SOFTWARE

Open Access

Hybrid natural language processing tool for semantic annotation of medical texts in Spanish

Leonardo Campillos-Llanos[1*], Ana Valverde-Mateos[2] and Adrián Capllonch-Carrión[3]

*Correspondence: leonardo.campillos@csic.es

[1] ILLA - CSIC (Spanish National Research Council), C/Albasanz 26-28, 28037 Madrid, Spain
[2] Medical Terminology Unit, Spanish Royal Academy of Medicine, C/Arrieta 12, 28013 Madrid, Spain
[3] Centro de Salud Retiro, Hospital Universitario Gregorio Marañon, C/Lope de Rueda, 43, 28009 Madrid, Spain

Abstract

Background: Natural language processing (NLP) enables the extraction of information embedded within unstructured texts, such as clinical case reports and trial eligibility criteria. By identifying relevant medical concepts, NLP facilitates the generation of structured and actionable data, supporting complex tasks like cohort identification and the analysis of clinical records. To accomplish those tasks, we introduce a deep learning-based and lexicon-based named entity recognition (NER) tool for texts in Spanish. It performs medical NER and normalization, medication information extraction and detection of temporal entities, negation and speculation, and temporality or experiencer attributes (Age, Contraindicated, Negated, Speculated, Hypothetical, Future, Family_member, Patient and Other). We built the tool with a dedicated lexicon and rules adapted from NegEx and HeidelTime. Using these resources, we annotated a corpus of 1200 texts, with high inter-annotator agreement (average F1 = 0.841% ± 0.045 for entities, and average F1 = 0.881% ± 0.032 for attributes). We used this corpus to train Transformer-based models (RoBERTa-based models, mBERT and mDeBERTa). We integrated them with the dictionary-based system in a hybrid tool, and distribute the models via the Hugging Face hub. For an internal validation, we used a held-out test set and conducted an error analysis. For an external validation, eight medical professionals evaluated the system by revising the annotation of 200 new texts not used in development.

Results: In the internal validation, the models yielded F1 values up to 0.915. In the external validation with 100 clinical trials, the tool achieved an average F1 score of 0.858 (± 0.032); and in 100 anonymized clinical cases, it achieved an average F1 score of 0.910 (± 0.019).

Conclusions: The tool is available at https://claramed.csic.es/medspaner . We also release the code (https://github.com/lcampillos/medspaner) and the annotated corpus to train the models.

Keywords: Medical natural language processing, Medical text mining tool, Named entity recognition, Deep learning in healthcare, Clinical trials, Spanish medical NLP

Introduction

The substantial volume of medical data contained within electronic health records (EHRs), articles or clinical trials represents a potential source of evidence and knowledge discovery [1, 2]. However, information is predominantly stored in a unstructured format, which poses challenges for effective extraction and analysis. The application of natural language processing (NLP) techniques has facilitated more efficient text mining in the medical domain [3].

Comprehensive NLP tools can extract relevant information and alleviate the manual curation of data by healthcare professionals, which is time-consuming and error-prone [4]. Such type of system can automate cohort definition tasks for clinical trials by extracting medical concepts from eligibility criteria [5], and can enhance the discovery of disease-drug pairs for drug-repurposing [6, 7]. Likewise, it might be used for information extraction from clinical cases, with the goal of enhancing data analysis in large volumes of data [8] and identifying phenotype variance [9]—a step towards precision medicine [10]. Potential applications range from automatically extracting co-occurrence of drug-disease pairs in real data [11, 12], detecting comorbidities, disease recurrence and risk factors (e.g., in dementia [13], cardiovascular [14, 15] or mental health [13, 16, 17])—or EHR-based pharmacovigilance [18]. The reuse of clinical data can also be employed to identify potential participants who meet the eligibility criteria of clinical trials [19]. In these tasks, the detection of negation and speculation is essential for accurately excluding pathological states and comorbidities that are not associated with specific conditions [20]. Additionally, the processing of temporal information and the history of medical conditions is critical for predicting clinical events and managing chronic diseases [21].

A challenge in implementing this type of NLP system is the considerable effort required to tailor it specifically to each use case [22]. Ideally, an open-source medical NLP tool should be flexible enough to allow other teams to customize it for their particular needs. Additionally, such a system should be capable of processing a wide range of information, encompassing disorders, treatments or age-groups, negated and speculated events, drug contraindications and history of conditions or procedures. Although there has been a rise in the number of teams conducting research in Spanish medical natural language processing (NLP), integrated systems remain limited. A notable exception is the framework developed by [23], which facilitates the expansion of medical terminologies, enabling the reuse of information available in EHRs for advanced data analysis.

In this context, we introduce MEDSPANER, a Medical Semantic Python-Assisted Named Entity Recognizer for the Spanish language. The tool performs medical named entity recognition (NER), medication information extraction, temporal entity annotation and detection of negation and speculation. The system also normalizes concepts according to Concept Unique Identifiers (CUIs) from the Unified Medical Language System (herein, UMLS) [24] or SNOMED CT codes [25], using a dedicated lexicon [26].

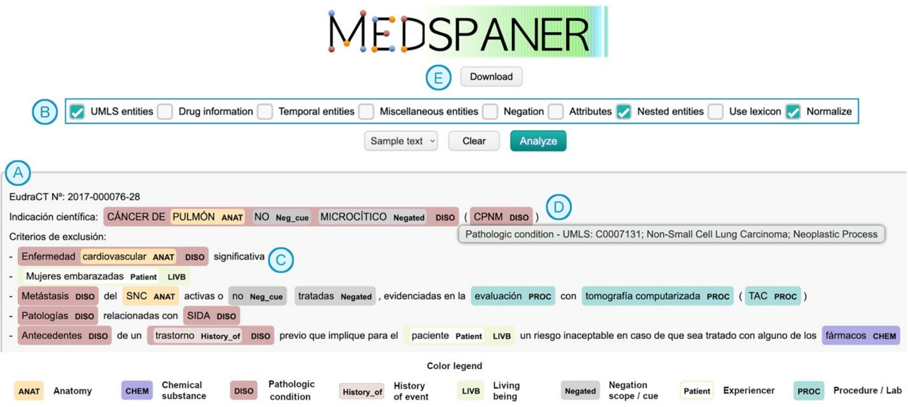
We developed the tool to automate the processing and analysis of clinical trials by medical professionals. While it was initially designed for this specific text genre, we are releasing it with the intention that it can also be applied to other medical sub-genres or adapted for additional tasks in the future. A demo video is provided as supplementary material. Our contributions are:

- An integrated tool that can be used with a UNIX-based terminal, or with a graphical user interface (GUI; Fig. 1). Annotations can be normalized to UMLS CUIs or SNOMED CT codes. The output can be `json` (which enables exporting data to a database such as Mongo DB) or the `ann` format (which enables loading the annotated files in the BRAT tool [27]).
- A set of fine-tuned Transformer-based models already available for the above-mentioned sub-tasks.
- An error analysis of the system's output, and a human evaluation by medical professionals, to show its strengths and limitations.
- An enriched annotation of the CT-EBM-SP corpus [28], used to develop the tool and to train the deep learning models. We also share the 200 annotated texts used in the human evaluation.

A supplementary graphical abstract summarizes the contents of this work [see Additional-File-1].

Background

NLP libraries in the Python language—e.g., Stanza [29] or spaCy [30]—currently perform fast automatic tokenization, sentence splitting and linguistic analysis (part-of-speech, syntax or semantic tagging). In addition, advances in neural-network-based frameworks [31, 32] provide modules to integrate the output of those linguistic analyses with contextual information available in pre-trained language models [33] or contextual embeddings [34]. In summary, the current NLP ecosystem enables the construction of complex pipelines for the processing of medical language. Some examples of Python libraries are those developed to lookup the UMLS Metathesaurus [35], for processing EHRs or other clinical texts [36–41].



☑ UMLS entities □ Drug information □ Temporal entities □ Miscellaneous entities □ Negation □ Attributes ☑ Nested entities □ Use lexicon ☑ Normalize

EudraCT Nº: 2017-000076-28

Indicación científica: CÁNCER DE PULMÓN ANAT NO Neg_cue MICROCÍTICO Negated DISO (CPNM DISO)

Pathologic condition - UMLS: C0007131; Non-Small Cell Lung Carcinoma; Neoplastic Process

Criterios de exclusión:

- Enfermedad cardiovascular ANAT DISO significativa
- Mujeres embarazadas Patient LIVB
- Metástasis DISO del SNC ANAT activas o no Neg_cue tratadas Negated , evidenciadas en la evaluación PROC con tomografía computarizada PROC (TAC PROC)
- Patologías DISO relacionadas con SIDA DISO
- Antecedentes DISO de un trastorno History_of DISO previo que implique para el paciente Patient LIVB un riesgo inaceptable en caso de que sea tratado con alguno de los fármacos CHEM

Fig. 1 Graphical User Interface. A text is copied on the text box (**A**) and several annotation options can be selected (**B**). The annotated entities are shown with different colors (**C**). Hovering over an entity shows information about its type and normalization (e.g. UMLS CUIs), if available (**D**). Results can be downloaded in BRAT format (**E**)

Multipurpose medical annotation tools feature medical entity annotation and classification—e.g., MedLEE [42], MetaMap [43] or MetaMap Lite [44]—and also negation and assertion detection—e.g., cTAKES [4]; or CLAMP [22]. For medical entity recognition, systems can use dictionaries [43]. These have the advantage of identifying entities in unsupervised contexts, and cope with infrequent term mentions, e.g., microbial species [45, 46]. Lexicon methods can be combined with rules and machine learning [4]. Recently, general models have been adapted or pre-trained on large scale biomedical and clinical data [47–50]. Recent experiments have tested dictionary-based approaches to feed a lexicon output into the neural model [51, 52].

With regard to temporal information [53], rules such as HeidelTime [54] or machine-learning classifiers have been applied, which yielded good results with clinical texts [55, 56]. Similar methods were used to process temporal entities and relations in clinical trials [57, 58]. For the detection of negation, uncertainty, temporality and experiencer attributes, methods have evolved from rules—e.g., NegEx [59] and ConText [60]—to supervised neural-network-models [61–65]. For the extraction of medication information, a similar transition has occurred from rule-based systems to neural-based or hybrid approaches [66, 67].

All the same, current advances of large language models (LLMs) and GPT-based methods are been applied to parse trial eligibility criteria and match patients to EHR data [68, 69]. Nonetheless, privacy issues to keep patients' protected information in each health institution poses an major limitation to commercial systems such as ChatGPT. In this context, Transformer-based models and open-source projects are a feasible alternative, given that they can be executed locally, and their results are competitive—sometimes with higher performance, according to some studies [70]. Combining such type of framework and a physician-in-the-loop to check the system output [71] seems a suitable approach.

Most tools are available for the English language and few tools are aimed at eligibility criteria of clinical trials [5] (no system exists for the Spanish language, to the best of our knowledge). Numerous teams have contributed to the task of medical entity recognition in Spanish; herein, we review only the works most related to ours (Table D.9 in Appendix D gathers more details). Initiatives include detecting drug effects in social media [72], drugs in clinical cases [73], or adverse drug reactions in EHRs by applying random forests (RFs) [74]. ICD-10 coding of disorders in EHRs was explored using RFs [75] or BERT models on radiology reports [76]. Also in radiology reports, a wide range of entities such as findings or body parts were annotated [77]. Detecting disabilities in Orphanet data was tested by means of SVM classifiers or Bi-LSTM networks [78] and BERT models [79]. In clinical referrals, entities were extracted by using a Bi-LSTM-CRF framework [80]; the same method combined with rules was applied to health reports of disease outbreaks [81]. Temporal entity recognition was performed by means of Bi-LSTM-CRF architectures in clinical narratives [82] or RoBERTa models in clinical cases [83]. As far as we know, detecting medication information (e.g. dosage) in Spanish was only explored on summaries of product characteristics by employing dictionary-based methods [84]. Negation detection was performed using syntactic methods on radiology reports [85], and negation/uncertainty was detected by applying Bi-LSTM-CRF or BERT-based models on product reviews, journal articles and clinical notes [62, 64, 65].

Overall, most methods have relied on Bi-LSTM networks or BERT models in the latest years. However, no comprehensive system exists for fast and actionable processing of that type of relevant information, and most works have focused on specific semantic aspects. No open-source tool has been developed that integrates annotation models for medical entities, negation and uncertainty, temporal information, and experiencer within a unified framework for Spanish medical texts. In addition, medical attributes of event temporality (e.g. `History_of`) or experiencer (e.g. `Patient` or `Family_member`) have not been considered in other NLP projects for the clinical literature in Spanish.

This work contributes with an integrated tool for medical language processing of texts in Spanish language. It extracts UMLS entities, medication-related data, temporal information, negation or speculation, experiencer or event temporality attributes, and miscellaneous entities such as observations, results or qualifiers. We combined a dedicated lexicon, the knowledge available in rule-based methods [54, 59] adapted to medical Spanish, and state-of-the-art Transformer-based models. Originally developed for clinical trials, the tool is now being released to be used with other medical sub-genres.

Implementation

The implementation involved several stages. The first steps were creating resources and rules for named entity recognition of the entities of interest. These resources enabled the pre-annotation of a task-specific corpus, which was revised by experts and then used to train deep learning (DL) models in a supervised setting. The MedLexSp lexicon of Spanish medical terms [26] with lemmas and form variants is used for domain entity recognition by employing a dictionary-based approach. The Unified Medical Language System (UMLS) [24] is the main source of domain terminology in MedLexSp. The tool normalizes concepts using SNOMED CT codes [25] or the UMLS concept unique identifiers (CUIs) recorded in MedLexSp.

For temporal entities, we developed rules by adapting HeidelTime [54]. For negation and speculation, we created rules by translating NegEx and ConText [59, 60] and reusing former work for Spanish [86]. Rules for medication-related information were manually developed and refined iteratively in corpus annotation. We also used the lists of administration routes and dosage forms from the Spanish Medicines Agency Nomenclátor [87]. We created rules for the following entity types: Dose, Concentration or Strength; Route or Mode of administration; and Dosage form.

The development of rules and the pre-annotation of domain texts were iterative processes: rules were redefined after analyzing the errors of the pre-annotated texts and after being revised by experts, then the improved rules were applied again to new batches of texts to revise, and so on. Figure 2 explains the workflow.

To develop the tools, we employed the Clinical Trials for Evidence-Based Medicine in Spanish (CT-EBM-SP) corpus [28]. This is a collection of 1200 texts about clinical trials (292173 tokens). A subset of 500 abstracts come from journals available in PubMed or the Scientific Electronic Library Online (SciELO). Another subset of 700 clinical trial announcements were published in the European Clinical Trials Register. The quality and consistency of the annotations were assessed by computing the inter-annotator agreement (IAA) between annotator pairs, which showed very good agreement (§ Annotation

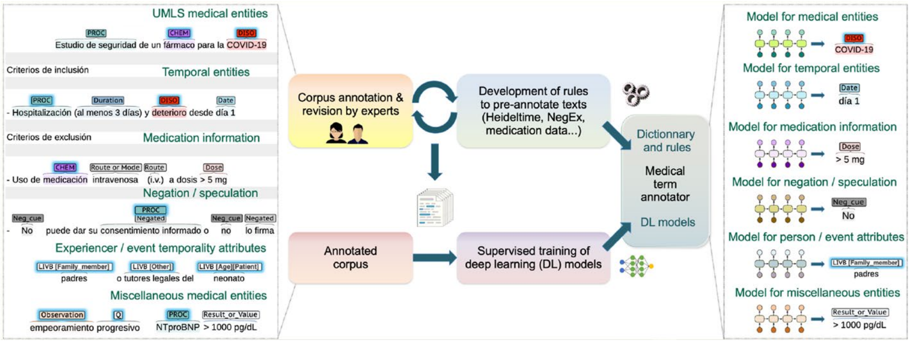
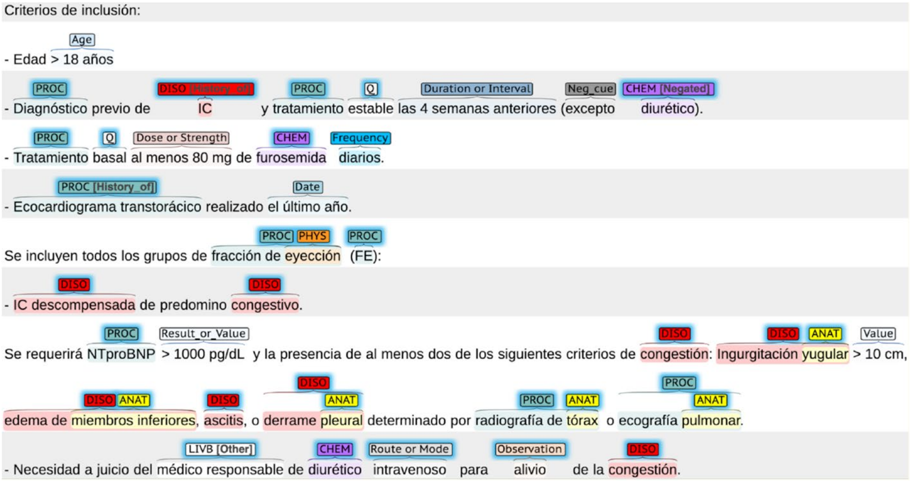

UMLS medical entities
PROC
CHEM
DISO
Estudio de seguridad de un fármaco para la COVID-19
Criterios de inclusión
Temporal entities
PROC
Duration
DISO
Date
- Hospitalización (al menos 3 días) y deterioro desde día 1
Criterios de exclusión
Medication information
CHEM
Route or Mode
Route
Dose
- Uso de medicación intravenosa (i.v.) a dosis > 5 mg
Negation / speculation
Neg_cue
PROC
Negated
Neg_cue
Negated
- No puede dar su consentimiento informado o no lo firma
Experiencer / event temporality attributes
LIVB [Family_member]
LIVB [Other]
LIVB [Age][Patient]
padres o tutores legales del neonato
Miscellaneous medical entities
Observation
Q
PROC
Result_or_Value
empeoramiento progresivo NTproBNP > 1000 pg/dL
Corpus annotation & revision by experts
Development of rules to pre-annotate texts (Heideltime, NegEx, medication data...)
Annotated corpus
Supervised training of deep learning (DL) models
Dictionnary and rules
Medical term annotator
DL models
Model for medical entities
DISO
COVID-19
Model for temporal entities
Date
día 1
Model for medication information
Dose
> 5 mg
Model for negation / speculation
Neg_cue
No
Model for person / event attributes
LIVB [Family_member]
padres
Model for miscellaneous entities
Result_or_Value
> 1000 pg/dL

Fig. 2 Workflow and system features



Criterios de inclusión:

- Edad > 18 años
- Diagnóstico previo de IC y tratamiento estable las 4 semanas anteriores (excepto diurético).
- Tratamiento basal al menos 80 mg de furosemida diarios.
- Ecocardiograma transtorácico realizado el último año.

Se incluyen todos los grupos de fracción de eyección (FE):

- IC descompensada de predomino congestivo.

Se requerirá NTproBNP > 1000 pg/dL y la presencia de al menos dos de los siguientes criterios de congestión: Ingurgitación yugular > 10 cm, edema de miembros inferiores, ascitis, o derrame pleural determinado por radiografía de tórax o ecografía pulmonar.

- Necesidad a juicio del médico responsable de diurético intravenoso para alivio de la congestión.

Fig. 3 Sample of the annotated corpus used to train the models

and IAA). The corpus and guidelines are available at the companion repository. Figure 3 is a sample of an annotated text.

We conducted an internal and external validation of the tool, as recommended [88]. The internal evaluation used a held-out test set, i.e. not used in model training. For the external validation, eight medical professionals revised the system annotations on 200 new texts (not used in system development): 100 clinical trial announcements, and also 100 anonymized clinical cases (to assess the performance of other medical sub-genres). The next sections provide more details of each component.

UMLS entities

In the first version of the corpus [28], only four UMLS semantic groups were annotated: anatomic entities (ANAT; e.g., *arm*), pharmacological and chemical substances (CHEM; e.g., *aspirin*), pathological conditions (DISO; e.g., *diabetes*) and laboratory, diagnostic and therapeutic procedures (PROC; e.g., *radiotherapy*). In the second version (distributed here), we annotated these entity types: medical devices (DEVI; e.g., *probe*), genes and genetic material (GENE; e.g., *BRAF*), physiological processes (PHYS; e.g., *breathing*) and living beings (LIVB, e.g., *patient*; note that this category also includes virus). Nevertheless, we did not use the GENE category to train the models, given the scarce number of instances and the low IAA scores achieved.

Temporal entities

We followed the TimeML annotation scheme [89] and annotated the following entity types: Date (e.g., *2022*), Duration (e.g., *dos horas*, 'two hours'), Frequency (called 'set' in TimeML; e.g., *semanalmente*, 'weekly') and Time (e.g., *noche*, 'night'). We did not merge entities beginning with *pre-/post-* prefixes, as in other annotation projects [90], which mark them as PREPOSTEX. PREPOSTEX generally gathers Date or Time entities that are ambiguous or unspecified (i.e. when there is not enough context to decide if *postoperative* refers to an amount of time expressed in days or hours). However, we kept the distinction between Date and Time in these expressions since it is more valuable for the clinical practice. Finally, we added the entity type Age (e.g., *18 years old*), given that age is an important factor for inclusion or exclusion in clinical trials. In addition, living being entities (LIVB) that also convey age information were annotated with the Age attribute (e.g., *adolescent*, *adults*).

Medication-related information

We considered the guidelines of previous corpora [66, 91, 92] and marked these types: Dosage form (Form, e.g., *píldora*, 'pill'), Route or Mode of administration (Route, e.g., *intravenoso*, 'intravenous') and Dose, Concentration or Strength (Dose, e.g., *125 mg*). We merged the categories of Dose and Concentration or Strength. This decision might provide coarse-grained results, but seemed more suitable for a general usage of the tool, which can be adapted if needed. Lastly, we annotated Contraindicated in some chemical and pharmacological entities (CHEM) or procedures (PROC). This is typically found in the exclusion criteria of trial announcements (e.g., *Pacientes con contraindicación a corticoesteroides*, 'Patients with a contraindication to corticosteroids'; the entity *corticosteroids* is Contraindicated).

Negation and speculation

We marked negation or speculation only on concepts or events. This choice was applied in other biomedical corpora [74, 93–96] and some annotation tools [4, 37]. We did not mark the full negated or speculated scope, as in other works [61, 63, 78, 97]. This criterion can be illustrated with the following sentence: *Los pacientes no habían recibido tratamiento antibiótico* ('The patients had not receive antibiotic treatment'). The scope of the negation would imply annotating 'had not receive antibiotic treatment'; however, we only annotated the concept 'antibiotic treatment' as Negated. This difference has impacted the performance of the NER models (see § Results). We marked negation and speculation cues (Neg_cue, Spec_cue), and events or entities within the scope of the negation or speculation were marked with Negated or Speculated, respectively. These were mostly marked on UMLS entities (ANAT, CHEM, DEVI, DISO, GENE, LIVB, PHYS and PROC).

Event temporality and Experiencer

Event temporality attributes are marked on entities referring to procedures and pathological or physiological conditions. These attributes specify whether the event occurred in the past (`History_of` or `Family_History_of`), if it will take place in the future (`Future`) or if it is `Hypothetical`. Experiencer attributes are only marked on `LIVB` entities referring to human beings. These attributes indicate if the entity is the experiencer of the condition or procedure (`Patient`), if it is a patient's relative (`Family_member`) or if he/she has another role (`Other`). We adapted the scheme from previous works [4, 60, 98] to the Spanish language. Importantly, we adopted entity and attribute types used in former annotation projects focused on EHR or clinical case data. This was motivated by the development of a comprehensive annotation framework aimed at improving the mapping of clinical patient data to relevant clinical trials. Nonetheless, we did not use the `Hypothetical` or `Family_History_of` attributes to train the models, due to the few instances in the corpus.

Miscellaneous entities

Lastly, we also annotated a set of clinical entities deemed necessary for the task: `CONC` (concepts), `Food/Drink` (*soy*), `Observation/Finding` (e.g. *relapse*), `Quantifier/Qualifier` (e.g. *at least 4, severe*) and `Result/Value` (e.g. < *3 UNL*).

Annotation and IAA

We applied the dictionary-based and rule-based tool to clinical trials texts to pre-annotate the data and then revise the output. Three experts annotated the data: a practicing medical doctor with 20 years of medical practice, a medical lexicographer with 18 years of experience in data curation and corpus annotation, and a computational linguist with 18 years of experience in data annotation and natural language processing. To define and learn the annotation guidelines, the experts annotated 12 texts, in several consensus rounds. We used the BRAT annotation tool [27]. After adequate agreement scores were reached, 112 texts were doubly annotated, again with meetings to fix disagreements and refine the guidelines iteratively. In the final step, the leading researcher annotated the remaining texts. Approximately 10% of the corpus (124 texts) was annotated by two or three experts.

We used the F-measure to assess the inter-annotator agreement (IAA) between pairs of doubly annotated sets. The F-measure is considered adequate for contexts where entities can have disparate spans [99]. The IAA values of four UMLS semantic groups (`ANAT`, `CHEM`, `DISO` and `PROC`) were reported in [28]. In the first version of the corpus, the IAA scores for these four entity types had an average F1 score of 0.856 (±0.048) in strict match (i.e. the full span and semantic class of the entity must match) after consensus annotations.

For the new annotated entities, the average IAA was an F1 score of 0.841% (±0.045) with consensus annotations (strict match). For attributes (`Age`, `Negated`, `Speculated`, `Contraindicated`, `Hypothetical`, `Future`, `Family_member`, `Patient` and `Other`), the IAA had an average F1 score of 0.881% (±0.032) with consensus (strict). Note that no IAA was computed for `Family_History_of`, because no instances appeared in the documents used for double annotation. IAA values were very good, given the large number of different entity types. However, IAA was lower for categories such as `Contraindicated`, `Form`, `Spec_cue` or `Speculated`, and these had an impact on the lower F1 results obtained with the tested models.

We annotated 86 389 entities (average of 71.99 per text) and 16 590 attributes (average of 13.82 per text). Figure 4 reports the number of annotations per entity type and the IAA scores. For named entity recognition, the annotations were converted to the standard BIO format (B stands for 'beginning' of entity; I, 'Inside', and O, 'Out'). The entity attributes were also processed to this format.

Transformer architecture in BERT models

We used the neural-network based Transformer architecture, as implemented in the Bidirectional Encoder Representations from Transformers (BERT) model [33]. The Transformer architecture replaces the recurrent neural network (e.g., Long Short-Term Memory models) with attention layers [100], and contain an encoder and a decoder (Fig. 5, a). Typically, 6 to 12 encoder layers and 6 to 12 decoder layers are stacked. The encoder converts the input sequence $(x_1, \ldots, x_n)$ into a sequence of continuous representations $(z = (z_1, \ldots, z_n))$. Encoding layers generate token embeddings and positional

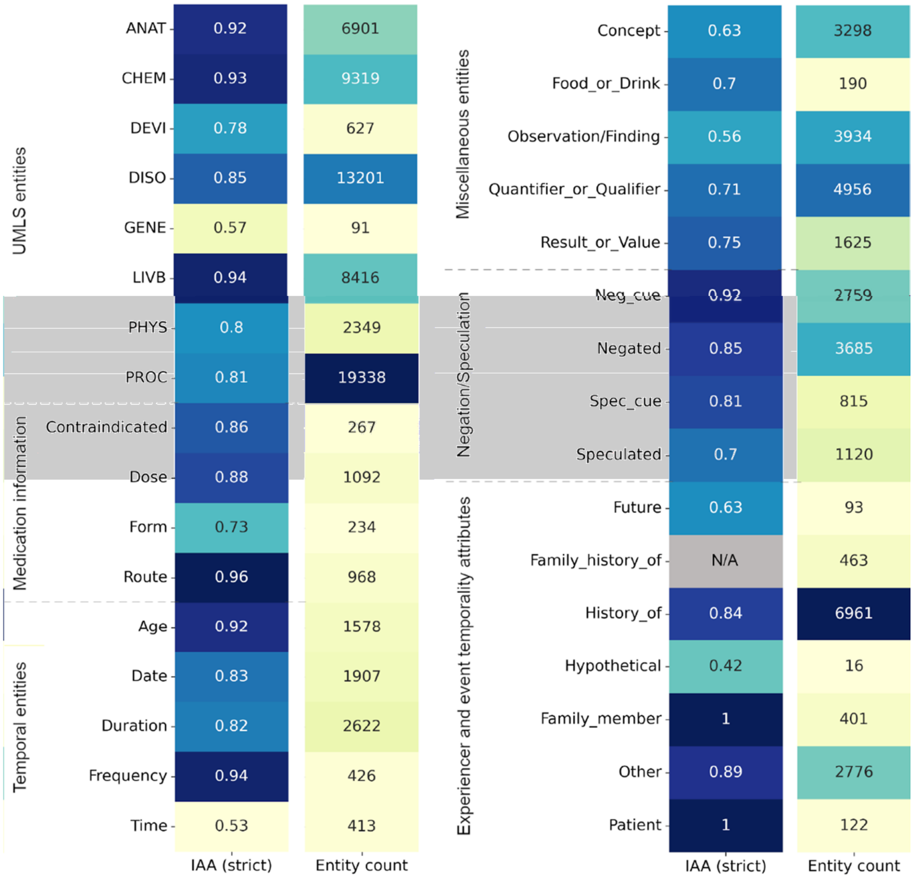


Fig. 4 Inter-annotator agreement (IAA) and count of each entity/attribute type

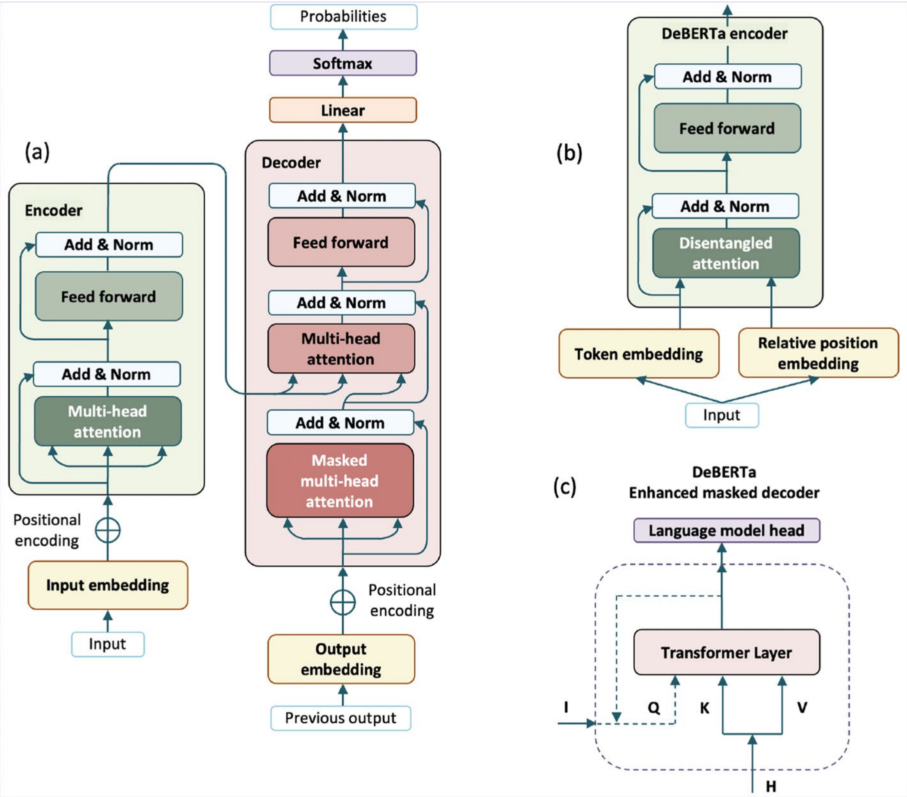


Fig. 5 Transformer architecture [100] (**a**); and modified encoder (**b**) and decoder (**c**) in DeBERTa [101]

embeddings, which bear the information about the location of each word and extract dependencies between them. Note that BERT and RoBERTa also learn representations of segment embeddings. Each layer has a self-attention mechanism that considers the entire context of the sequence, and weighs the importance of each word in relation to others. The multi-head self-attention component employs a set of *attention heads*, which are run in parallel to calculate Query (Q), Key (K) and Value (V) matrices to compute attention scores. The final attention output is passed through a feed-forward neural network with two fully connected layers and ReLU activation in between. A layer normalization is applied, and residual connections are added to the self-attention and the feed-forward components, which help stabilize training. With these z representations, the decoder outputs a sequence of symbols $(y_1, \ldots, y_n)$. The decoder consists of similar components to those in the encoder; since BERT models do not rely on the decoder for inference, due to space limits, we omit an exhaustive explanation and refer to [100] for more details. Once the sequence passes through all decoder layers, the output is fed into a linear layer followed by a SoftMax activation, which produces probability distributions over the vocabulary to predict the next token.

BERT, and its variants such as RoBERTa [102] or DeBERTa [101], are the most widely-used models with the attention mechanism for sequence labeling tasks. These models' architecture is composed of a stacked multi-layer bidirectional Transformer encoder, and separate the training in two steps. The first is pre-training a large task-independent language model, with a masked language model objective (i.e. the model has to predict a subset of hidden words). This pre-trained model is, secondly, fine-tuned with data for each task (e.g., named entity recognition). BERT models have an input length limit of 512 tokens at a time (vectors are truncated or padded to that fixed length) and may have 12 or 24 layers.

Figure 6 depicts the Transformer-based processing of a sentence, as implemented in BERT [33] or RoBERTa [102]. First, a `[CLS]` symbol is added at the beginning of every input sentence (for *classification* purposes) and a `[SEP]` (*separator*) token is put to separate each segment from the next one. The input text is split into word units (typically, at punctuation signs and white space characters) and sub-word units, using the `WordPiece` tokenizer [103]. Sub-words are generally used if the original words were not seen in the pre-training corpus, as it occurs in the medical domain: e.g., *homocisteína* ('homocysteine') is split in *homo*, *##ci*, *##ste*, *##ína*. Then, a Transformer block processes each sentence element and represents it with token, segment and position embeddings (Fig. 7 depicts this process for BERT and RoBERTa models). The segment embedding is a value corresponding to the sentence or phrase where each token occurs, separated by the `[SEP]` symbol. In Fig. 7, two segments appear: *niveles de homocisteína en pacientes en hemodiálisis* ('homocysteine levels in hemodialysis patients', segment A) and *Ensayo clínico* ('Clinical trial', segment B). Every token in the same segment has the same segment embedding vector. The input representation is combined by summing the token, segment and position embeddings to obtain a hidden representation with contextual information; note that DeBERTA models encode the input with two vectors (for token and relative position), as explained in the next section. Lastly, a token classification head for named entity recognition—made up of a dense layer and a SoftMax activation—outputs the semantic class and span of the named entities.

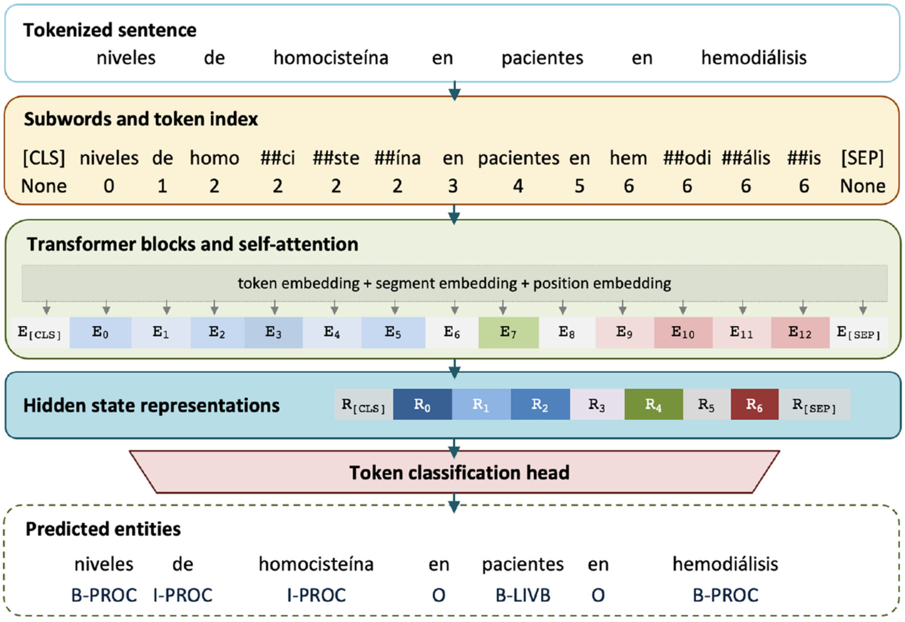

Tokenized sentence
niveles de homocisteína en pacientes en hemodiálisis
Subwords and token index
[CLS] niveles de homo ##ci ##ste ##ína en pacientes en hem ##odi ##ális ##is [SEP]
None 0 1 2 2 2 2 3 4 5 6 6 6 6 None
Transformer blocks and self-attention
token embedding + segment embedding + position embedding
$E_{[CLS]}$ E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12} $E_{[SEP]}$
Hidden state representations
$R_{[CLS]}$ R_0 R_1 R_2 R_3 R_4 R_5 R_6 $R_{[SEP]}$
Token classification head
Predicted entities
niveles de homocisteína en pacientes en hemodiálisis
B-PROC I-PROC I-PROC O B-LIVB O B-PROC

Fig. 6 General scheme of Transformer-based medical NER, as implemented in BERT [33]

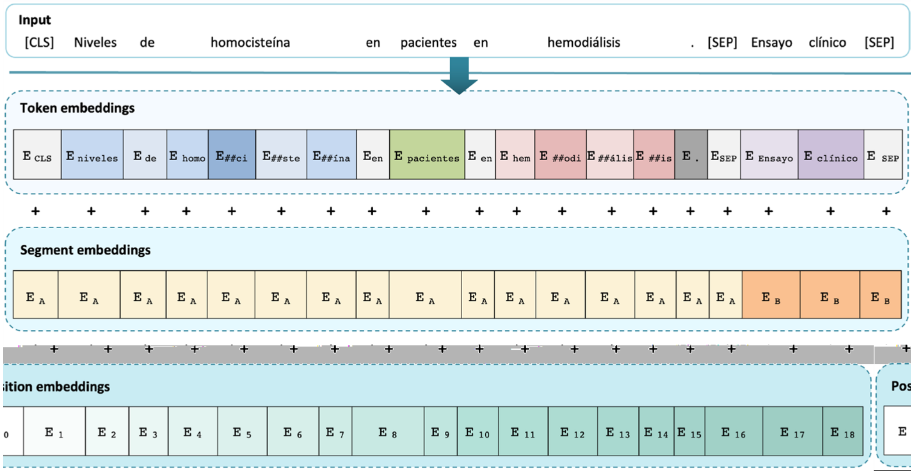


Fig. 7 Token, segment and position embeddings in the BERT and RoBERTa Transformer encoder [33]

Training of models

The annotated corpus was used in a supervised setting to train Transformer-based DL models. We fine-tuned six model variants for: (1) UMLS semantic groups; (2) medication information; (3) temporal entities; (4) negation and speculation; (5) miscellaneous medical entities; and (6) experiencer and temporality attributes.

First, we tested domain-specific RoBERTa-based [102] pre-trained models available for medical Spanish. RoBERTa (Robustly Optimized BERT Pre-training Approach) is a model variant that was pre-trained for longer (125000 additional steps), with a larger batch size (8000 samples per batch), using dynamic masking, and on a larger dataset (160 gigabytes) than the collection used to train BERT. RoBERTa also removes the Next Sentence Prediction task used for pre-training BERT, which makes it a more optimized model. We tested the `bsc-bio-ehr-es` model, which further pre-trained RoBERTa with medical texts in Spanish from SciELO, Wikipedia, EMEA or PubMed, and also with clinical data from EHRs [49]. This model uses the Byte-Pair Encoding (BPE) tokenizer, which is more efficient for subword tokenization. To analyze the influence of bilingual, domain pre-training data, we also tested EriBERTa [104] (base version), a bilingual RoBERTa-based bilingual model. EriBERTa base was pre-trained on an English and Spanish corpus. The English medical pre-training data include texts from EMEA, PubMed and ClinicalTrials.gov; and subsets of the Spanish collection come from SciELO, EMEA, PubMed, SNOMED CT and a Spanish clinical cases corpus. Lastly, we compared both models to the RoBERTa XLM Spanish Clinical model (hereafter, CLIN-X-ES) [50] large version. CLIN-X-ES is derived from the XML RoBERTA multilingual model (originally pre-trained on 2.5 terabytes of the CommonCrawl corpus for 100 languages), by continuous pre-training on a corpus of medical texts from SciELO, MedlinePlus, EMEA or PubMed. This model implements the XLM-R tokenizer and computes the input on subword level (not on word level); also, the cross-sentence context is incorporated in the input. Lastly, a Conditional Random Field (CRF) layer computes the output.

Second, we assessed multilingual, general-domain models, to examine further whether the difference in performance across models might be related to domain-specific or multilingual pre-training data. We evaluated the multilingual BERT (from here on, mBERT) base model, which was pre-trained on a corpus derived from Wikipedia in 104 languages. We also tested the Decoding-enhanced BERT with Disentangled Attention (DeBERTa) model, which presents two innovations in the BERT architecture. First, DeBERTa features a disentangled attention mechanism: each word is represented by separate vectors (not only by one sum vector, as in the original BERT) to encode the content and relative position separately (Fig. 5 b). With this approach, the self-attention layer models the dependency between near tokens in a better way. Second, DeBERTa features an enhanced mask decoder for pre-training (Fig. 5 c; *Q* stands for 'Query'; *K*, for 'Key'; *V*, for 'Value'; and *I*, for 'Input'; and *H*, for 'Hidden state'). The enhanced mask encoder employs the information about the content and the position for masked language modeling. Information about absolute positions is introduced in the decoding layer (before the SoftMax layer) to predict masked tokens during pre-training. By means of this technique, the model gives more importance to the absolute position of words. In our experiments, we specifically used the multilingual DeBERTa (hereafter, mDeBERTa) model vs 3 [101] base version, which is a multilingual, general-domain model pre-trained on the CC-100 corpus (CommonCrawl for over 100 languages).

Table 1 summarizes the characteristics of the Transformer-based models we tested and the pre-training details. To fine-tune and release the models, we used Transformers Hugging Face [32], which facilitates the reuse and adaptation of available models, and the replication of experiments. We used the `AutoModelForTokenClassification` class for NER. All the tested models are shared at the Hugging Face hub.

As baselines for comparison, we applied the rules we developed for temporal annotation and medication information extraction. We also tested the annotated corpus on a Bidirectional Long Short-Term Memory (Bi-LSTM) architecture with Conditional Random Fields (CRF) for sequence labeling. We used the implementation in the FLAIR framework, which models the linguistic context through *contextual embeddings* [34], but does not include any local nor global attention mechanism [105, 106]. The FLAIR implementation has one bidirectional LSTM layer with 1024 hidden states each (totaling 2048, 1024 for each direction).

To select the best configurations, we used the train (60%) and development splits (20%) from [28] (respectively, 720 and 240 texts). Once the best models were obtained, we tested then on a held-out set (20%, 240 texts). Because preliminary experiments for negation and speculation yielded poor results, we also used the NUBEs corpus [61] in

Table 1 Characteristics of the Transformer-based models and pre-training details (medical-domain models are italized in rows 2-4)

Model	PT corpus size	#A	#H	#L	#P	#V
RoBERTa EHR (**bsc-bio-ehr-es**)	>1B tok	12	768	12	125M	52K
EriBERTa (**EriBERTa-base**)	900M tok	12	768	12	125M	50K
CLIN-X-ES (**xlm-roberta-large-spanish-clinical**)	790MB	16	1024	24	550M	250K
mBERT (`bert-base-multilingual-cased`)	2.5T	12	768	12	110M	110K
mDeBERTa (`mdeberta-v3-base`)	2.5T	12	768	12	190M	250K

A: attention heads; *B*: billion; *H*: hidden size; *K*: thousand; *L*: number of layers; *M*: million;
MB: megabytes; *P*: parameters; *PT*: pre-training; *T*: terabytes; *Tok*: tokens; *V*: vocabulary size

Table 2 Data splits to train and test each model (listed in the header)

	UMLS entities, temporal entities, miscellaneous entities, temporality/experiencer attributes		Medication information		Negation/speculation	
	#Texts	**#Tokens**	**#Texts**	**#Tokens**	**#Texts**	**#Tokens**
Train	720	175203	720	175203	7739	693271
Train (CC)	820	205011	1085	285876	7839	723079
Dev	240	58670	240	58670	240	58670
Test	240	58300	240	58300	240	58300
Test (HE)	200	27332	200	27332	200	27332

CC: clinical cases; *HE*: human evaluation

Table 3 Fine-tuning hyperparameters of the tested models

Model	B	Mx Ep	LR	Optim	Pat	Seed
Bi-LSTM-CRF (FLAIR)	16	100	0.1	SGD	5	Random
RoBERTa, EriBERTa, mBERT and mDeBERTA vs 3	16	20	2e-05	Adam	5	{100, 200, 300, 400, 500}
CLIN-X-ES	8	30	2e-05	Adam	5	{100, 200, 300, 400, 500}

B: 'batch'; *LR*: 'fine-tune learning rate'; *Mx Ep*: 'maximum number of epochs'; *Pat*: 'Patience'; *Optim*: 'Optimizer'; *SGD*: 'stochastic gradient descent'

our data. We used train/dev/test splits and added them to the corresponding splits in our corpus. For the negation/speculation model, we manually converted the format and labels from the NUBEs corpus in order to fit our criteria of labeling negation and speculation (e.g., label names were changed).

Table 2 describes the data splits used to train, validate and test the models (in the internal evaluation and in the human evaluation, *HE*); we include the count of tokens and texts in the training data enriched with the NUBEs corpus (+518068 tokens). We also count the subsets of augmented data for applying the RoBERTa EHR and the CLIN-X-ES models to clinical cases (§Extending the models to clinical cases). These augmented data contain 100 clinical cases (*CC*, +29808 tokens) and summaries of product characteristics (SPCs, +51057 tokens).

We used a Zotac Geforce RTX 3090 GPU of 24 GB RAM for the experiments. The models were trained with early stop with a patience of 5 (i.e. the training stopped if the F1 score did not improve after 5 epochs). Table 3 shows the hyperparameters.

Annotation pipeline

Given an unstructured medical text (e.g., a clinical trial announcement or a clinical case), stage 1 involves pre-processing (e.g., changing characters that cause problems to the Transformer-based models), sentence splitting, tokenization and part-of-speech tagging by means of spaCy or Stanza (we used spaCy in this work). Then, the UMLS medical entities are detected with the dictionary (MedLexSp) [26] or with a BERT-based model. A list of exceptions can be applied to the output. For example, a specific semantic group can be excluded if it is unnecessary for a task: e.g. ACTI (activity entities), which are recognized only with the dictionary. Stage 2 involves named entity recognition of temporal entities, medication information, miscellaneous clinical entities, and attributes expressing negation/speculation or temporality/experiencer. This can be customized according to the user's needs (the output of stage 1 can be passed to stage 3 straightaway) and is processed sequentially with dedicated models. Lastly, stage 3 converts the output to JSON or BRAT format. Figure C.17 (Appendix) outlines the steps.

Extending the models to clinical cases

To test the generalizability of the trained models to other medical text types, we used 200 clinical cases already anonymized and prepared by medical residents or under a Creative Commons License.[1] We utilized 100 texts (comprising 29808 tokens) to further fine-tune the RoBERTa EHR and CLIN-X-ES models and adapt them to the new subdomain; and 100 texts (27332 tokens) for a human evaluation by medical professionals.

For the medication information model applied to clinical cases, we also included in the fine-tuning data 265 text samples (51057 tokens) of summaries of product characteristics (SPCs) from the Spanish Drug Information Center (CIMA). We selected SPCs from a subset of drugs included in the WHO Model List of Essential Medicines [107]. We used excerpts from sections corresponding to the Medication name, Dosage form, Administration mode and Contraindications. We also applied a data-augmentation technique for synonym replacement [108] using UMLS Concept Unique Identifiers available in MedLexSp [26]. For example, full forms (*intravenoso*, 'intravenous') were replaced with abbreviations/acronyms (*IV*). Some contexts required revision to fix gender agreement and replacement errors. In summary, we trained the models for medication information entities with the CT-EBM-SP corpus, 100 clinical cases and 265 text samples from SPCs (Table 2).

After inspecting the output of the models, we found that the RoBERTa EHR models were the most suitable for integration into the annotation pipeline of the medical entity recognizer (see §Results); accordingly, medical professionals evaluated only these models. We did not fine-tune the EriBERTa, mBERT or mDeBERTa models with the 100 clinical cases nor with the SPCs for the current version of the tool.

Evaluation

Internal validation

We computed precision (P, or positive predictive value), recall (R, or sensitivity) and F1 with the `seqeval` library [109]:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2PR}{P + R}$$

(*TP*: true positives; *FP*: false positives; *FN*: false negatives). We report micro-average scores because of the classification imbalance: e.g., anatomic entities (`ANAT`) or devices (`DEVI`) are scarce compared to chemical instances (`CHEM`). Results are reported on exact match at named-entity level (i.e. a true positive is counted if the models' prediction and reference match in scope and class). We provide the average and standard deviation of

[1] The Andalusian Society of Family and Medicine (SAMFyC) authorized us to use the cases they published. We also used clinical cases aimed at educational purposes from https://hipocampo.org/ and https://urgenciasbidasoa.wordpress.com/casos-clinicos-3/ [Accessed: 2024-23-02].

five experimental rounds with different initialization seeds. Carbon emissions of the models were estimated with a calculator online [110].

External validation (human evaluation)

As a use case in a real setting, eight medical professionals revised the annotations obtained with the tool (using the RoBERTa EHR models) on 200 new texts not employed for system development. First, 100 texts (29851 tokens) announcing trials published in 2022-23 were downloaded from EudraCT after developing the system. Second, to test the generalizability to other medical sub-genres, the evaluators checked the annotation of 100 clinical cases with a Creative Commons license (27332 tokens). Three medical doctors, four medical interns and one nurse participated; their average age was 33.0 years and the average years of medical practice was 6.88 years. Evaluators were instructed about the entity types and scope. The evaluation texts are available at the companion repository. Participation was remunerated afterwards.

Results

Results of the internal validation

Table 4 shows the results of the compared methods. Rules for medication information extraction and temporal entity recognition performed below the neural-network-based models, although the recall for the medication information entities (R = 0.786) was higher than that of the Bi-LSTM-CRF model (R = 0.740). For their part, Transformer-based models showed better performance compared to the Bi-LSTM-CRF model across all subtasks. The RoBERTa EHR and EriBERTa models performed similarly well, the latter tending to achieve slightly higher F1 scores. When comparing the performance of multilingual general-domain models, mBERT tended to yield lower scores than the RoBERTa-based models, but the mDeBERTa vs 3 model outperformed the other models for all tasks. However, the CLIN-X-ES models exhibited comparable performance, and outperformed the rest of RoBERTa-based models, excepting for the recognition of temporal entities. In Appendix B, Figure B.8 plots the F1 measure per model (averaged over 5 experimental runs); and Table B.8 presents the results per entity type of the RoBERTa EHR models, which were used in the human evaluation.

Regarding the training time of the clinical trial texts, the RoBERTA EHR model needed fewer training epochs; the estimated carbon footprint of each experimental round for the RoBERTa EHR models was of 63 g (eq. CO_2) per model (this value and the following are the average of 5 rounds). The EriBERTa models required a similar number of training epochs, and given their shorter training times, they had an estimated carbon footprint values of 33 g (eq. CO_2) per model. The CLIN-X-ES models also needed similar training epochs, but each one had longer training times; their estimated carbon footprint was of 209 g (eq. CO_2) per model. The mBERT and mDeBERTa models showed a similar trend, with fewer training epochs (although each epoch in mDeBERTa was longer). The estimated carbon footprint of each mBERT model was of 62 g (eq. CO_2), and the estimated carbon footprint of each mDeBERTa model was of 115 g (eq. CO_2). Lastly, the Bi-LSTM-CRF models needed more training epochs and longer training times; their estimated

Table 4 Comparison models (average ± standard deviation); medical-domain models are italized;general models are underlined; best results in bold (*P*: 'precision'; *R*: 'recall')

	UMLS entities			**Negation / speculation**		
	P	**R**	**F1**	**P**	**R**	**F1**
Bi-LSTM-CRF	0.815 (± 0.003)	0.813 (± 0.002)	0.814 (± 0.002)	0.845 (± 0.002)	0.776 (± 0.006)	0.809 (± 0.004)
mBERT	0.875 (± 0.004)	0.887 (± 0.005)	0.881 (± 0.001)	0.855 (± 0.006)	0.852 (± 0.010)	0.853 (± 0.004)
RoBERTa EHR	0.878 (± 0.003)	0.894 (± 0.003)	0.886 (± 0.002)	0.855 (± 0.005)	0.864 (± 0.009)	0.859 (± 0.006)
EriBERTa	0.881 (± 0.005)	0.896 (± 0.002)	0.889 (± 0.003)	0.861 (± 0.008)	0.871 (± 0.005)	0.866 (± 0.006)
CLIN-X-ES	**0.906** (± **0.005**)	0.911 (± 0.008)	0.909 (± 0.003)	**0.871** (± **0.008**)	**0.874** (± **0.007**)	**0.873** (± **0.005**)
mDeBERTa	0.904 (± 0.001)	**0.916** (± **0.003**)	**0.910** (± **0.002**)	0.861 (± 0.009)	0.871 (± 0.009)	0.866 (± 0.006)
	Temporal entities			Medication information		
	P	R	F1	P	R	F1
Rules	0.824	0.841	0.833	0.628	0.786	0.698
Bi-LSTM-CRF	0.899 (± 0.007)	0.859 (± 0.005)	0.879 (± 0.006)	0.895 (± 0.002)	0.740 (± 0.015)	0.810 (± 0.009)
mBERT	0.874 (± 0.007)	0.862 (± .006)	0.868 (± 0.004)	0.856 (± 0.007)	0.823 (± 0.011)	0.839 (± 0.004)
RoBERTa EHR	0.900 (± 0.011)	0.900 (± 0.007)	0.900 (± 0.001)	0.873 (± 0.022)	0.871 (± 0.015)	0.872 (± 0.004)
EriBERTa	**0.922** (± **0.004**)	0.908 (± 0.011)	**0.915** (± **0.007**)	0.874 (± 0.024)	0.828 (± 0.007)	0.850 (± 0.010)
CLIN-X-ES	0.899 (± 0.017)	0.895 (± 0.005)	0.897 (± 0.010)	0.883 (± 0.021)	0.863 (± 0.015)	0.872 (± 0.008)
mDeBERTa	0.909 (± 0.009)	**0.918** **(± 0.006)**	0.913 (± 0.005)	**0.897** **(± 0.012)**	**0.885** **(± 0.014)**	**0.891** **(± 0.008)**
	Miscellaneous entities			Experiencer/Temporality attributes		
	P	R	F1	P	R	F1
Bi-LSTM-CRF	**0.721** (± **0.006**)	0.536 (± 0.005)	0.613 (± 0.004)	0.891 (± 0.007)	0.816 (± 0.004)	0.852 (± 0.001)
mBERT	0.674 (± 0.027)	0.623 (± 0.015)	0.647 (± 0.006)	0.868 (± 0.023)	0.831 (± 0.014)	0.849 (± 0.005)
RoBERTa EHR	0.685 (± 0.008)	0.669 (± 0.004)	0.677 (± 0.003)	0.877 (± 0.009)	0.835 (± 0.008)	0.856 (± 0.006)
EriBERTa	0.703 (± 0.017)	0.666 (± 0.009)	0.684 (± 0.006)	0.890 (± 0.007)	0.848 (± 0.008)	0.868 (± 0.002)
CLIN-X-ES	0.715 (± 0.014)	**0.672** (± **0.016**)	**0.692** (± **0.007**)	0.894 (±0.013)	0.863 (± 0.010)	0.878 (± 0.005)
mDeBERTa	0.702 (± 0.011)	0.670 (± 0.007)	0.686 (± 0.004)	**0.898** **(± 0.009)**	**0.882** **(± 0.008)**	**0.890** **(± 0.005)**

carbon footprint was of 263 g (eq. CO_2) per model. Table A.6 in Appendix A includes all the details.

With regard to the annotation times, we compared both RoBERTa-based models using the 100 new EudraCT texts and the 100 clinical cases for the human evaluation. Speed of processing was measured in a laptop with a CPU Intel Core i7 processor (2.8 GHz), 16 GB of RAM and Mac OS. Table A.7 in Appendix A shows that the RoBERTa EHR and EriBERTa models were faster compared to and mDeBERTa and CLIN-X-ES; we did not compare them with the Bi-LSTM or the mBERT models, given the worse performance of the latter models.

We inspected the errors made by these models on a subset of the test split (§Discussion). The CLIN-X-ES models made critical errors in medication and temporal information, and the mDeBERTa models did not label several medical abbreviations and acronyms. Moreover, these models had longer annotation times and a higher environmental impact of fine-tuning. We thus selected the RoBERTa EHR models to be evaluated by medical professionals (note that the EriBERTa model was publicly released after the human evaluation has conducted). In the current version of the medical entity recognition tool, we integrated the RoBERTa EHR models; however, we release the rest of fine-tuned models in the Hugging Face hub.

Results of the human evaluation

Table 5 shows the comparison between the evaluators' revisions and the system output of 100 new texts from EudraCT and 100 clinical cases. The average F1 score (strict match) of the tool was of 0.858 (± 0.032) on the 100 trial announcements, and of 0.910 (± 0.019) on the 100 clinical cases. Interestingly, performance on the 100 cases was higher than on the 100 new EudraCT texts. The fact that the RoBERTa EHR model was trained on clinical data may account for this outcome.

System predictions regarding the seven UMLS entity types and the temporal annotations were in line with the results obtained in our held-out test set. In the 100 new EudraCT trials, results for medication information and negation or speculation were moderately below our scores in the test set; but note that their standard deviation scores were higher. The system performed very well in some texts, but not in other ones. A similar trend is observed in models for miscellaneous medical entities and experiencer and temporality attributes. The models for medication information, negation and speculation, miscellaneous medical entities, and experiencer and temporality attributes seem to vary widely in a real-world setting.

Upon analyzing the discrepancies between the system's predictions and the human evaluations, we observed that the tool failed to annotate certain Route entities that were missing in the training data (e.g., *systemic*, which caused inconsistencies across annotators). Negation and speculation caused several misinterpretations regarding the scope of `Negated` or `Speculated`. In particular, contexts involving laboratory tests, or where the scope is unclear, were often ambiguous for human evaluators. Several mismatches affected the PHYS and Observation categories, especially in entities expressing normal findings (e.g., *normoperfundido*, 'normal perfusion'), and generally in the 100 clinical cases. The `PROC` semantic group also caused ambiguities in diagnostic tools or scales (e.g., *ECOG*) or observable entities that can be interpreted as an observation (e.g., *left ventricular ejection fraction*).

Table 5 Comparison of system predictions and human evaluation of 100 new trials and 100 clinical cases

	Strict			**Relaxed**		
	P	**R**	**F1**	**P**	**R**	**F1**
EudraCT						
UMLS entities	0.908 (± 0.037)	0.879 (± 0.049)	0.893 (± 0.042)	0.965 (± 0.018)	0.928 (± 0.029)	0.946 (± 0.020)
Temporal entities	0.901 (± 0.098)	0.899 (± 0.086)	0.899 (± 0.085)	0.933 (± 0.062)	0.933 (± 0.066)	0.931 (± 0.049)
Medication information	0.838 (± 0.155)	0.793 (± 0.158)	0.803 (± 0.140)	0.934 (± 0.122)	0.879 (± 0.161)	0.893 (± 0.112)
Negation / speculation	0.785 (± 0.153)	0.819 (± 0.095)	0.795 (± 0.125)	0.892 (± 0.126)	0.912 (± 0.056)	0.899 (± 0.090)
Miscellaneous entities	0.801 (± 0.075)	0.714 (± 0.073)	0.754 (± 0.068)	0.872 (± 0.064)	0.776 (± 0.061)	0.820 (± 0.055)
Experiencer / temporality	0.897 (± 0.069)	0.854 (± 0.093)	0.873 (± 0.074)	0.939 (± 0.113)	0.908 (± 0.091)	0.918 (± 0.093)
Overall	0.876 (± 0.029)	0.842 (± 0.039)	0.858 (± 0.032)	0.942 (± 0.020)	0.897 (± 0.028)	0.919 (± 0.019)
		Clinical cases				
UMLS entities	0.946 (± 0.027)	0.929 (± 0.023)	0.937 (± 0.024)	0.971 (± 0.020)	0.949 (± 0.014)	0. 960 (± 0.015)
Temporal entities	0.955 (± 0.063)	0.952 (± 0.056)	0.953 (± 0.059)	0.988 (± 0.018)	0.986 (± 0.016)	0.987 (± 0.014)
Medication information	0.948 (± 0.905)	0.945 (± 0.104)	0.946 (± 0.096)	0.989 (± 0.033)	0.990 (± 0.033)	0.989 (± 0.027)
Negation / speculation	0.938 (± 0.063)	0.968 (± 0.031)	0.952 (± 0.042)	0.962 (± 0.059)	0.992 (± 0.014)	0.976 (± 0.034)
Miscellaneous entities	0.757 (± 0.085)	0.878 (± 0.060)	0.812 (± 0.072)	0.796 (± 0.068)	0.933 (± 0.033)	0.858 (± 0.048)
Experiencer / temporality	0.759 (± 0.126)	0.880 (± 0.110)	0.811 (± 0.106)	0.803 (± 0.117)	0.932 (± 0.094)	0.859 (± 0.092)
Overall	0.896 (± 0.023)	0.924 (± 0.018)	0. 910 (± 0.019)	0.926 (± 0.018)	0.955 (± 0.010)	0.940 (± 0.012)

The medical professionals also gave us some feedback. Some aspects revolved around the scope of the annotation, which needed discontinuous entity marking (e.g., *pregnant or lactating patients*). However, discontinuous annotations are currently not supported in this version of the tool.

Discussion

Overall, Transformer-based models achieved the highest scores compared to our manually-crafted rules or the Bi-LSTM-CRF framework. In the comparison between the RoBERTa-based models, we observed that a bilingual, domain-specific model (EriBERTa) slightly surpassed the monolingual medical model RoBERTa EHR (except for the recognition of medication information). This suggests that cross-lingual transfer learning has a positive impact. Indeed, CLIN-X-ES outperformed both models, and a plausible explanation might be that it was pre-trained on the multilingual XLM RoBERTa model and further pre-trained on Spanish medical corpora. Interestingly, we identified contradictory findings when comparing multilingual general models to monolingual domain models. Although mBERT tended to show a lower performance, the mDeBERTa model generally outperformed RoBERTa ERH and CLIN-X-ES, which were both trained with medical data. One explanation could be the specific architecture of the mDeBERTa model, namely the disentangled attention mechanism and the use of absolute position information, which enables the model to learn better dependencies between words. To confirm whether this performance is due to the multilingual pre-training data or to the model architecture, a medical mDeBERTa model should also be tested in our data, which is, however, currently unavailable. Nonetheless, the mDeBERTa and CLIN-X-ES models did not extract key medical entities, abbreviations and acronyms for our task; therefore, we used the RoBERTa ERH models in the external evaluation by medical professionals.

We analyzed the errors in the output of the RoBERTa EHR model for the human evaluation. Figure B.9 in Appendix B includes a detailed error analysis based on an standard taxonomy [111]; Figures B.10-B.15 include examples; and Figure B.16 shows the confusion matrices of predicted and gold standard labels per category. Many errors involved the scope of `DISO`, `PROC`, `Duration`, `Observation`, `Negated` or `Speculated` entities. These occurred when the model labeled modifiers as part of the entity (e.g., *small children*), or vice versa. False negatives occurred in classes with scarce instances in the corpus (`ANAT`, `DEVI`, `PHYS`, `Contraindicated`, `Form`, `Food` or `Future`, and lexical negation/speculation cues). Proper names and acronyms caused both false positives (e.g., *dialyzer AN69ST®*) and false negatives (*GE*, 'experimental group'). The semantic class was often wrong between `CHEM` and `PROC`, `Date` and `Time`, `DISO` and `Observation`, or `Negated` and `Speculated`.

This analysis revealed that a contributing factor to the observed errors is the insufficient number of annotated samples for specific categories within the training dataset. Another source of errors occurred under the zero-shot setting, when unseen instances in the training data (e.g., acronyms or brand names) were unannotated or misclassified. These would be the major weaknesses of the current version of the tool, especially if false negative errors cause valuable information loss. Errors affecting the scope of entities seem to arise owing to ambiguous contexts, in particular of negated and speculated phrases. Nonetheless, human evaluators did not find severe errors in this respect; and in the case of multi-word entities, missing tokens might be considered a minor error if the entity head is labeled (e.g. *gastric bleeding* vs. *upper gastric bleeding*). The error analysis also provided us with insights into future improvements. To alleviate the false negative and zero-shot issues, we would need to conduct more annotations—specifically for the infrequent classes—and train the models with enriched annotated data. To improve the detection of negation and speculation, we could try methods based on syntax [112] or data augmentation [113].

We next present a comparative analysis of our results with those reported in other studies for similar tasks. We refer to recent works for a comparison of systems on EHRs [114] and an analysis of encoder-based models on clinical datasets [115]. Table D.9 in Appendix D includes specific figures reported by the authors, with an exact match criterion, excepting the work by [67], who reported relaxed-match results. Replicating other teams' results was out of the scope of this work.

In general, direct comparison of our results is not feasible due to the substantial variability among annotation schemes. We used a subset of UMLS semantic groups; however, the UMLS subset of cTAKES differs from our scheme (e.g., we did not separate signs/symptoms and disorders into different categories). For English clinical trial data, the UMLS was not used [116–118]; and Criteria2Query [5] was evaluated on the OMOP scheme [119].

Regarding temporal entities, our results with a Transformer model are slightly higher than those reported by other teams. However, we used the `Age` class and the TimeML scheme. Other groups restricted to the four categories of TimeML and achieved competitive results using rule-based methods (generally based on HeidelTime [120, 121]), machine-learning (as implemented in the cTAKES' temporal expression extractor [55]) and neural models [82, 83].

Our annotation scheme also differs from other studies regarding drug-related entities. We merged Dose and Strength, and did not include classes such as Condition or Reason, considered in other works [66, 92]. Results cannot be compared to those by other teams who applied rule-based methods [122], deep learning-based algorithms [66] or ensemble methods [67]. Nonetheless, to the best of our knowledge, no similar work has been done on medication data in Spanish texts using deep learning.

Regarding negation and speculation, we annotated this type of information exclusively for concepts or events, which is similar to the cTAKES approach [4]. Although we did not annotate the scope of negation or speculation, most state-of-the-art works achieved similar outcomes. Recognizing cues generally achieves higher scores compared to scopes, and this is consistent across medical reports [61, 64], medical literature [62] or clinical trials [5]. This trend is also observed with neural-network-based approaches in Spanish, French or Brazilian Portuguese [63]. We also obtained higher scores recognizing negation rather than speculation.

As for event temporality attributes (`History_of`, `Family_History_of` and `Future`), comparison with other studies is challenging, as these classes are frequently merged into more general categories. Some examples are status attributes in [4] and historical attributes in ConText [60]. Few projects considered experiencer attributes (`Patient`, `Other` and `Family_member`), but consistent with our findings, these generally demonstrated high scores.

Miscellaneous medical entities show a broader variability in results. Observations or Findings did not achieve the highest scores, which is consistent with the results reported by other teams [77, 95, 123]. This class can be confused with disorders or results, and models tend to perform poorly. The same trend occurs in quantifier/qualifier entities, which may include a wide range of adjectives or expressions that other schemes represent with more specific classes: e.g. `Multiplier` [117] or `Modifier` [123]. In contrast, entities expressing results or values have high scores.

Limitations

Among our limitations, we only used texts about clinical trials to develop the tool. We trained new models with 100 anonymized clinical cases and tested them on 100 different cases, but we need to confirm the performance for other text genres, especially real EHRs. Moreover, our annotation scheme might be coarse for some tasks. The `DISO` category merged signs, symptoms and disorders; and `Dose` includes both dosage and concentration or strength. We largely depend on the UMLS categories, which might be inadequate for other tasks, and the tool is not compatible with other standards, e.g., OHDSI OMOP [119]. However, the tool can be used to pre-annotate texts using labels that can be adapted to specific tasks. Further improvement is required in the recognition of drug-related information; achieving this will necessitate additional annotated data. Discontinuous entities (i.e. separated by non-annotated tokens) were discarded to train the models: e.g., *vacuna anual contra la gripe* ('influenza virus [..] vaccine'). Furthermore, new architectures and language models need to be tested. Lastly, although the tool supports concept normalization to UMLS CUIs or SNOMED CT codes, it does not provide sense disambiguation. In the future, we plan to evaluate the normalization performance and will address relation extraction tasks by annotating the corpus with relations.

Usage

First, download the models from the Hugging Face hub. Then, import the `AutoModelForTokenClassification` class and load the NER model; for example, to annotate texts with the RoBERTa `bsc-bio-ehr-es` model for UMLS semantic groups, use: `model = "roberta-es-clinical-trials-umls-7sgs-ner"` (Figure C.18). Along with the graphical user interface (Fig. 1), the code can be executed via an UNIX-based terminal (Figure C.19) either for a single document or a batch of files included in a folder. A configuration file (Figure C.20) facilitates the specification of annotation options: e.g., use of the lexicon, temporal entities, negation or speculation, or output format of the annotations (`JSON` or BRAT `ann`). The code is available at: https://github.com/lcampillos/medspaner.

Conclusions

We have introduced a comprehensive NLP tool designed to automate the processing of clinical trials in Spanish and demonstrated its efficacy in extracting medical information from clinical cases. The Medical Semantic Python-Assisted Named Entity Recognizer (MEDSPANER) is open source, supports both lexicon- and Transformer-based annotation of medical entities, and also normalizes entities to UMLS Concept Unique Identifiers or SNOMED CT codes. The tool integrates Transformer-based NER models that can be adapted to other sub-genres (e.g., clinical reports or medical journal articles). In addition, the models for negation or speculation and temporal entity recognition can be re-adapted for non-medical domains. We conducted an internal validation, with F1 scores up to 0.915 (strict match). We also performed an external validation in which eight medical professionals evaluated the system annotations on 100 new clinical trials texts (average F1 = 0.858) and 100 anonymized clinical cases (with an average F1 = 0.910). To the best of our knowledge, this is one of the few tools for comprehensive processing of Spanish medical texts, including the processing of medication information, experiencer and event temporality. The tool can be adapted to other languages for which similar text data and resources exist. We make the tool available online and distribute the code. Because new language models will be released, which are expected to surpass the results presented herein, we created a space at the Hugging Face Hub to share these and future models. We also release in a companion repository the annotated corpus for training the models.

Appendix A. Training and annotation details

Table 6 Training epochs per model (average of five rounds and standard deviation)

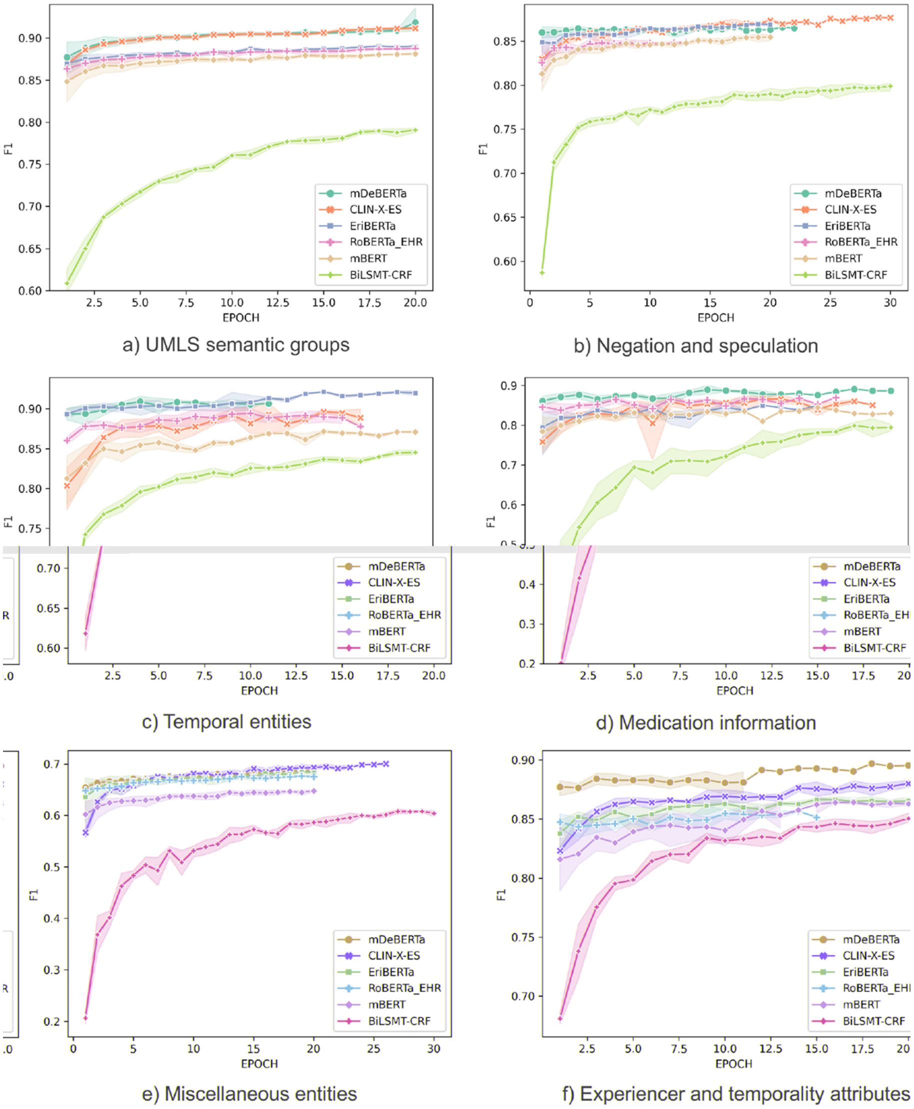
Model	UMLS entities	Medication data	Temporal entities	Negation/ Speculation	Miscellaneous entities	Experiencer/ Temporality
Bi-LSTM-CRF	95.60 (± 5.37)	68.40 (± 7.47)	68.20 (± 7.29)	80.60 (± 15.07)	78.80 (± 9.44)	71.20 (± 7.73)
RoBERTa EHR	17.00 (± 2.83)	14.20 (± 3.63)	14.00 (± 2.24)	10.80 (± 1.92)	16.80 (± 3.56)	10.80 (± 4.09)
EriBERTa	15.50 (± 4.12)	10.40 (± 3.78)	12.80 (± 6.72)	17.25 (± 5.50)	15.60 (± 6.07)	16.40 (± 3.36)
CLIN-X-ES	17.00 (± 2.83)	14.60 (± 3.65)	13.60 (± 4.22)	21.80 (± 5.36)	19.40 (± 6.07)	16.40 (± 4.83)
mBERT	14.75 (± 6.18)	11.40 (± 5.22)	14.00 (± 5.61)	16.00 (± 6.16)	18.20 (± 2.49)	11.60 (± 2.07)
mDeBERTa vs 3	17.80 (± 4.92)	11.60 (± 5.68)	9.80 (± 2.28)	15.00 (± 6.93)	18.00 (± 2.74)	11.60 (± 4.83)

Table 7 Annotation times (seconds per text) on 100 new trial texts and 100 clinical cases

Text type	Average bytes/text	Average tokens/text	RoBERTa EHR	EriBERTa	CLIN-X-ES	mDeBERTa vs 3
EudraCT	2074.92 (± 763.67)	298.51 (± 115.71)	24″	24″	96″	153″
Clinical cases	1945.18 (± 593.39)	278.31 (± 80.55)	28″	29″	111″	129″

For each model architecture, we timed the annotation process in which six models were applied (for UMLS entities, medication information, temporal entities, negation/speculation, miscellaneous entities and experiencer/event temporality attributes)

Appendix B. Detailed results and error analysis



a) UMLS semantic groups

b) Negation and speculation

c) Temporal entities

d) Medication information

e) Miscellaneous entities

f) Experiencer and temporality attributes

Fig. 8 F1 measure (averaged over 5 runs) over training iterations (the shaded area represents standard deviation)

UMLS entities

Scope errors (type 5) were the most frequent. These especially affected adjectives that were not annotated in the gold standard, or vice versa (e.g., *hemorragia digestiva alta*, 'upper gastric bleeding'; Fig. 10). They also occurred in proper names. Scope errors often appeared in compound entities, when the model predicted two separate entities (but one entity was marked the gold standard), or vice versa (e.g., 'recurrent VTE in pregnancy'). The scope caused many ambiguities to the human annotators.

Table 8 Results of the RoBERTa EHR model per entity type*

UMLS entities	**Precision**	**Recall**	**F1**	**Temporal entities**	**Precision**	**Recall**	**F1**
ANAT	0.728	0.686	0.706	Age	0.926	0.947	0.936
[308]	(± 0.030)	(± 0.030)	(± 0.025)	[372]	(± 0.013)	(± 0.009)	(± 0.010)
CHEM	0.917	0.923	0.920	Date	0.931	0.895	0.913
[2932]	(± 0.005)	(± 0.008)	(± 0.005)	[412]	(± 0.015)	(± 0.014)	(± 0.013)
DEVI	0.645	0.791	0.711	Duration	0.918	0.893	0.905
[134]	(± 0.018)	(± 0.047)	(± 0.027)	[629]	(± 0.014)	(± 0.019)	(± 0.010)
DISO	0.890	0.903	0.896	Frequency	0.780	0.885	0.829
[3065]	(± 0.008)	(± 0.003)	(± 0.003)	[73]	(± 0.043)	(± 0.008)	(± 0.024)
LIVB	0.949	0.959	0.954	Time	0.722	0.809	0.762
[1685]	(± 0.004)	(± 0.006)	(± 0.002)	[113]	(± 0.068)	(± 0.042)	(± 0.052)
PHYS	0.766	0.765	0.765	Miscellaneous	Precision	Recall	F1
[308]	(± 0.021)	(± 0.012)	(± 0.008)				
PROC	0.842	0.871	0.856	Concept	0.644	0.612	0.627
[4154]	(± 0.002)	(± 0.004)	(± 0.002)	[764]	(± 0.016)	(± 0.019)	(± 0.009)
Medicationdata	Precision	Recall	F1	Food/Drink	0.692	0.733	0.712
				[27]	(± 0.049)	(± 0.071)	(± 0.058)
Contraindicated.	0.818	0.816	0.812	Observation	0.626	0.617	0.621
[76]	(± 0.047)	(± 0.104)	(± 0.049)	[822]	(± 0.015)	(± 0.010)	(± 0.010)
Dose	0.824	0.843	0.833	Quant./Qual.	0.700	0.661	0.680
[314]	(± 0.039)	(± 0.020)	(± 0.021)	[1202]	(± 0.015)	(± 0.020)	(± 0.008)
Form	0.932	0.884	0.907	Result/Value	0.828	0.910	0.867
[74]	(± 0.030)	(± 0.020)	(± 0.019)	[394]	(± 0.013)	(± 0.005)	(± 0.007)
Route	0.916	0.899	0.907	Experiencer/temp. attr.	Precision	Recall	F1
[288]	(± 0.037)	(± 0.022)	(± 0.017)				
Negation/Speculation	Precision	Recall	F1	Family_member	0.721	0.920	0.808
				[20]	(± 0.048)	(± 0.027)	(± 0.034)
Neg_cue	0.955	0.958	0.957	Other	0.852	0.805	0.828
[2484]	(± 0.006)	(± 0.006)	(± 0.005)	[120]	(± 0.019)	(± 0.015)	(± 0.011)
Negated	0.829	0.837	0.833	Patient	0.949	0.921	0.935
[3160]	(± 0.005)	(± 0.014)	(± 0.008)	[1462]	(± 0.003)	(± 0.005)	(± 0.003)
Spec_cue	0.834	0.859	0.846	Future	0.640	0.620	0.629
[756]	(± 0.021)	(± 0.017)	(± 0.007)	[70]	(± 0.040)	(± 0.059)	(± 0.045)
Speculated	0.708	0.719	0.713	History_of	0.742	0.667	0.703
[1008]	(± 0.019)	(± 0.016)	(± 0.016)	[647]	(± 0.021)	(± 0.016)	(± 0.010)

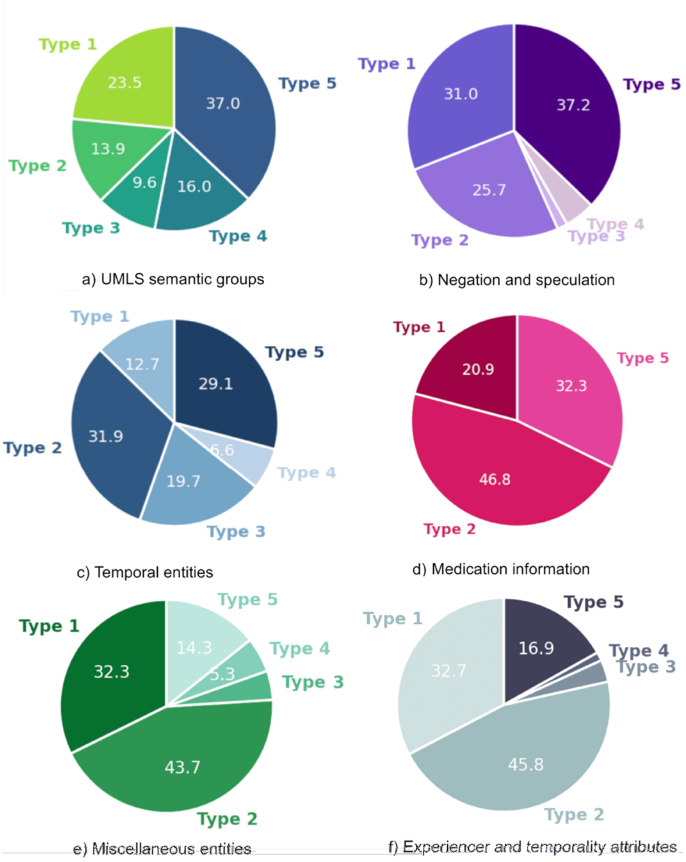
* The number of instances of each type in the test set (support) appears in square brackets (below each entity name in the first column); *temp. att.* stands for 'temporality attributes'

False positives (FPs, type 1 errors) abounded. These were related to polysemous entities (e.g., *brazo*, 'arm', was labeled as `ANAT` when it referred to each trial cohort). Some proper names and abbreviations were mislabeled; e.g., `CHEM` in *AN69ST*®, or *FCI* ('informed consent document') misinterpreted as `DEVI`. Most false negatives (FNs, type 2 errors) were owing to the lack of enough instances in the training corpus, namely anatomical entities, physiology terms and medical devices. Some infrequent abbreviations were not recognized (e.g., *GE*, 'experimental group', `LIVB`). Wrong labels (type 3 and 4 errors) were due to ambiguous entities: e.g., *muestras* can refer to the anatomic samples (`ANAT`) or to the sampling procedure (`PROC`).

Medication information entities

Many errors were FPs (type 1). In particular, the model incorrectly labeled as `Dose` some number expressions and values of lab results (e.g., *blood concentrations of tacrolimus of 5-15 μ g/L*). This mostly affected the eligibility criteria of trial announcements and was the main cause of the model's decreased performance. `Route` FPs involved words expressing spatial concepts related to procedures. Semantically-related terms expressing the location of a drug agent also caused FPs (e.g., *oral corticosteroids with topical action*).

However, the most frequent error were FNs (type 2). These mostly affected `Form` entities, since some dosage form terms were infrequent in the training data (e.g., *film, gel, mouthwashes*). Likewise, the model's performance was not high on the `Contraindicated` class, given the scarce number of entities in the corpus. Errors in this category caused both FNs and FPs (e.g., the model also annotated `Contraindicated` in some diseases, whereas our annotation scheme only focused on medications or procedures).



a) UMLS semantic groups

b) Negation and speculation

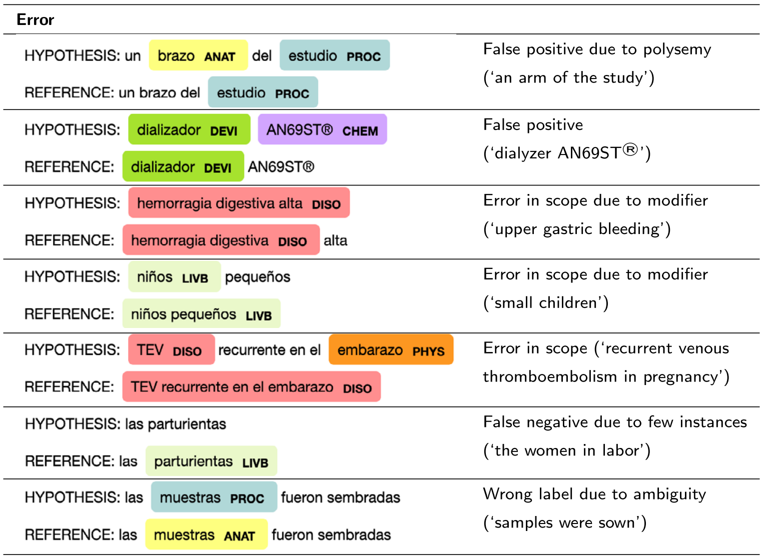
c) Temporal entities

d) Medication information

e) Miscellaneous entities

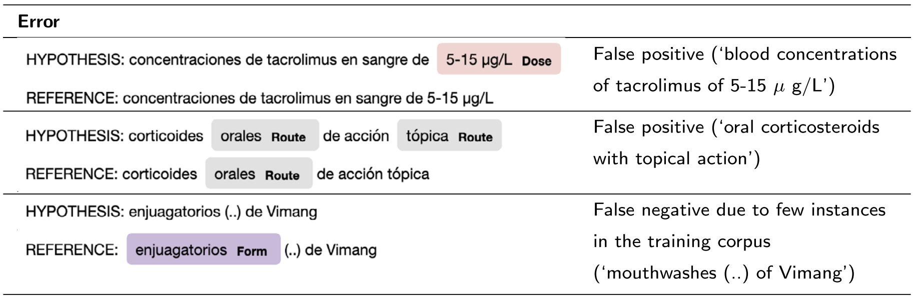
f) Experiencer and temporality attributes

Fig. 9 Errors (%) with RoBERTa EHR models classified according to [111]. Type 1: false positives; Type 2: false negatives; Type 3: same span, label mismatch; Type 4: overlapping span, wrong label; Type 5: overlapping span, right label



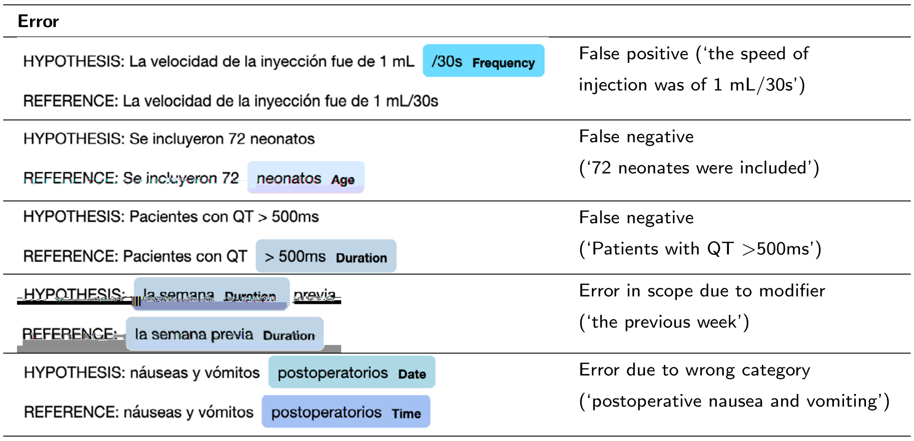
Error	
HYPOTHESIS: un brazo ANAT del estudio PROC REFERENCE: un brazo del estudio PROC	False positive due to polysemy ('an arm of the study')
HYPOTHESIS: dializador DEVI AN69ST® CHEM REFERENCE: dializador DEVI AN69ST®	False positive ('dialyzer AN69ST®')
HYPOTHESIS: hemorragia digestiva alta DISO REFERENCE: hemorragia digestiva DISO alta	Error in scope due to modifier ('upper gastric bleeding')
HYPOTHESIS: niños LIVB pequeños REFERENCE: niños pequeños LIVB	Error in scope due to modifier ('small children')
HYPOTHESIS: TEV DISO recurrente en el embarazo PHYS REFERENCE: TEV recurrente en el embarazo DISO	Error in scope ('recurrent venous thromboembolism in pregnancy')
HYPOTHESIS: las parturientas REFERENCE: las parturientas LIVB	False negative due to few instances ('the women in labor')
HYPOTHESIS: las muestras PROC fueron sembradas REFERENCE: las muestras ANAT fueron sembradas	Wrong label due to ambiguity ('samples were sown')

Fig. 10 Samples of NER errors of UMLS semantic groups (translation in brackets)



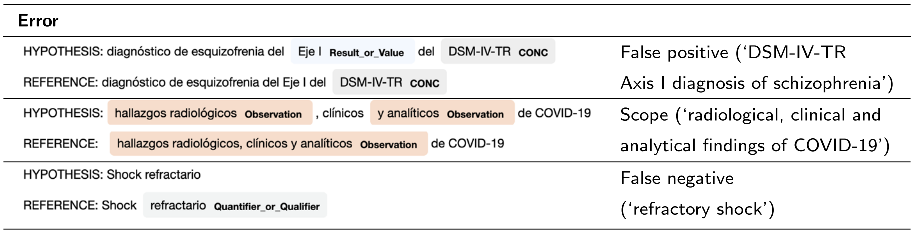
Error	
HYPOTHESIS: concentraciones de tacrolimus en sangre de 5-15 µg/L Dose REFERENCE: concentraciones de tacrolimus en sangre de 5-15 µg/L	False positive ('blood concentrations of tacrolimus of 5-15 μ g/L')
HYPOTHESIS: corticoides orales Route de acción tópica Route REFERENCE: corticoides orales Route de acción tópica	False positive ('oral corticosteroids with topical action')
HYPOTHESIS: enjuagatorios (..) de Vimang REFERENCE: enjuagatorios Form (..) de Vimang	False negative due to few instances in the training corpus ('mouthwashes (..) of Vimang')

Fig. 11 Samples of NER errors of medication information (translation in brackets)



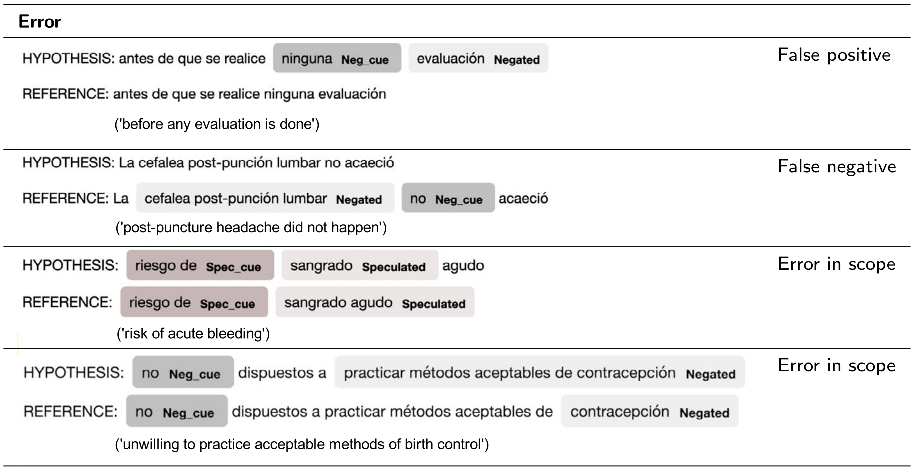
Error	
HYPOTHESIS: La velocidad de la inyección fue de 1 mL /30s Frequency REFERENCE: La velocidad de la inyección fue de 1 mL/30s	False positive ('the speed of injection was of 1 mL/30s')
HYPOTHESIS: Se incluyeron 72 neonatos REFERENCE: Se incluyeron 72 neonatos Age	False negative ('72 neonates were included')
HYPOTHESIS: Pacientes con QT > 500ms REFERENCE: Pacientes con QT > 500ms Duration	False negative ('Patients with QT >500ms')
HYPOTHESIS: la semana Duration previa REFERENCE: la semana previa Duration	Error in scope due to modifier ('the previous week')
HYPOTHESIS: náuseas y vómitos postoperatorios Date REFERENCE: náuseas y vómitos postoperatorios Time	Error due to wrong category ('postoperative nausea and vomiting')

Fig. 12 Samples of NER errors of temporal entities (translation in brackets)



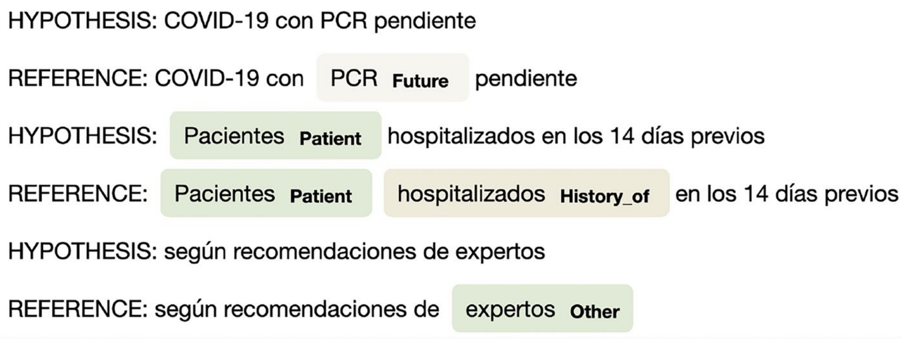
Error	
HYPOTHESIS: diagnóstico de esquizofrenia del Eje I Result_or_Value del DSM-IV-TR CONC REFERENCE: diagnóstico de esquizofrenia del Eje I del DSM-IV-TR CONC	False positive ('DSM-IV-TR Axis I diagnosis of schizophrenia')
HYPOTHESIS: hallazgos radiológicos Observation , clínicos y analíticos Observation de COVID-19 REFERENCE: hallazgos radiológicos, clínicos y analíticos Observation de COVID-19	Scope ('radiological, clinical and analytical findings of COVID-19')
HYPOTHESIS: Shock refractario REFERENCE: Shock refractario Quantifier_or_Qualifier	False negative ('refractory shock')

Fig. 13 Samples of NER errors of miscellaneous entities (translation in brackets)



Error	
HYPOTHESIS: antes de que se realice ninguna Neg_cue evaluación Negated REFERENCE: antes de que se realice ninguna evaluación ('before any evaluation is done')	False positive
HYPOTHESIS: La cefalea post-punción lumbar no acaeció REFERENCE: La cefalea post-punción lumbar Negated no Neg_cue acaeció ('post-puncture headache did not happen')	False negative
HYPOTHESIS: riesgo de Spec_cue sangrado Speculated agudo REFERENCE: riesgo de Spec_cue sangrado agudo Speculated ('risk of acute bleeding')	Error in scope
HYPOTHESIS: no Neg_cue dispuestos a practicar métodos aceptables de contracepción Negated REFERENCE: no Neg_cue dispuestos a practicar métodos aceptables de contracepción Negated ('unwilling to practice acceptable methods of birth control')	Error in scope

Fig. 14 Samples of NER errors of negation and speculation (translation in brackets)



HYPOTHESIS: COVID-19 con PCR pendiente

REFERENCE: COVID-19 con PCR Future pendiente

HYPOTHESIS: Pacientes Patient hospitalizados en los 14 días previos

REFERENCE: Pacientes Patient hospitalizados History_of en los 14 días previos

HYPOTHESIS: según recomendaciones de expertos

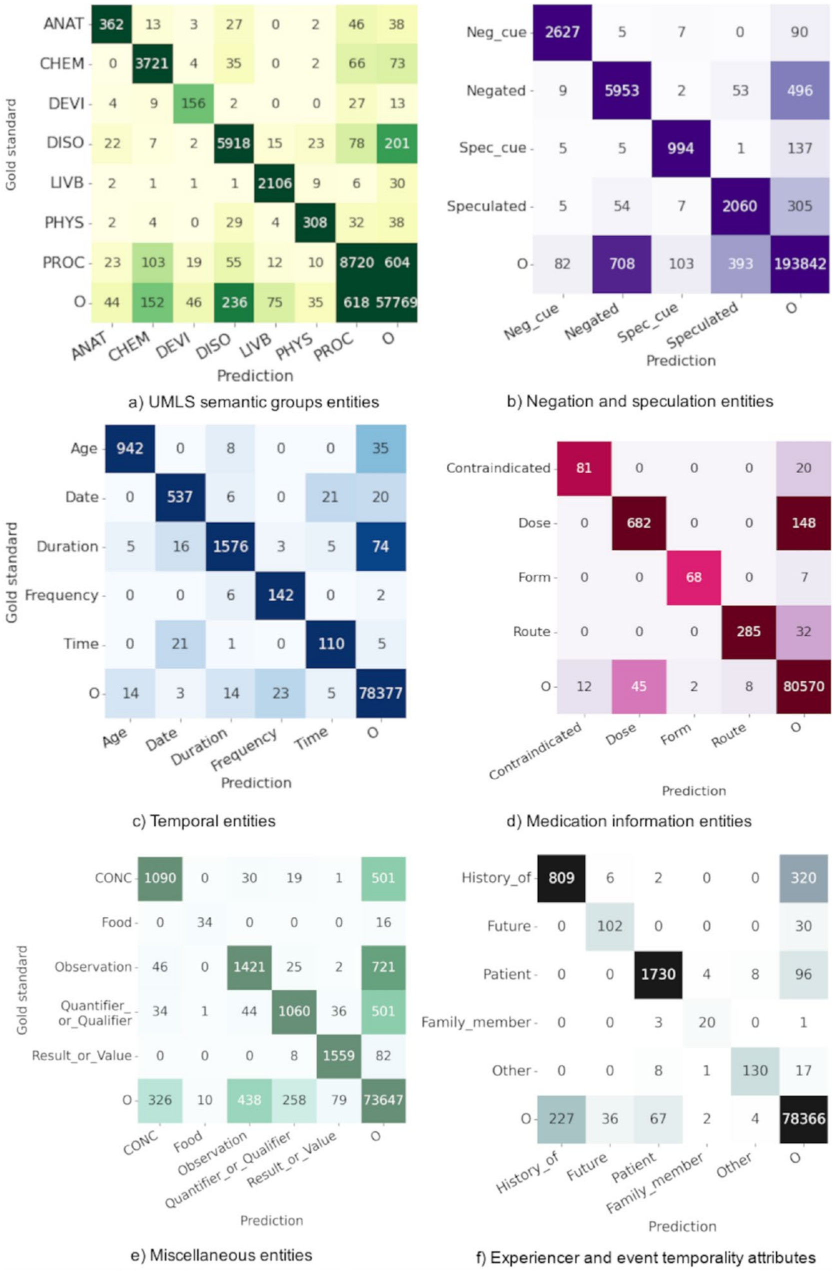
REFERENCE: según recomendaciones de expertos Other

Fig. 15 Samples of NER errors of event temporality and experiencer attributes (translation in brackets)

In most contexts, this was not a serious error. Scope errors (type 5) were very abundant and mostly affected quantifiers in `Dose` entities (e.g., *superior a 10 mg*, '<u>above</u> 10 mg'). Mismatch errors (type 3 and 4) were sporadic (Fig. 11).

Temporal entities

Most errors were FNs (type 2), namely in entities with low frequency in our training data (e.g., *pre-inclusion, QT > 500ms*; Fig. 12). Also, the model did not annotate `Age` in some `LIVB` entities expressing age groups (e.g., *neonate*). The second most common error affected the scope (type 5), namely when the model predicted a shorter span of the full entity (e.g., *la semana previa*, 'the previous week').



	ANAT	CHEM	DEVI	DISO	LIVB	PHYS	PROC	O
ANAT	362	13	3	27	0	2	46	38
CHEM	0	3721	4	35	0	2	66	73
DEVI	4	9	156	2	0	0	27	13
DISO	22	7	2	5918	15	23	78	201
LIVB	2	1	1	1	2106	9	6	30
PHYS	2	4	0	29	4	308	32	38
PROC	23	103	19	55	12	10	8720	604
O	44	152	46	236	75	35	618	57769

a) UMLS semantic groups entities

	Neg_cue	Negated	Spec_cue	Speculated	O
Neg_cue	2627	5	7	0	90
Negated	9	5953	2	53	496
Spec_cue	5	5	994	1	137
Speculated	5	54	7	2060	305
O	82	708	103	393	193842

b) Negation and speculation entities

	Age	Date	Duration	Frequency	Time	O
Age	942	0	8	0	0	35
Date	0	537	6	0	21	20
Duration	5	16	1576	3	5	74
Frequency	0	0	6	142	0	2
Time	0	21	1	0	110	5
O	14	3	14	23	5	78377

c) Temporal entities

	Contraindicated	Dose	Form	Route	O
Contraindicated	81	0	0	0	20
Dose	0	682	0	0	148
Form	0	0	68	0	7
Route	0	0	0	285	32
O	12	45	2	8	80570

d) Medication information entities

	CONC	Food	Observation	Quantifier_or_Qualifier	Result_or_Value	O
CONC	1090	0	30	19	1	501
Food	0	34	0	0	0	16
Observation	46	0	1421	25	2	721
Quantifier_ or_Qualifier	34	1	44	1060	36	501
Result_or_Value	0	0	0	8	1559	82
O	326	10	438	258	79	73647

e) Miscellaneous entities

	History_of	Future	Patient	Family_member	Other	O
History_of	809	6	2	0	0	320
Future	0	102	0	0	0	30
Patient	0	0	1730	4	8	96
Family_member	0	0	3	20	0	1
Other	0	0	8	1	130	17
O	227	36	67	2	4	78366

f) Experiencer and event temporality attributes

Fig. 16 Confusion matrices of entity sub-classes, obtained with RoBERTa EHR Transformer-based model

Numerous errors were due to label mismatch (type 3), especially in `Date` and `Time` entities with *pre- / post-* affixes. Their meaning depends on the context: e.g., *postoperatorio* tends to be `Date` (as in *dolor postoperatorio*, 'postoperative pain'), but it is `Time` when it refers to a period shorter than 24 h. Another confusion affected `Date` and `Duration`, which can be interpreted as a point in time, an interval or a temporal container (e.g. *the fifth week, over the past few years*). Those entity types were ambiguous for the human annotators (the IAA score of `Time` was low).

FPs (type 1) occurred in `LIVB` entities mislabeled as `Age` (e.g., *student*). Some numbers in protocol identifiers were mistaken as `Date` (e.g., *NOPHO-DBH AML 2012 protocol*). Other errors were due to tokenization (e.g., *min-1* in `Frequency`) or inconsistencies in the gold standard.

Miscellaneous entities

FNs (type 2) were the most recurrent errors, and tended to appear in `Food_or_Drink`, `Observation` or `Quantifier_or_Qualifier` entities (e.g. *refractory shock*; Fig. 13). These entities are highly variable and some instances in the test set might not occur in the training data. FPs (type 1) are the second most prevalent errors, and affected `Result_or_Value` entities very often. Especially, when an expression contains numbers or typographical characters used in measurements: e.g. *DSM-IV-TR Axis I diagnosis*. Finally, scope errors (type 5) were common in `Observation` entities. Often, the model did not label some parts of the entity, or tagged discontinuous expressions (while one entity was tagged in the gold standard): e.g. *radiological, clinical and analytical findings*. This can be related to the high variability of word forms and also to annotation inconsistencies.

Negation and speculation

The main source of errors was related to scope (type 5). These occurred when the model did not label modifiers (or vice versa): e.g., *riesgo de sangrado agudo* ('risk of acute bleeding'). At times, the model did not label the negation or speculation cue when postponed after the focus event (e.g., *post-puncture headache did not happen*).

FPs (type 1) also affected the scope. In some contexts, the model annotated the full scope, but only the focus of negation or speculation was marked in the reference (e.g., *unwilling to practice acceptable methods of birth control*). Other FPs were due to the modifier *negative*, in cases such as *Serological tests negative against syphilis*. According to our criteria, we only labeled the negated entity (*syphilis*), since the focus of the negation is the test result (Fig. 14).

Another cause of FPs were negated entities in a wider negated context. In the gold standard, we did not annotate them since negating a negated entity makes a positive statement: e.g., in *patients without any hearing impairment except age-related pathology*, only *hearing impairment* was marked as negated in the gold standard. These were ambiguous to human annotators, but we tried to preserve the semantic coherence of the eligibility criteria.

Contexts where medical knowledge is needed caused errors. In *Forms of psoriasis other than chronic plaque-type (e.g., pustular, erythrodermic and guttate psoriasis)*, the model labeled the entities between brackets as negated. However, these are not hyponyms nor co-referent of *plaque psoriasis*, they should not be negated. Stylistic usage of negation was problematic; the model mislabeled contexts where no event was in fact negated: e.g., *antes de que se realice ninguna evaluación* ('before any evaluation is done').

FNs (type 2) were the third main cause of error. These were mainly due to the lack of training instances. Lexical negation caused many errors since the model did not recognize verbs or nouns negating an event. Compared to syntactic negation cues, lexical cues vary to a greater degree: e.g., *cesar* ('to quit'), *descartar* ('to rule out'), *retirada de* ('withdrawal of'), *supresión* ('elimination'), etc. Indeed, the human annotators hesitated about whether they should mark them or not. The larger number of errors in speculation cues can be explained because more variants exist in markers (compared to negation cues; e.g., *dudosamente*, 'doubtfully'; *potencial*, 'potential'), and in completive clauses with speculation verbs (e.g., *creemos que*, 'we believe that'; *evaluar si*, 'to evaluate whether'). This variance also affected verbs in conditional mood, which can also express uncertainty (e.g., *habría que descartar*, 'should be discarded').

Event temporality and experiencer attributes

FNs (type 2) represented almost half of the errors, and these tend to occur in attributes with few instances in the test set: e.g. `Future` is missing in *COVID-19 with pending PCR*. FNs also occurred in entities with higher variability. For example, the `Other` person attribute might be a medical professional, a legal representative or a researcher: e.g. it is missing in *according to experts recommendations*. The `History_of` attribute was missing in contexts without the standard cue words, or if these appeared far from the focus word in the sentence.

FPs (type 1) were the second most common error type. They often involved temporality attributes, when the model predicted `Future` or `History_of` in a present condition. Lastly, the experiencer attribute was also mismatched in contexts where a medical professional was the participant in a clinical trial (Fig. 15).

Appendix C. Functioning and usage of the tool

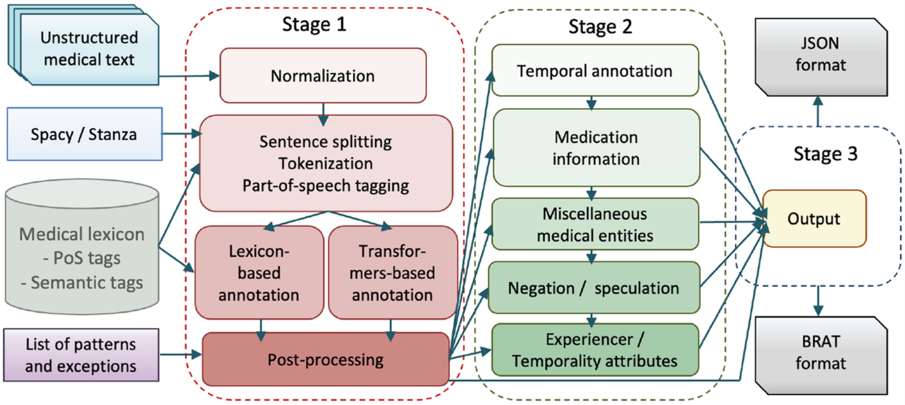
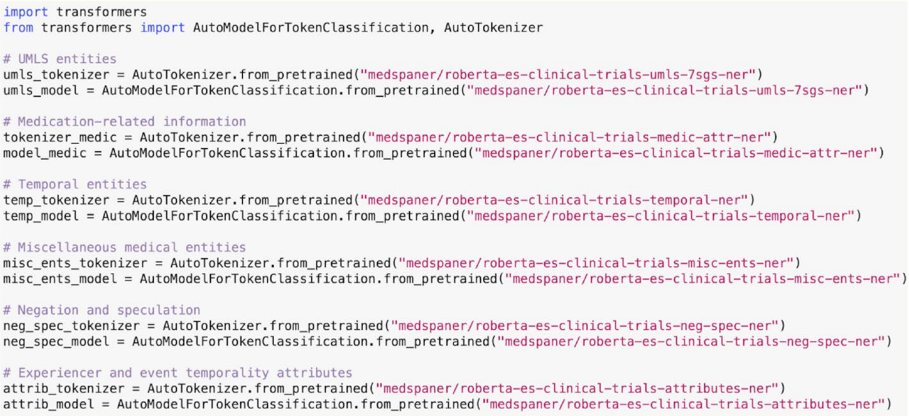
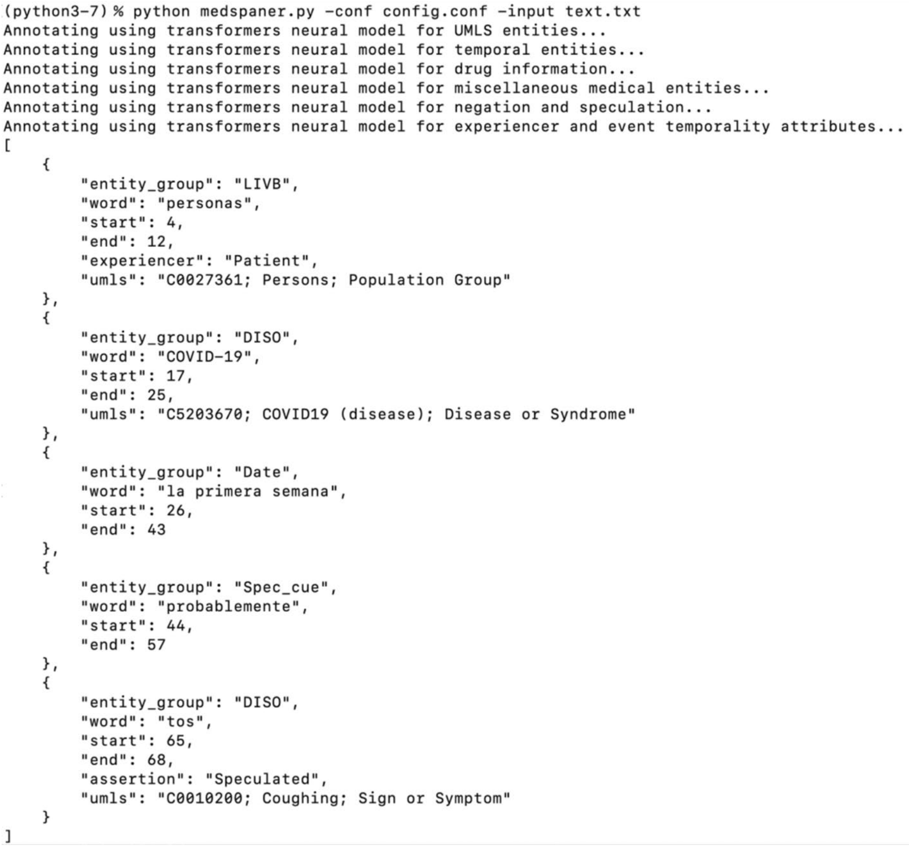


Fig. 17 Processing pipeline



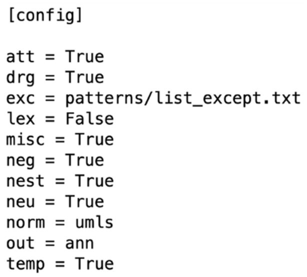
```
import transformers
from transformers import AutoModelForTokenClassification, AutoTokenizer

# UMLS entities
umls_tokenizer = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-umls-7sgs-ner")
umls_model = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-umls-7sgs-ner")

# Medication-related information
tokenizer_medic = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-medic-attr-ner")
model_medic = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-medic-attr-ner")

# Temporal entities
temp_tokenizer = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-temporal-ner")
temp_model = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-temporal-ner")

# Miscellaneous medical entities
misc_ents_tokenizer = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-misc-ents-ner")
misc_ents_model = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-misc-ents-ner")

# Negation and speculation
neg_spec_tokenizer = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-neg-spec-ner")
neg_spec_model = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-neg-spec-ner")

# Experiencer and event temporality attributes
attrib_tokenizer = AutoTokenizer.from_pretrained("medspaner/roberta-es-clinical-trials-attributes-ner")
attrib_model = AutoModelForTokenClassification.from_pretrained("medspaner/roberta-es-clinical-trials-attributes-ner")
```

Fig. 18 Sample code to import the Transformer-based models



```
(python3-7) % python medspaner.py -conf config.conf -input text.txt
Annotating using transformers neural model for UMLS entities...
Annotating using transformers neural model for temporal entities...
Annotating using transformers neural model for drug information...
Annotating using transformers neural model for miscellaneous medical entities...
Annotating using transformers neural model for negation and speculation...
Annotating using transformers neural model for experiencer and event temporality attributes...
[
    {
        "entity_group": "LIVB",
        "word": "personas",
        "start": 4,
        "end": 12,
        "experiencer": "Patient",
        "umls": "C0027361; Persons; Population Group"
    },
    {
        "entity_group": "DISO",
        "word": "COVID-19",
        "start": 17,
        "end": 25,
        "umls": "C5203670; COVID19 (disease); Disease or Syndrome"
    },
    {
        "entity_group": "Date",
        "word": "la primera semana",
        "start": 26,
        "end": 43
    },
    {
        "entity_group": "Spec_cue",
        "word": "probablemente",
        "start": 44,
        "end": 57
    },
    {
        "entity_group": "DISO",
        "word": "tos",
        "start": 65,
        "end": 68,
        "assertion": "Speculated",
        "umls": "C0010200; Coughing; Sign or Symptom"
    }
]
```

Fig. 19 Annotation output from the UNIX-based terminal



```
[config]

att = True
drg = True
exc = patterns/list_except.txt
lex = False
misc = True
neg = True
nest = True
neu = True
norm = umls
out = ann
temp = True
```

Fig. 20 Configuration file with annotation options*. *`att`: annotate experiencer/temporality attributes; `drg`: annotate drug information; `exc`: use list of excluded words; `lex`: use lexicon; `misc`: annotate miscellaneous entities; `neg`: detect negation/speculation; `nest`: output nested entities; `neu`: UMLS neural models; `norm`: `umls` (if normalization to UMLS CUIs) or `snomed` (if normalization to SNOMED CT codes); `out`: output format (`ann` for BRAT, or `json`); `temp`: annotate temporal entities.

Appendix D: Comparison of system results with similar works

Table 9 Comparison of F-score in similar tasks (corpora in brackets); results from [22] referenced from [114]

	UMLS & medical entities	**Negation/speculation**
[4]	0.715	0.943 (negation), 0.859 (status)
[22]	0.70 [i2b2]	0.63 (negation)
[5]	0.795	0.984 (negation)
[124]	0.855 [i2b2]	0.905 [i2b2] (negation)
	0.874 [In-house corpus]	0.899 [In-house corpus] (negation)
[61]		0.955 (neg. cue), 0.89 (neg. scope)
[NUBEs]		0.829 (spec. cue), 0.746 (spec. scope)
[62]		0.85 (negation)
[95]	0.798 [Spanish CWLC]	
[63]		0.963 & 0.975 (cue, French)
		0.765 & 0.880 (scope, French)
[116]	0.658 [Chia] & 0.785 [FRD]	
[64]		0.95 (neg. cue), 0.92 (neg. scope)
[NUBEs]		0.84 (spec. cue), 0.80 (spec. scope)
[64]		0.90 (neg. cue), 0.84 (neg. scope)
[In-house cancer corpus]		0.81 (spec. cue), 0.74 (spec. scope)
[65]		0.786 (negation, best model)
[DIANN corpus]		0.765 (negation, authors' model)
[125]	0.660 (disorders)	
[E3C corpus, Spanish]		
	Temporal entities	Medication information
[55]	0.726 [THYME] & 0.762 [i2b2]	
[120]	0.838 [i2b2]	
[82]	0.889 [Spanish BARR]	
[121]	0.824 (scope), 0.783 (class) [HourGlass]	
	0.851 (scope), 0.831 (class) [TempEval2]	
[83]	0.761 (strict), 0.912 (relaxed) [E3C]	

Table 9 (continued)

	UMLS & medical entities	Negation/speculation
[126]	0.986 (Age), 0.934 (Frequency), 0.991 (Date)	0.970 (Dose)
[In-house cancer corpus]	0.898 (Duration), 0.933 (Implicit date)	
[122]		0.91–0.95
[66]		0.899
[67]		0.921 [n2c2 task 2]
	Miscellaneous entities	Experiencer/event temp.
[60]		0.760 (Historical);
[inhouse corpus]		1.000 (Experiencer)
[77]	0.800 (Finding) [SpRadIE corpus]	
[95]	0.622 (Finding)	0.918 (Family_member) [CWLC]
[117]	0.267 (Observation); 0.462 (Qualifier);	
	0.735 (Value) [Chia]	
[123]	0.747 (Observation); 0.618 (Modifier)	0.416 (Family_member) [LEAF]
	0.967 (Value) [Chia]	

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	BidirectionalLong-Short Term Memory
BIO	Begin, Inside, Out
CRF	Conditional Random Fields
CT	Clinical trials
CUI	Concept Unique Identifier
EBM	Evidence-Based Medicine
EHR	Electronic health record
EudraCT	European Clinical Trials Register
FN	False negative
FP	False positive
GPU	Graphical processing unit
H	Hidden state
I	Input
IAA	Inter-annotator agreement
K	Key
LLM	Large language model
M	Mean
NER	Named entity recognition
NLP	Natural language processing
OOV	Out-of-vocabulary
P	Precision
Q	Query
R	Recall
REEC	Repositorio Español de Estudios Clínicos
SciELO	Scientific Electronic Library Online
SD	Standard deviation
SG	Semantic group
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
TP	True positive
V	Value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05949-6.

Supplementary Material 1.

Acknowledgements

We are really thankful to Dr. Antonio Moreno-Sandoval for giving us access to the computer facilities of the Computational Linguistics Laboratory (Autonomous University of Madrid) and Dr. Jordi Porta for his help with the GPU. We also thank the Spanish Medicines Agency (AEMPS) and Dr. Jesús Santamaría for their support with the first stage of the human evaluation; Dr. Jonathan Heras for his valuable comments; and SAMFyC, who allowed us using their published clinical cases. We greatly thank the medical professionals who evaluated the texts and provided us with feedback to improve the system: Beatriz Arribas, Jorge Collado, Roberto Crespo, Daniel Duménigo, Hermenegildo Martínez, Rosa López, Noelia Martín and Javier Pumares. We also gratefully thank the creators of the NUBEs corpus and the pre-trained models we integrated in the tool. Lastly, we deeply thank the anonymous reviewers for their valuable comments to improve this work and the final version of the manuscript.

Availability and requirements

Project name: MEDSPANER. Code page: https://github.com/lcampillos/medspaner. Demo page: https://claramed.csic.es/medspaner . Trained models available at: https://huggingface.co/medspaner . Operating system(s): UNIX-based (including Mac OS). Programming language: Python vs 3.7 or higher. Other requirements: spaCy (vs. 3.3.1), Transformers (vs. 4.16.2); the full list of requirements is provided in the companion repository. License: General Public License 3 (AGPL—3.0). Any restrictions to use: MEDSPANER uses the MedLexSp lexicon, which is distributed upon signing license available at: https://digital.csic.es/handle/10261/270429.

Author Contributions

LCL conceptualized the annotation scheme, collected the texts, annotated data, analyzed the results, conducted the experiments, programmed the software and prepared the manuscript. AVM contributed to the creation of the annotation scheme, annotation criteria and guidelines, doubly annotated some sets and reviewed the manuscript. ACC helped to create the annotation scheme and guidelines according to his medical knowledge, helped to engage the health professionals who participated in the external evaluation, doubly annotated some sets and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The first version of MEDSPANER was developed under the Marie Curie COFUND grant agreement no. 713366 (InterTalentum UAM). The training of the Transformer-based models and their integration into the new version was conducted under the CLARA-MeD project (PID2020-116001RA-C33), funded by MICIU/AEI/10.13039/501100011033/, in project call: "Proyectos I+D+i Retos de investigación". The funding bodies did not take part in the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

Data Availibility Statement

All the resources supporting this article are available at the companion repository: https://github.com/lcampillos/medspaner. A demonstration system of the web interface is available at: https://claramed.csic.es/medspaner . The annotated corpus to train the models can be downloaded at: https://zenodo.org/records/13880599. The fine-tuned models are available at: https://huggingface.co/medspaner. .

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no Conflict of interest.

Received: 28 February 2024 Accepted: 30 September 2024
Published online: 08 January 2025

References

1. Data MC. Secondary Analysis of Electronic Health Records. Springer Nature; 2016.
2. Chen X, Xie H, Cheng G, Poon LK, Leng M, Wang FL. Trends and features of the applications of natural language processing techniques for clinical trials text analysis. Appl Sci. 2020;10(6):2157.
3. Xu H, Demner-Fushman D. Natural Language Processing in Biomedicine. Cham: Springer; 2024.
4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA. 2010;17(5):507–13.
5. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. JAMIA. 2019;26(4):294–305.
6. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol. 2015;33(4):342–5.
7. Álvarez-Pérez A, Prieto-Santamaría L, Ugarte-Carro E, Otero-Carrasco B, Ayuso-Muñoz A, Rodríguez-González A. Exploring disease-drug pairs in Clinical Trials information for personalized drug repurposing. Proc of 26th IEEE Computer-based Med Syst Conf. 2023;p. 2023–05.
8. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med. 2023;155: 106649.
9. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. Mol Case Stud. 2015;1(1): a000372.

10. Gupta NS, Kumar P. Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. Comput Biol Med. 2023;162: 107051.
11. Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Sci Data. 2014;1(1):1–9.
12. Zhou M, Zheng C, Xu R. Combining phenome-driven drug-target interaction prediction with patientsâ€™ electronic health records-based clinical corroboration toward drug discovery. Bioinform. 2020;36:i436–44.
13. Noori A, Magdamo C, Liu X, Tyagi T, Li Z, Kondepudi A, et al. Development and evaluation of a natural language processing annotation tool to facilitate phenotyping of cognitive status in electronic health records: diagnostic study. JMIR Med Inform. 2022;24(8): e40384.
14. Ashburner JM, Chang Y, Wang X, Khurshid S, Anderson CD, Dahal K, et al. Natural language processing to improve prediction of incident atrial fibrillation using electronic health records. J Am Heart Assoc. 2022;11(15): e026014.
15. Zheng C, Ms Lee, Bansal N, Go AS, Chen C, Harrison TN, et al. Identification of recurrent atrial fibrillation using natural language processing applied to electronic health records. Eur Heart J. 2023;10(1):77–88.
16. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, et al. Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. Schizophr Bull. 2021;47(2):405–14.
17. Fu S, Lopes GS, Pagali SR, Thorsteinsdottir B, LeBrasseur NK, Wen A, et al. Ascertainment of delirium status using natural language processing from electronic health records. J Gerontol. 2022;77(3):524–30.
18. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf. 2017;40:1075–89.
19. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. 2015;15:1–10.
20. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS ONE. 2014;9(11): e112774.
21. Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. J Biomed Inform. 2022;126: 103980.
22. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP-A toolkit for efficiently building customized clinical natural language processing pipelines. JAMIA. 2018;25(3):331–6.
23. Espinosa L, Tello J, Pardo A, Hernández-Medrano I, Ureña A, Salcedo I, et al. SAVANA: a global information extraction and terminology expansion framework in the medical domain. Proces Leng Nat. 2016;57:23–30.
24. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70.
25. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279.
26. Campillos-Llanos L. MedLexSp - A medical lexicon for Spanish medical natural language processing. J Biomed Semant. 2023;14(1):2.
27. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: Proc. of 13th EACL Conf. - Demonstrations; 2012. p. 102–107.
28. Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A, Moreno-Sandoval A. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. BMC Med Inform Decis Mak. 2021;21(1):1–19.
29. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python natural language processing toolkit for many human languages. Proc of 58th ACL. 2020;p. 101–108.
30. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2017. To appear. Available from: https://spacy.io/.
31. Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, et al. AllenNLP: A Deep Semantic Natural Language Processing Platform. In: Proc. of Workshop for NLP Open Source Software. Melbourne, Australia; 2018. p. 1–6.
32. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Proc. of 2020 EMNLP; 2020. p. 38–45.
33. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. of 2019 NAACL Conf., vol. 1. Minneapolis, MN, USA, 2-7 June; 2019. p. 4171–86.
34. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. Proc of 27th Int Conf Comp Ling Santa Fe, NM, USA, 20-26 Aug. 2018;p. 1638–49.
35. Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, SIGIR; 2016. p. 1–4.
36. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. In: Proc. AMIA Symp.. vol. 2021. American Medical Informatics Association; 2021. p. 438.
37. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artif Intell Med. 2021;117: 102083.
38. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. JAMIA. 2021;28(9):1892–9.
39. Li I, You K, Tang X, Qiao Y, Huang L, Hsieh CC, et al. EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts. arXiv preprint arXiv:2204.06604. 2022;.
40. Krishnamoorthy S, Jiang Y, Buchanan W, Singh A, Ortega J. CLPT: A Universal annotation scheme and toolkit for clinical language processing. In: Proc. 4th Clinical NLP Workshop; 2022. p. 1–9.
41. Neuraz A, Vaillant G, Arias C, Birot O, Huynh KT, Fabacher T, et al. Facilitating phenotyping from clinical texts: the medkit library. arXiv preprint arXiv:2409.00164. 2024;.
42. Friedman C. A broad-coverage natural language processing system. In: Proc. AMIA Symp. American Medical Informatics Association; 2000. p. 270.
43. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010;17(3):229–36.

44. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. JAMIA. 2017;24(4):841–4.
45. Cook HV, Pafilis E, Jensen LJ. A dictionary-and rule-based system for identification of bacteria and habitats in text. In: Proc. of 15th BioNLP; 2016. p. 50–55.
46. Cook HV, Bērziņš R, Leal C, Cejuela JM, Jensen LJ. Creation and evaluation of a dictionary-based tagger for virus species and proteins. In: Proc. of. 16th BioNLP; 2017. p. 91–98.
47. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. 2020;36(4):1234–40.
48. Rojas M, Dunstan J, Villena F. Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing. In: Proc. of 4th Clinical NLP Workshop; 2022. p. 87–92.
49. Carriño CP, Llop J, Pàmies M, Gutiérrez A, Armengol J, Silveira J, et al. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In: Proc. of 21st BioNLP; 2022. p. 193–199.
50. Lange L, Adel H, Strötgen J, Klakow D. CLIN-X: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. Bioinform. 2022;38(12):3267–74.
51. Chen Y, Zhou C, Li T, Wu H, Zhao X, Ye K, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. J Biomed Inform. 2019;96: 103252.
52. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. J Biomed Inform. 2020;102: 103356.
53. Madkour M, Benhaddou D, Tao C. Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput Methods Prog Biomed. 2016;128:52–68.
54. Strötgen J, Gertz M. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In: Proc. of 5th Int. Workshop Semant. Eval.; 2010. p. 321–324.
55. Miller T, Bethard S, Dligach D, Lin C, Savova G. Extracting time expressions from clinical text. In: Proc. of 15th BioNLP; 2015. p. 81–91.
56. Najafabadipour M, Zanin M, Rodríguez A, Torrente M, Nuñez B, Bermudez JL, et al. Reconstructing the patient's natural history from electronic health records. Artif Intell Med. 2020;105: 101860.
57. Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In: Proc. AMIA Symp.. vol. 2011. American Medical Informatics Association; 2011. p. 843.
58. Boland MR, Tu SW, Carini S, Sim I, Weng C. EliXR-TIME: a temporal knowledge representation for clinical research eligibility criteria. Proc AMIA summits translat sci. 2012;2012:71.
59. Chapman W, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301–10.
60. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomedical Inform. 2009;42(5):839–51.
61. Lima S, Perez N, Cuadros M, Rigau G. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In: Proc. 12th LREC. Marseille, France; 2020. p. 5772–5781.
62. Rivera R, Martinez P. The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study. JMIR Med Inform. 2020;8(12): e18953.
63. Dalloux C, Claveau V, Grabar N, Silva LE, Cabral C, Bonescki Y, et al. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. Nat Lang Eng. 2021;27(2):181–201.
64. Solarte-Pabón O, Montenegro O, Torrente M, González AR, Provencio M, Menasalvas E. Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach. PeerJ Comp Sci. 2022;8: e913.
65. Fabregat H, Duque A, Martinez-Romo J, Araujo L. Negation-based transfer learning for improving biomedical named entity recognition and relation extraction. J Biomed Inform. 2023;138: 104279.
66. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A, et al. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Med Inform. 2021;9(3): e17934.
67. Alfattni G, Belousov M, Peek N, Nenadic G, et al. Extracting drug names and associated attributes from discharge summaries: Text mining study. JMIR Med Inform. 2021;9(5): e24678.
68. Park J, Fang Y, Ta C, Zhang G, Idnay B, Chen F, et al. Criteria2query 3.0: leveraging generative large language models for clinical trial eligibility query generation. Available at SSRN 4637800;.
69. Peikos G, Symeonidis S, Kasela P, Pasi G. Utilizing ChatGPT to Enhance Clinical Trial Enrollment. arXiv preprint arXiv: 2306.02077. 2023;.
70. Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, et al. Zero-shot clinical entity recognition using ChatGPT. arXiv preprint arXiv:2303.16416. 2023;.
71. Hamer DMd, Schoor P, Polak TB, Kapitan D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. arXiv preprint arXiv:2304.07396. 2023.
72. Segura-Bedmar I, Martínez P, Revert R, Moreno-Schneider J. Exploring Spanish health social media for detecting drug effects. In: BMC Med. Inform. Decis. Mak.. vol. 15. BioMed Central; 2015. p. 1–9.
73. Gonzalez-Agirre A, Marimon M, Intxaurrondo A, Rabal O, Villegas M, Krallinger M. PharmaCoNER: Pharmacological Substances, Compounds and Proteins Named Entity Recognition Track. In: Proc. of 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 1–10.
74. Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. On the creation of a clinical gold standard corpus in Spanish: mining adverse drug reactions. J Biomed Inform. 2015;56:318–32.
75. Atutxa A, Pérez A, Casillas A. Machine learning approaches on diagnostic term encoding with the ICD for clinical documentation. IEEE J Biomed & Health Inform. 2017;22(4):1323–9.
76. Chizhikova M, López-Úbeda P, Collado-Montañez J, Martín-Noguerol T, Díaz-Galiano MC, Luna A, et al. CARES: a Corpus for classification of Spanish Radiological reports. Comput Biol Med. 2023;154: 106581.
77. Cotik V, Filippo D, Roller R, Uszkoreit H, Xu F. Annotation of entities and relations in Spanish radiology reports. In: Proc. of RANLP; 2017. p. 177–184.

78. Fabregat H, Araujo L, Martinez J. Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases. Comput Methods Prog Biomed. 2018;164:121–9.
79. Goenaga I, Andres E, Gojenola K, Atutxa A. Advances in monolingual and crosslingual automatic disability annotation in Spanish. BMC Bioinform. 2023;24(1):1–18.
80. Báez P, Bravo-Marquez F, Dunstan J, Rojas M, Villena F. Automatic extraction of nested entities in clinical referrals in Spanish. ACM Trans Comput Healthc. 2022;3(3):1–22.
81. Dellanzo A, Cotik V, Lozano Barriga DY, Mollapaza Apaza JJ, Palomino D, Schiaffino F, et al. Digital surveillance in Latin American diseases outbreaks: information extraction from a novel Spanish corpus. BMC Bioinform. 2022;23(1):1–22.
82. Ruiz A, López JL, González I, Ruiz B. Sequence Time Expression Recognition in the Spanish Clinical Narrative. In: 2019 IEEE 32nd Int. Symp. Computer-Based Med. Syst. (CBMS). IEEE; 2019. p. 331–336.
83. Sánchez-de Castro A, Araujo L, Martínez J. RoBERTime: a novel model for the detection of temporal expressions in Spanish. Proces Leng Nat. 2023;70(1):39–51.
84. Moreno I, Boldrini E, Moreda P, Romá-Ferri MT. Drugsemantics: a corpus for named entity recognition in Spanish summaries of product characteristics. J Biomed Inform. 2017;72:8–22.
85. Cotik V, Stricker V, Vivaldi J, Rodríguez Hontoria H. Syntactic methods for negation detection in radiology reports in Spanish. In: Proc. of 15th BioNLP. Association for Computational Linguistics; 2016. p. 156–165.
86. Campillos-Llanos L, Martınez P, Segura-Bedmar I. A preliminary analysis of negation in a Spanish clinical records dataset. In: Proc. 1st Workshop Spanish Negation (NEGES); 2017. .
87. AEMPS. Nomenclátor de Prescripción. [Retrieved on 2024-02-23]. 2023;Available from: www.aemps.gob.es.
88. Manlhiot C, van den Eynde J, Kutty S, Ross HJ. A primer on the present state and future prospects for machine learning and artificial intelligence applications in cardiology. Can J Cardiol. 2022;38(2):169–84.
89. Pustejovsky J, Castano JM, Ingria R, Sauri R, Gaizauskas RJ, Setzer A, et al. TimeML: Robust specification of event and temporal expressions in text. New Dir Quest Answ. 2003;3:28–34.
90. Styler WF, Bethard S, Finan S, Palmer M, Pradhan S, De Groen PC, et al. Temporal annotation in the clinical domain. TACL. 2014;2:143–54.
91. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. JAMIA. 2010;17(5):514–8.
92. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. JAMIA. 2020;27(1):3–12.
93. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. JAMIA. 2011;18(5):552–6.
94. Thompson P, Daikou S, Ueno K, Batista-Navarro R, Tsujii J, Ananiadou S. Annotation and detection of drug effects in text for pharmacovigilance. J Cheminform. 2018;10(1):1–33.
95. Báez P, Villena F, Rojas M, Durán M, Dunstan J. The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish. In: Proc. of 3rd Clinical NLP workshop; 2020. p. 291–300.
96. Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I, et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA open. 2021;4(2):ooab025.
97. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinform. 2008;9(11):1–9.
98. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Lang Resour Eval. 2018;52:571–601.
99. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. JAMIA. 2005;12(3):296–8.
100. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Proc of Advanc Neural Inf Process Syst. 2017;p. 5998–6008.
101. He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In: Proc. of ICLR; 2023. .
102. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. 2019;.
103. Schuster M, Nakajima K. Wordpiece Tokenization. In: IEEE Int. Conf. Acoustics, Speech & Signal Process.. vol. 1; 2012. p. 5149–5152.
104. de la Iglesia I, Atutxa A, Gojenola K, Barrena A. EriBERTa: A Bilingual Pre-Trained Language Model for Clinical Natural Language Processing. arXiv preprint arXiv:2306.07373. 2023.
105. Bahdanau D. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.
106. Luong T, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. In: Màrquez L, Callison-Burch C, Su J, editors. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 1412–1421.
107. WHO. WHO model list of essential medicines: 18th list, April 2013. World Health Organization; 2021.
108. Phan U, Nguyen N. Simple Semantic-based Data Augmentation for Named Entity Recognition in Biomedical Texts. In: Proc. of 21st BioNLP; 2022. p. 123–129.
109. Nakayama H. seqeval: A Python framework for sequence labeling evaluation; 2018.
110. Lannelongue L, Grealey J, Inouye M. Green algorithms: quantifying the carbon footprint of computation. Adv Sci. 2021;8(12):2100707.
111. Nejadgholi I, Fraser KC, de Bruijn B. Extensive error analysis and a learning-based evaluation of medical entity recognition systems to approximate user experience. In: Proc. of 19th BioNLP. Online; 2020. p. 177–186.
112. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform. 2015;54:213–9.
113. Hosseini A, Reddy S, Bahdanau D, Hjelm RD, Sordoni A, Courville A. Understanding by Understanding Not: Modeling Negation in Language Models. In: Proc. of 2021 Conf. NAACL; 2021. p. 1301–1312.

114. Lossio-Ventura JA, Sun R, Boussard S, Hernandez-Boussard T. Clinical concept recognition: evaluation of existing systems on EHRs. Front Artif Intell. 2022;5:1051724.
115. García Subies G, Barbero Jiménez Á, Martínez Fernández P. A comparative analysis of Spanish Clinical encoder-based models on NER and classification tasks. J Am Med Inf Assoc. 2024. https://doi.org/10.1093/jamia/ocae054.
116. Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proc. 12th ACM Conf. Bioinform., Comp. Biol. & Health Inform.; 2021. p. 1–6.
117. Kury F, Butler A, Yuan C, Fu Lh, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Sci data. 2020;7(1):1–11.
118. Tseo Y, Salkola M, Mohamed A, Kumar A, Abnousi F. Information extraction of clinical trial eligibility criteria. arXiv preprint arXiv:2006.07296. 2020;.
119. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574.
120. Grouin C, Grabar N, Hamon T, Rosset S, Tannier X, Zweigenbaum P. Eventual situations for timeline extraction from clinical reports. JAMIA. 2013;20(5):820–7.
121. Navas M, Rodríguez V. Annotador: a temporal tagger for Spanish. J Intell Fuzzy Syst. 2020;39(2):1979–91.
122. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. JAMIA. 2010;17(1):19–24.
123. Dobbins NJ, Mullen T, Uzuner Ö, Yetisgen M. The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria. Sci Data. 2022;9(1):490.
124. Bhatia P, Celikkaya B, Khalilia M, Senthivel S, Comprehend medical: a named entity recognition and relationship extraction web service. In,. 18th IEEE Int. Conf. Mach. Learn. & Applications (ICMLA). IEEE. 2019;2019:1844–51.
125. Zanoli R, Lavelli A, do Amarante DV, Toti D. Assessment of the E3C corpus for the recognition of disorders in clinical texts. Nat Lang Eng. 2023;30(4):851–69.
126. Solarte-Pabón O, Montenegro O, García-Barragán A, Torrente M, Provencio M, Menasalvas E, et al. Transformers for extracting breast cancer information from Spanish clinical narratives. Artif Intell Med. 2023;143: 102625.

Publisher's Note