Simple extraction of tooth (procedure)|* are two separate entries, with the latter being a child node of the former. This close similarity caused the manual annotation process to take longer, since the annotators had to compare both codes and explore the concept tree to make sure they were choosing the most appropriate one.

Future Work.

To summarize, this paper has presented the first systematic effort that specifically focuses on NER and entity linking strategies for clinical procedures in clinical case reports in Spanish, with the possibility of adaptation to other languages.

Future efforts will explore the association of procedures with other clinical entity types annotated in the same or related corpora, like diseases, symptoms, medications, or occupations. Additionally, with the aim to further improve the performance of NER systems for diseases and procedures, we are working on the development and release of new annotated collections focused on specific medical areas like cardiology, rheumatology, occupational health, toxic habits and rare diseases. We aim to link work on procedures with upcoming corpora related to anatomical entities, medical devices, biomedical materials and implants. Finally, we also plan to continue working on the multilingual silver standards by performing a formal evaluation of the annotation transfer quality as well as the usefulness of the data itself and of systems trained on it, as well as expand it to new languages.

Acknowledgments

We acknowledge the Encargo of Plan TL (SEDIA) to BSC for funding. Due to the relevance of medical procedures for implants/devices specially in the case cardiac diseases this project is supported by the European Union's Horizon Europe Coordination & Support Action under Grant Agreement No 101058779 (BIOMATDB) and DataTools4Heart Grant Agreement No. 101057849. We also acknowledge the support from the AI4PROFHEALTH project (PID2020-119266RA-I00).

References

[1] P. Patel, D. Davey, V. Panchal, P. Pathak, Annotation of a large clinical entity corpus, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,

Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2033–2042. URL: https://aclanthology.org/D18-1228. doi:10.18653/v1/D18-1228.

[2] P. L. Schuyler, W. T. Hole, M. S. Tuttle, D. D. Sherertz, The umls metathesaurus: representing different views of biomedical concepts., Bulletin of the Medical Library Association 81 (1993) 217.

[3] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities, Bioinformatics 39 (2023). URL: https://doi.org/10.1093/bioinformatics/btad161. doi:10.1093/bioinformatics/btad161.

[4] L. Campillos-Llanos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, A. Névéol, A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot), Language Resources and Evaluation 52 (2018). doi:10.1007/s10579-017-9382-y.

[5] I. Lerner, N. Paris, X. Tannier, Terminologies augmented recurrent neural network model for clinical named entity recognition, Journal of Biomedical Informatics 102 (2020). URL: https://hal.science/hal-02428771. doi:10.1016/j.jbi.2019.103356.

[6] L. Oliveira, A. Peters, A. Silva, C. Gebeluca, Y. Gumiel, L. Cintho, D. Carvalho, S. Hasan, C. Moro, Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks, Journal of Biomedical Semantics 13 (2022). doi:10.1186/s13326-022-00269-1.

[7] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine, BMC medical informatics and decision making 21 (2021) 1–19.

[8] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources (2022).

[9] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, A. Miranda-Escalada, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioasq 2022: The tenth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 337–361.

[10] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources, Procesamiento del Lenguaje Natural (2022).

[11] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

[12] A. Miranda-Escalada, E. Farré-Maduell, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages

Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
[13] S. Lima-López, E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, M. Krallinger, Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts, Procesamiento del Lenguaje Natural 67 (2021) 243–256. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6393.
[14] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
[15] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, R. T. Murasaki, E. Primo-Peña, C. Bojo Canales, G. Paliouras, M. Krallinger, et al., Overview of bioasq 2021-mesinesp track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials, CEUR Workshop Proceedings, 2021.
[16] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioasq 2021: the ninth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 239–263.
[17] P. Ivanov, A. Aksenova, T. Asamov, S. Boytcheva, Leveraging Biomedical Ontologies for Clinical Procedures Recognition in Spanish at BioASQ MedProcNER, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[18] E. Zotova, A. García-Pablos, M. Cuadros, G. Rigau, VICOMTECH at MedProcNER 2023: Transformers-based Sequence-labelling and Cross-encoding for Entity Detection and Normalisation in Spanish Clinical Texts, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[19] S. Ateia, U. Kruschwitz, Is ChatGPT a Biomedical Expert? - Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[20] M. Chizhikova, J. Collado-Montañez, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Coming a long way with pre-trained transformers and string matching techniques: clinical procedure mention recognition and normalization, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[21] S. Vassileva, G. Grazhdanski, S. Boytcheva, I. Koychev, Fusion @ BioASQ MedProcNER: Transformer-based Approach for Procedure Recognition and Linking in Spanish Clinical Text, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[22] T. Almeida, R. A. A. Jonker, R. Poudel, J. M. Silva, S. Matos, Discovering medical procedures in Spanish using Transformer models with MCRF and Augmentation, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/

paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[25] S. Alrowili, V. Shanker, BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 221–227. URL: https://www.aclweb.org/anthology/2021.bionlp-1.24.
[26] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv:2004.05150 (2020).
[27] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, 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, Online, 2021, pp. 4228–4238. URL: https://aclanthology.org/2021.naacl-main.334. doi:10.18653/v1/2021.naacl-main.334.
[28] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.
[29] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, arXiv preprint arXiv:2105.14398 (2021).
[30] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, arXiv preprint arXiv:2005.00239 (2020).